Structural Studies of Multidomain Proteins of the Immunoglobulin Superfamily

Thesis Presented for the Degree of

Doctor of Philosophy

By

Mark Karl Boehm

Department of Biochemistry and Molecular Biology

Royal Free Campus

Royal Free and University College Medical School

University College London

September 1998

This work is dedicated to the memory of

Mr Clement Wheeler-Bennett

Abstract

Carcinoembryonic antigen (CEA) is an important tumour marker. It is heavily glycosylated and contains seven immunoglobulin (Ig) domains. Its solution structure was determined by X-ray and neutron scattering. X-rays showed it has a radius of gyration (R_G) of 8.0 nm and a cross-sectional radius of gyration (R_{XS}) of 2.1 nm. Neutron data showed that CEA is monomeric. Models of CEA were built using domains from homologous structures. The models which best-fitted the experimental data had elongated "zig-zag" structures and inter-domain orientations similar to those in CD2.

IgA is the most abundant class of human immunoglobulin. IgA1 contains two Fabs joined by hinges to an Fc, and two C-terminal tailpieces. Solution scattering showed that IgA1 has an R_G of 6.11-6.20 nm and IgA1 lacking tailpieces has an R_G of 5.84-6.16 nm. It was predicted that the hinges are extended and that the tailpieces are compact. Automated curve-fitting modelling confirmed both of these predictions, and showed that IgA1 is "T-shaped", in which the tailpieces fold back against the Fc. This structure is unlike those observed for IgG proteins.

MFE-23 is an anti-CEA single chain Fv antibody that is used for targeting tumours. It contains two domains joined by a flexible linker. The MFE-23 crystal structure was solved by molecular replacement, and the final model had an R-factor of 19.0% at 2.4 Å resolution. Its antigen-binding loops have well-defined structures. In the structure, MFE-23 forms dimers. Neutron scattering showed that MFE-23 exists as monomers below 1 mg/ml and oligomers at higher concentrations.

The interaction between MFE-23 and CEA was modelled using lattice contacts in the MFE-23 crystal. In this model, the antigen-binding loops formed contacts with the first two domains of CEA, there was good surface complementarity, appropriate electrostatic contacts, and no steric conflicts with CEA carbohydrate.

The models will be most useful for studying the functions of these proteins.

Acknowledgements

Foremost, I would like to thank my supervisor Prof. S. J. Perkins for the considerable time and energy he has dedicated to the work presented in this thesis over the past four years. I would also like to thank the many members of Prof. Perkin's group, who have provided assistance to many aspects of this work, most especially during data collection, computing and sample preparation.

For the work on CEA and MFE-23, I am indebted to the members of the Clinical Oncology Department at the Royal Free and University College Medical School. In particular, I wish to express my gratitude to Dr P. A. Keep and Mr J. D. Thornton for their work on protein purification and crystallization. In addition, the input and guidance from Prof. R. H. J. Begent and Dr K. A. Chester has been invaluable. For the IgA solution scattering studies, I would like to thank Prof. M. A. Kerr and Dr J. M. Woof from the Department of Molecular and Cellular Pathology at the University of Dundee for providing the purified protein samples and for their critical assistance in preparing this part of the thesis. Access to X-ray and neutron facilities has been essential for the solution scattering work. I must therefore thank the Biotechnology and Biological Sciences Research Council for providing such facilities, and Dr E. Towns-Andrews, Mrs S. Slawson, Mr A. Gleeson, Dr P. A. Timmins, Dr R. K. Heenan and Dr S. M. King for generous instrumental support. The X-ray crystallography work was carried out at the Randall Institute, Kings College, London. Accordingly, I would like to thank Dr Maninder Sohi, Dr Tommy Wan and Dr Adam Corper for crystallizing MFE-23, collecting diffraction data, and guiding me through the data processing and model building process, respectively. I offer thanks to Dr Brian Sutton for making these facilities available and for his critical contribution to this project.

I would like to give special thanks to the Clement-Wheeler Bennett trust for providing me with a studentship, and for the continual interest that they have shown in this. And, I also thank Prof. K. D. Bagshawe for overseeing this studentship.

Finally, I would like to acknowledge my friends and family for providing encouragement, support and a welcome reminder of reality throughout.

Contents **Page**

Contents: Figures **Page**

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Contents: Tables **Page**

List of Abbreviations

ADEPT	Antibody-directed enzyme prodrug therarpy
BGP	biliary glycoprotein
CCP4	Collaborative Computational Project, number 4
CD	circular dichroism
CDR	complementarity-determining region
CEA	carcinoembryonic antigen
cDNA	complementary DNA
CDR	complementarity-determining region
CGM	CEA gene family members
CPU	central processing unit
DNA	deoxyribonucleic acid
FPLC	fast performance liquid chromatography
FT-IR	Fourier transform infrared
Fuc	Fucose
Gal	Galactose
GlcNAc	N-acetyl glucosamine
GPI	glycosyl phosphatidyl-inositol
HEL	hen egg-white lysozyme
ICAM-2	intercellular adhesion molecule-2
Ig	immunoglobulin
IgSF	immunoglobulin superfamily
IL-1	interleukin-1
ILL	Institute Laue-Langevin
LINAC	linear accelerator
Man	Mannose
MFE	Man from Eden
MHC	major histocompatibility complex
mRNA	messenger RNA
NCA	non-specific cross-reacting antigen
NeuNAc	N-acetyl neuraminic acid
NMR	nuclear magnetic resonance
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDB	protein databank
PSG	pregnancy-specific glycoprotein
RAL	Rutherford Appleton Laboratory
RMS	root mean squared
RNA	ribonucleic acid
ScFv	single-chain Fv
SCR	structurally conserved region
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SH	src homology
SPECT	Single-photon emission computerized tomography
SRS	Synchrotron Radiation Source
TcR	T-cell receptor
VCAM-1	vascular cell adhesion molecule-1

Amino Acid Abbreviations

Amino acid	3 letter format	1 letter format
alanine	Ala	A
arginine	Arg	R
aspartate	Asp	D
asparagine	Asn	N
cysteine	Cys	C
glutamate	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histidine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Trp	W
tyrosine	Tyr	Y
valine	Val	V

Chapter 1

The Immunoglobulin Fold:
An Important Domain of Protein Structure

1.1. Introduction to protein structure

During the course of evolution, an immensely diverse array of proteins has arisen to perform the myriad of functions associated with living organisms. Detailed structural knowledge of a protein provides an improved understanding of its biological function and the atomic coordinates for more than 7,000 experimentally determined protein structures have been deposited in the Brookhaven Protein Data Bank (PDB; March 1998; Bernstein *et al.*, 1977). Arguably the most important feature of globular protein structure is the domain (Wetlaufer, 1973).

1.1.1. Basic principles of domain structure

A domain is a region of contiguous polypeptide that folds independently of the remainder of the sequence and it represents a basic subunit within the protein tertiary structure. Inspection of the cores of globular domains reveals that they consist almost entirely of non-polar, hydrophobic amino acids such as leucine, valine, phenylalanine and tryptophan, while their solvent-accessible surfaces are composed mainly of polar, hydrophilic sidechains. This partitioning of globular proteins is a result of the hydrophobic effect, which is entropic and describes the association of non-polar groups in order to reduce their contact with water. Van der Waals interactions favour close atomic packing, and in globular domain structures the sidechains of buried core residues generally pack more tightly and are more ordered than the sidechains of surface residues (for recent review; Levitt *et al.*, 1997).

Upon formation of a hydrophobic core, hydrogen bonds are formed between buried polar groups. Of major importance are the hydrogen bonds between mainchain amide and carboxyl groups that produce α-helix and β-sheet secondary structures. An α-helix is localised to a region of consecutive residues, and is formed by a hydrogen bond between each mainchain carboxyl group and the amide group which is found 3.6 residues towards the C-terminal end of the polypeptide. A β-sheet is formed by reciprocal hydrogen bonds between the mainchain carboxyl and amide groups on separate regions of polypeptide chain that run parallel or antiparallel to each other. It is often useful to describe a globular domain structure in terms of its secondary structure packing or topology. The strict mainchain geometries required to form α-helices and

β-sheets can be considered to produce rigid structural elements which pack, by means of sidechain interactions, to form the domain core. The folding pattern of a globular protein is then the arrangements and connections of these secondary structure elements (Chothia & Finkelstein, 1990).

Protein structures may be stabilized by the formation of salt bridges between oppositely charged amino acid residues. Alternatively, structures may be stabilized by covalent disulphide bridges between pairs of cysteine residues.

The amino acid sequence of a domain must have the appropriate physical and chemical properties to fold into a compact and functional three-dimensional structure. The critical sequence length necessary for the formation of a globular domain seems to be about 40 amino acids, e.g. the epidermal growth factor domain (Montelione *et al.*, 1987). The upper domain size is estimated to be in the region of 400 residues (Islam *et al.*, 1995), e.g. cytochrome p450eryf (Cupp-Vickery & Poulos, 1995).

1.1.2. Domains and protein evolution

The domain represents the fundamental unit of protein evolution. It is proposed that once a domain gene arose, selection forces caused its propagation throughout evolution by continuous duplication, shuffling and modification events. According to this theory, all naturally-occurring proteins have originated from a relatively small number of primordial domain genes and the existence of ubiquitous folds, such as the nucleotide-binding fold (Rossman *et al.*, 1974), the flavin-binding fold (Correll *et al.*, 1993) and the haem-binding domains (Vasudevan *et al.*, 1991), supports this view. Domains that share a common ancestry will have similar fold structures. If an unknown domain structure can be shown to be related to a known structure, homology modelling enables qualified predictions to be made about its three-dimensional structure (for recent review; Srinivasan *et al.*, 1996). Identifying evolutionary relationships between proteins is therefore an invaluable area of protein structure research.

Sequence similarities are used to identify related domains and the concepts of families and superfamilies are used to describe these relationships. Dayhoff and

colleagues assigned two proteins to the same family if their sequences were more than 50% identical, and two proteins to the same superfamily if they had more than 30% sequence identity (Dayhoff *et al.*, 1972). It is well established that 30% identity represents a threshold, above which domains are confidently predicted to adopt the same fold structure (Sander & Schneider, 1991; Flores *et al.*, 1993; Hilbert *et al.*, 1993), but related domains that have the same fold structure may have identities below this threshold. For example, globin fold domains can have sequence identities as low as 15% (Lesk & Chothia, 1980). Therefore, it is often more convenient to use a more flexible definition in which proteins are classed in the same family if they are evidently related, regardless of their sequence identity, and families are grouped in the same superfamily if homology is demonstrable to overlap their member proteins (Doolittle, 1981). If the identity is low, an evolutionary relationship may be established by refining the sequence comparison using matrices to quantify amino acid differences, e.g. the Dayhoff matrices (Dayhoff *et al.*, 1978). Alternatively, esoteric knowledge such as functional similarities or important sequence motifs may be used to identify related sequences. It is generally estimated that there are about 1000 protein superfamilies (Chothia, 1992).

1.1.3. Domain structure comparisons

Classifying domains by comparisons of their three-dimensional domain structures aids the understanding of protein structures. At one level, structural classifications instil a welcome degree of order to the thousands of different structures in the PDB, but the information obtained from them has also been used for several alternative functions. For example, structural classifications have allowed general rules of protein folding patterns to be elucidated (Chothia & Finkelstein, 1990). They may also allow the identification of important sequence motifs that are not apparent from sequence data alone. Perhaps the best demonstration of the value of rigorous structural classification is the success that has been achieved in modelling the three-dimensional structures of the antigen-binding domains of antibodies using template loop structures (for recent review; Rees *et al.*, 1996).

In the past few years several databases have been established that classify

domain structures in the Brookhaven PDB, e.g. CATH (Orengo *et al.*, 1997), FSSP (Holm & Sander, 1998) and SCOP (Murzin *et al.*, 1995). The basic principle of each database is to provide a hierarchical system for categorizing domain structures. In the FSSP database, an exhaustive all-against-all alignment of three-dimensional structures is performed, and folds are systematically grouped according to the similarity score generated by the alignment algorithm. The CATH and the SCOP databases both classify proteins using combinations of automatic procedures and visual inspection, and in both databases the classification ranges from the predominant type of secondary structure at the lowest similarity level, to domain families which have significant sequence similarities, even though the number and definitions of hierarchical levels differ between the two databases. All three databases are routinely updated and are accessible via the World Wide Web.

1.1.4. Domain "superfold" structures

The classification of domain structures has identified the existence of domains which have similar fold structures but do not exhibit an evolutionary relationship. This phenomenon was first recognised in the 1970's when similarities were observed between the domain structures in flavodoxin and lactate dehydrogenase (Rao & Rossman, 1973). After systematic classification of the PDB, Orengo *et al.* (1994) noted nine domain "superfold" structures which are adopted by domains with neither sequence nor functional similarity, and that 32% of domains in the PDB database adopted one of these "superfold" structures. These nine "superfolds" are shown in Figure 1.1 and are termed the α/β doubly wound fold, the triosephosphate isomerase barrel, the split $\alpha\beta$ sandwich, the Greek-key immunoglobulin fold, the α-up-down fold, the globin fold, the jelly roll, the trefoil fold and the ubiquitin $\alpha\beta$ roll. One explanation of "superfolds" is that they represent extra stable structures that can tolerate high levels of sequence mutations, which mask the relationships between domains (divergent evolution). However, an alternative explanation is that, because of its high stability, a "superfold" could have evolved independently on more than one occasion (convergent evolution).

The evolution of many different domains from a single primordial gene is known as divergent evolution and the opposite of divergent evolution is convergent evolution.

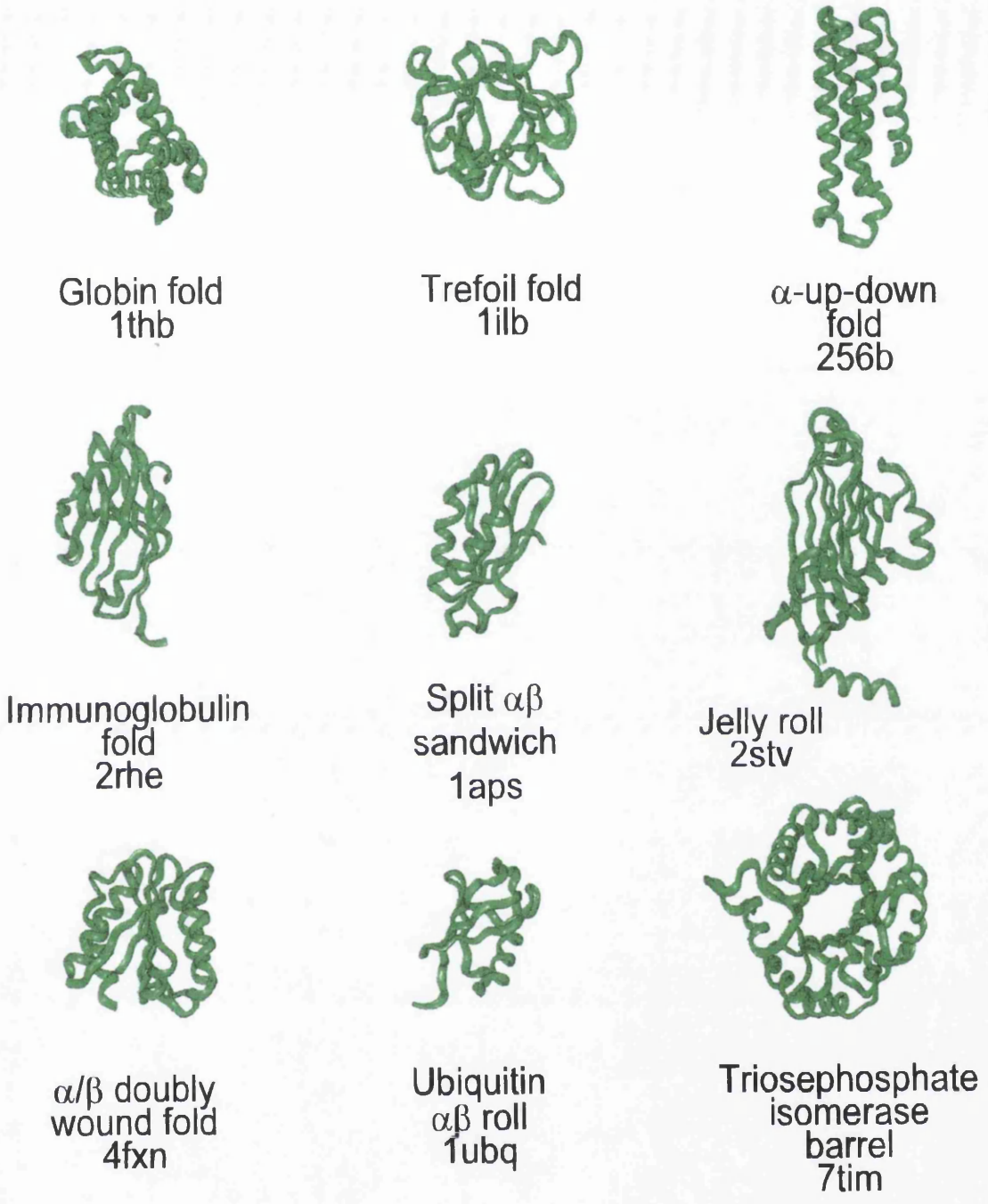


Figure 1.1. The nine "superfolds". An example for each of the nine "superfold" structures is shown as a ribbon representation. The PDB accession code for each structure is given (Adapted from Orengo *et al.*, 1994).

It has long been known that different enzymes have evolved independently to catalyze the same reaction. For example, serine proteases, superoxide dismutases, sugar kinases and alcohol dehydrogenases have all evolved on more than one occasion (Doolittle, 1994). A more striking type of convergence is mechanistic convergence. The crystal structures of chymotrypsin and subtilisin revealed that, despite having different fold structures, these two serine proteases both utilize a triad of a histidine, an aspartic acid and a serine to catalyze their reactions. Identical domain folds which have originated from different ancestor genes would represent an even more remarkable example of convergent evolution. However, it is impossible to determine whether a "superfold" has evolved by convergence or whether any evolutionary relationship has been obscured by highly divergent evolution.

It is observed that in each of the "superfolds" a high percentage of sequential secondary structures lie adjacent in the tertiary structure. Consequently, it has been proposed that this confers simplistic many-direction folding pathways, which could result in extra-stability (Orengo *et al.*, 1994). The existence of "superfold" structures has prompted the development of analogy modelling techniques. This type of modelling relies on identifying whether a domain sequence without structural homologues is compatible with a seemingly unrelated structure. An example of the successful application of analogy modelling has been the prediction that the von Willebrand factor type A domain adopts a similar fold to the GTP binding domain of ras-p21 (Edwards & Perkins, 1995; reviewed in Perkins *et al.*, 1998b).

1.1.5. Multidomain and multichain proteins

A polypeptide chain may fold into a single domain or into many domains strung together. The sequence of titin, a human muscle filament protein, predicts a 3,000,000 Da polypeptide containing more than 240 globular domains (Labeit & Kolmerer, 1995). It is often convenient to consider multidomain proteins as being constructed from the fusion of distinct structural and functional subunits (Campbell & Downing, 1994; Doolittle, 1995). This is particularly true for "mosaic" proteins that consist of more than one type of domain fold. It is a feature of certain biological pathways that their constituent proteins are dominated by a few domain fold types which perform the

characteristic functions of the pathway. For example, blood clot haemostasis proteins have a high incidence of several domain types including serine protease and epidermal growth factor-like domains (Pathy, 1993), while SH type 1 and type 2 domains are often found in combination in intracellular signalling proteins (Koch *et al.*, 1991; Waksman *et al.*, 1992).

A protein may also consist of more than one polypeptide chain. This protein tertiary structure may be important for fashioning active sites by the association of domains from more than one chain. Disulphide bridges and non-covalent interactions between domain surfaces are utilized in holding chains together.

1.2. The immunoglobulin superfamily

1.2.1. Overview

A major domain type found in protein structures is the globular fold that was first observed in immunoglobulin (Ig) proteins. The immunoglobulins are produced in vertebrates to bind antigens as part of the adaptive immune response to infection. A common ancestry for the immunoglobulins was evident from the first polypeptide chains to be sequenced. Both the internal homology within Ig chains and high sequence identity between light and heavy Ig chains lead to the proposition that the immunoglobulins evolved from a primordial gene encoding approximately 110 amino acids (Hill *et al.*, 1966). Edelman showed that the homology regions within Ig chains corresponded to a regular distribution of intrachain disulphide bridges, and concluded that the homology regions formed compact domains with similar three-dimensional structures (Edelman, 1970). This was confirmed when the first structure from an immunoglobulin was determined (Poljak *et al.*, 1973). The immunoglobulins consist of two types of domains, which are known as the V-set and the C1-set Ig folds. In both of these folds, the core consists of two anti-parallel β-sheets, which pack closely to form a sandwich structure. The two folds share a common folding pattern, as evidenced by identical connections between their constituent β-strands. The V-set fold has two more β-strands than the C1-set fold, and this is distinguishable at the sequence level: V-set domains have approximately 65 to 70 residues between the conserved disulphide bridge compared to 55 to 60 residues in C1-set domains, and each fold type has a characteristic

pattern of conserved amino acids (Williams & Barclay, 1988).

It became apparent that the Ig fold has a wider significance for protein structures when sequence similarities identified Ig fold domains in proteins other than immunoglobulins. The immunoglobulin superfamily (IgSF) is used to describe proteins containing Ig fold domains. The first non-immunoglobulin member of the IgSF to be identified was β_2-microglobulin (Peterson *et al.*, 1972). It was predicted to contain a single C1-set domain, and this was confirmed upon determination of its crystal structure (Becker & Reeke, 1985). β_2-microglobulin forms the light chain of major histocompatibility complex (MHC) class I antigen presentation molecules, and the heavy chain of MHC class I molecules also contains a C1-set domain (Orr *et al.*, 1979; Bjorkman *et al.*, 1987). It was therefore proposed that Ig fold domains are confined to proteins which play a role in immune recognition. However, the sequence of Thy-1, a cell surface molecule which is not involved in immune recognition, predicted that it consists of a single V-set Ig fold (Williams & Gagnon, 1982). It was subsequently discovered that there is a widespread occurrence of Ig fold domains in cell surface proteins (Williams, 1987; Williams & Barclay, 1988).

The sequences of cell surface IgSF proteins revealed that many putative Ig domains are intermediate between the V-set and C1-set folds (Williams, 1987). The classification of this third type of Ig fold was based on the patterns of residues that are conserved between domains of the same set. This Ig fold was termed the C2-set fold because its sequence length is closer to C1- than to V-set folds. When X-ray crystal structures were determined for the cell surface IgSF proteins CD4 (Ryu *et al.*, 1990; Wang *et al.*, 1990) and CD2 (Jones *et al.*, 1992), domains which had been classified as C2-set were indeed shown to have a distinct fold structure. This fold is similar to the C1-set fold except that one of the β-strands has switched sheets, and the topology of this fold has generally been accepted as the C2-set fold structure. However, the crystal structure of the single-domain IgSF member, telokin, revealed a fourth variant of the Ig fold (Holden *et al.*, 1992). This novel Ig fold shares structural characteristics with both the V-set and C1-set Ig fold structures, and consequently it has been termed the "intermediate" or I-set Ig fold (Harpaz & Chothia, 1994). It has been proposed that it

represents the ancestor of the other three Ig fold structures (Chothia & Jones, 1997). This view is supported by the occurrence of I-set domains in bacteria (Bateman *et al.*, 1996), marine sponges (Gamulin *et al.*, 1994) and *Caenorhabditis elegans* (Fong *et al.*, 1996). The sequence characteristics that are important for determining the I-set fold structure were characterised by detailed examination of the telokin structure, and many of the Ig fold domains which were initially classified as C2-set have subsequently been predicted to have the I-set fold (Williams *et al.*, 1988; Harpaz & Chothia, 1994).

The immunoglobulin superfamily covers a vast number of proteins. It has been predicted, for example, that 40% of the cell surface proteins from human leukocytes contain Ig fold domains (Barclay *et al.*, 1993), while it is claimed that in animals Ig-like domains represent the most widespread subunit of protein structure (Doolittle & Bork, 1993). Therefore, in the presence of such a large group of proteins, it is important to consider the structural mechanisms by which functional diversity is generated from the Ig fold structure.

1.2.2. Structural features of the Ig fold

The Ig fold is an all β-class protein fold which consists of two antiparallel β-sheets packing face-to-face to form a sandwich structure. The two β-sheets exhibit aligned packing, in which the direction of all β-strands is essentially either up or down the fold (Figure 1.2a). In β-strands, the mainchain conformation results in the sidechains of consecutive residues lying on opposite sides of the mainchain and therefore it is common to observe hydrophobic residues which are buried in the core of the fold, alternating with surface exposed polar residues. The buried residues at the interface between the two sheets are aligned and, because the β-strand mainchain conformation causes a right-handed twisting of the antiparallel β-sheets, the two sheets pack with an angle of about -30° between the directions of their strands in order to accommodate the aligned packing of sidechains (Figure 1.2a; Chothia & Janin, 1981).

The polypeptide chain traverses the length of the fold an odd number of times (either 7 or 9 times) so that its N-terminal and C-terminal residues lie on opposite sides

Figure 1.2. (Overleaf) The structural sets of the immunoglobulin superfamily.

(*a*) Ribbon diagrams showing a representative structure of each of the four fold sets; the V-set structure is from the light chain of the J539 mouse IgA Fab (PDB code: 2fbj; residues L1 to L107), the C1-set structure is from the heavy chain of the J539 Fab (H121 to H220), the C2-set structure is the second domain of human CD4 (3cd4; 96 to 178) and the I-set structure is the first domain of human VCAM-1 (1vca; A1 to A89). The four structures are orientated in the same way, with their N-terminal residues at the top and their DEBA faces at the front. The β-strands, as defined by DSSP (Kabsch & Sander, 1983) are shown as thick arrows.

(*b*) Two-dimensional topology diagrams for each of the four Ig fold sets. An arrow is used to represent each of the 7 to 10 strands (A, A', B, C, C', C", D, E, F, and G) and its N- to C-terminal direction. The strands are not drawn to scale, but the arrangements of the strands and their connections are correctly depicted. Strands (D)EBA form the front sheet and strands (A')GFC(C'C") form the back sheet and the broken line indicates the edge of the fold, about which the back sheet can be considered to wrap back against the front face. The C' strand which is often observed in the C1- and I-set structures has been omitted for simplicity.

(*c*) The structure that is usually conserved in each of the four fold sets (adapted from Chothia & Jones, 1997). Solid circles represent β-sheet residues, open circles represent residues in conserved loops and horizontal broken lines represent mainchain hydrogen bonds. The β-strand nomenclature used in (*b*) is maintained, but it should be noted that the order of the back sheet strands has been reversed. This is to illustrate the manner in which the two sheets interact, by which the (D)EBA sheet sits directly on top of the (C")C'CFG(A') sheet. The positions usually occupied by the cysteine residues on strands B and F that form the intersheet disulphide bridge are shown as squares.

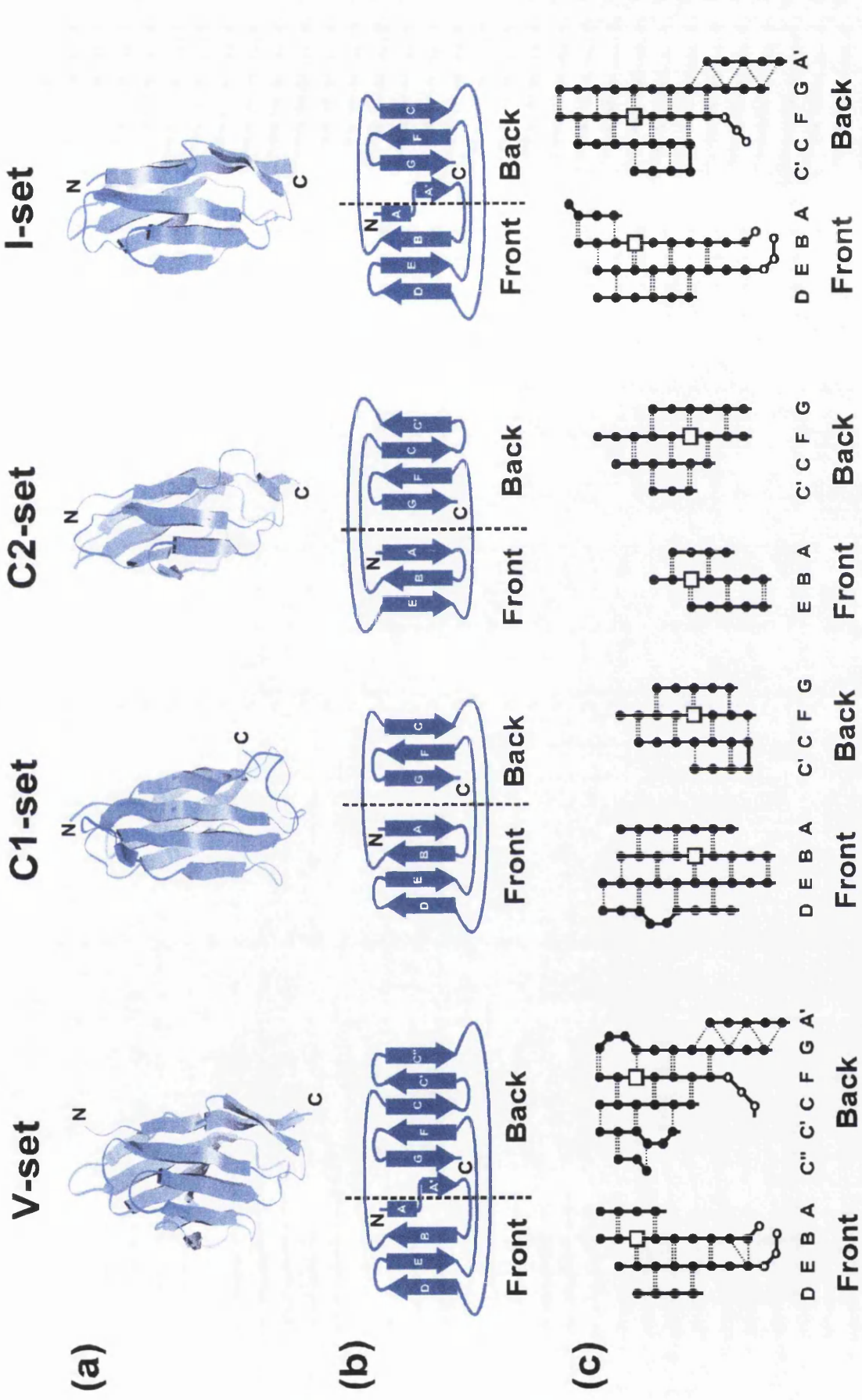


Figure 1.2. The structural sets of the immunoglobulin superfamily (legend on page 11).

of the fold. The Ig fold consists of 7 to 10 β-strands which are labelled sequentially from A for the N-terminal strand to G for the C-terminal strand, and additional strands are labelled A', C' and C", depending where they occur in the sequence. Of the ten possible strands observed in Ig fold structures, strands D, E, B and A are arranged in that order to form one sheet, while strands A', G, F, C, C' and C" occur in that order to form the second sheet (Figure 1.2).

1.2.3. The four classes of Ig fold

The four classes of Ig fold are characterized according to which of the ten possible β-strands they possess. An example of each of the four Ig sets is depicted as a ribbon diagram in Figure 1.2a. Two-dimensional topology diagrams for the four types of Ig fold are shown in Figure 1.2b (Bork *et al.*, 1994). In Figure 1.2c, the conserved β-sheet residues of the four Ig fold sets are represented according to the method of Chothia and co-workers (Harpaz & Chothia, 1994; Chothia & Jones, 1997). The IgSF structures in the Brookhaven database, as identified by SCOP, CATH, and FSSP, are summarised in Figure 1.3, and the corresponding references are listed in Table 1.1.

The V-set domain from the light chain of the J539 IgA Fab fragment is shown in Figure 1.2a (Suh *et al.*, 1986). In the V-set fold the polypeptide chain traverses the length of the domain nine times (Figure 1.2b). The V-set Ig fold has a maximum of ten β-strands, and the extra strand arises from the polypeptide chain switching sheets after strand A to form the A' strand. The ten strands are arranged as a four strand DEBA β-sheet, packed against a six strand A'GFCC'C" β-sheet. Both β-sheets are antiparallel except for the short parallel β-ladder between strand A' and strand G (Figure 1.2c). V-set domains with the full complement of ten β-strands are found in immunoglobulins, T-cell receptor (TcR) proteins and P_0 (Figure 1.3). Variations of this fold have also been observed. When the structure of the first two domains of CD4 was solved, the first domain was found to be similar to the V-set fold from immunoglobulins but lacking β-strand A. The third domain of CD4 and the first domain of CD2 also have this variant of the V-set fold. The V-set fold of CD8 has the A strand but, although the mainchain follows the approximate direction of the A' strand, mainchain hydrogen-bonds are not made with strand G, and it therefore lacks the A' strand.

Figure 1.3. (Overleaf) The known immunoglobulin superfamily structures. Immunoglobulin (IgSF) structures were identified using CATH (September 1997 release; Orengo *et al.*, 1997), FSSP (30th March 1998; Holm & Sander, 1998) and SCOP (February 1998 release; Murzin *et al.*, 1995). The chain and domain content is shown for each protein. The four-character PDB accession code is given for each structure, and a line is used to depict the domains it contains, and the source (B = bovine, C = *Caenorhabditis elegans*, H = human, M = mouse, R = rat and T = turkey) and the method of structure determination (N = NMR, X = X-ray crystallography) are shown in parentheses. The Ig fold structures were classified according to SCOP, except for the interleukin-1 receptor (Il-1R) where the domain assignments of Schreuder *et al.* (1997) were used.

The known Ig fold structures are represented by coloured circles: V-set structures are yellow, C1-set structures are blue, C2-set structures are magenta and I-set structures are green. The N- and the C-termini of the proteins are indicated and the common B-F disulphide bridge is labelled if it is present. Many of the proteins are bound to the plasma membrane either by means of a transmembrane peptide or a glycosyl phosphatidyl-inositol (GPI) anchor, as indicated.

IgG was selected as a representative antibody structure and its domains are labelled using alternative scheme that is commonly used for antibodies: the light chain V- and C1-set domains are labelled V_L and C_L respectively, the heavy chain V-set domain is labelled V_H and its three C1-set domains are labelled in the N- to C-terminal direction as C_H1, C_H2 and C_H3. The antibody fragments, the Fab and the Fc, are also labelled. The Fab is the association of a light chain with the V_H-C_H1 domain pair and the Fc is the association of the C_H2 and C_H3 domains from two heavy chains. There are numerous structures of IgG Fab fragments or their component domains so only two examples are given.

Similarly there are numerous MHC class I structures so again only two representative structures have been listed. In the MHC class I and class II structures, the polymorphic N-terminal domains, which are shown as red triangles, do not have the Ig fold. CD1 and the rat neonatal Fc receptor have similar structures to MHC class I molecules.

In NCAM and VCAM-1, domains whose structures have not been solved are coloured black. The unknown NCAM domains have been predicted to be I-set Ig folds or fibronectin type III (FnIII) structures (Chothia & Jones, 1997). The unknown VCAM-1 domains are labelled as C2-set domains (Brümmendorf & Rathjen, 1995) although it is possible that some of these domains have I-set structures. Of the muscle proteins, telokin consists of a single I-set domain whereas titin and twitchin contain many domains of different types. Structures are only known for the 5th and 27th Ig fold domains of titin and for the 18th Ig fold of twitchin.

Figure 1.3. The known immunoglobulin superfamily structures (legend on page 14).

Table 1.1. Summary of the IgSF protein entries in the Brookhaven Protein Databank

PDB Code	Protein	Reference
1ao7	TcR	Garboczi *et al.*, 1996
1bec	TcR	Bentley *et al.*, 1995
1cd1	CD1	Zeng *et al.*, 1997
1cd8	CD8	Leahy *et al.*, 1992
1cdb	CD2	Withka *et al.*, 1993
1cdh, 1cdi	CD4	Ryu *et al.*, 1994
1cdj, 1cdu, 1cdy	CD4	Wu *et al.*, 1996
1cid	CD4	Brady *et al.*, 1993
1dlh	MHC class 2	Stern *et al.*, 1994
1fc1, 1fc2	Ig	Deisenhofer., 1981
1frt	Neonatal Fc receptor	Burmeister *et al.*, 1994b
1fru	Neonatal Fc receptor	Burmeister *et al.*, 1994a
1gya	CD2	Wyss *et al.*, 1995
1hnf	CD2	Bodian *et al.*, 1994
1hng	CD2	Jones *et al.*, 1992
1hoc	MHC class 1	Young *et al.*, 1994
1hsc	MHC class 1	Madden *et al.*, 1992
1iea	MHC class 2	Fremont *et al.*, 1996
1igt	Ig	Harris *et al.*, 1997
1igy	Ig	Harris *et al.*, 1998
1itb	IL-1 receptor	Vigers *et al.*, 1997
1mlb	Ig	Braden *et al.*, 1994
1neu	P_0	Shapiro *et al.*, 1996
1nct, 1ncu	Titin	Pfuhl *et al.*, 1997
1seb	MHC class 2	Jardetzky *et al.*, 1994
1tcr	TcR	Garcia *et al.*, 1996
1tit, 1tiu	Titin	Improta *et al.*, 1996
1tlk	Telokin	Holden *et al.*, 1992
1tnm, 1tnn	Titin	Pfuhl *et al.*, 1995
1vca	VCAM-1	Jones *et al.*, 1995
1vsc	VCAM-1	Wang *et al.*, 1995
1wit	Twitchin	Fong *et al.*, 1996
1zxq	ICAM-2	Casasnovas *et al.*, 1997
2fb4	Ig	Marquart *et al.*, 1980
2ncm	NCAM	Thomsen *et al.*, 1996
3cd4	CD4	Garrett *et al.*, 1993

The C1-set Ig fold from the heavy chain of the J539 IgA Fab fragment is shown in Figure 1.2a (Suh *et al.*, 1986). In the C1-set Ig fold the polypeptide chain traverses the length of the domain seven times (Figure 1.2b). The C1-set Ig fold has seven or eight β-strands, and the additional strand arises as a short C' strand as the polypeptide chain switches sheets between strands C and D (Figure 1.2c). It should be noted that in the seven-stranded C1-set structures the mainchain between strands C and D roughly follows the path of the C' strand but the appropriate hydrogen-bonds cannot be detected by conventional means, such as DSSP (Kabsch & Sander, 1983). The β-strands are arranged as a four strand DEBA β-sheet which is packed against a three or four strand GFC(C') β-sheet. Both β-sheets are fully antiparallel (Figure 1.2c). The first C1-set (C_H1) and the third C1-set (C_H3) domains of the IgG heavy chain have been observed to have seven β-strands, whereas the eight strand variant has been observed in the second C1-set domain (C_H2) of the IgG heavy chain, in light Ig chains (C_L), in the light and heavy chains of both classes of MHC classes, CD1 and the rat neonatal Fc receptor which are both similar to MHC class I, and in the β-chain of the TcR (Figure 1.3). SCOP also classifies the C-terminal domain of the TcR α-chain as a C1-set domain, but the GFC(C') β-sheet is not formed correctly and instead the polypeptide chain loosely follows the direction of the β-strands without forming mainchain hydrogen bonds (PDB code: 1tcr; Garcia *et al.*, 1996).

The C2-set Ig fold of the second domain of CD4 is shown in Figure 1.2a (Garrett *et al.*, 1993). In the C2-set Ig fold the polypeptide chain traverses the length of the domain seven times (Figure 1.2b). The C2-set fold is characterised by a short polypeptide connection between the end of strand C and the beginning of strand E, relative to the C1-set fold that causes the D strand observed in C1-set structures to switch sheets and thereby form a C' strand. Consequently this structure has also been called the switch or S-type Ig fold (Bork *et al.*, 1994). The C2-set Ig fold has seven β-strands which are arranged as a three strand EBA β-sheet packed against a four strand GFCC' β-sheet. Both β-sheets are antiparallel (Figure 1.2c). The second and fourth domains of CD4, and the second domains of CD2, vascular cell adhesion molecule-1 (VCAM-1) and intercellular adhesion molecule-2 (ICAM-2) have all been shown to have C2-set Ig fold structures (Figure 1.3). Extra β-sheet outside of the core

EBA|GFCC' fold has also been observed in C2-set fold structures, e.g. CD4 (PDB code: 3cd4; Wang *et al.*, 1990), VCAM-1 (PDB code: 1vca; Jones *et al.*, 1995) and ICAM-2 (Casasnovas *et al.*, 1997). In these structures, the loop between strands A and B sometimes forms mainchain hydrogen-bonds with strand G to form a short parallel β-sheet, but it has a different conformation to the A' strand observed in V-set domains.

The I-set Ig fold of VCAM-1 is shown in Figure 1.2a (Jones *et al.*, 1995). In the I-set Ig fold the polypeptide chain traverses the length of the domain seven times (Figure 1.2b). The I-set domain has nine β-strands, and like the V-set fold one extra strand arises because the polypeptide chain switches sheets after strand A to form the A' strand. The second extra strand is as a short C' strand formed as the polypeptide chain switches sheets between strands C and D, which is analogous to the C1-set structure (Figure 1.2c). The nine strands are arranged as a four strand DEBA β-sheet that is packed against a four strand A'GFCC' β-sheet. Both β-sheets are antiparallel except for the short parallel β-ladder between strand A' and strand G (Figure 1.2c). The first domains of VCAM-1, ICAM-2, NCAM, and telokin, and certain domains from titin and twitchin have I-set Ig fold structures (Figure 1.3). In analogy to variants observed for C1-set domain structures, the I-set domain from ICAM-2 lacks the C' strand (Casasnovas *et al.*, 1997). The three domains from two recently published structures of the interleukin-1 (IL-1) receptor have similar topologies to the I-set fold structure (Schreuder *et al.*, 1997; Vigers *et al.*, 1997). The second domain of the IL-1 receptor has eight of the nine I-set fold strands, but lacks strand A which is analogous to the CD2 and CD4 variants of the V-set fold. The first domain of the IL-1 receptor lacks strands A and C'.

1.2.4. The structural core of the Ig fold

An appealing view of the Ig fold is one of a rigid structural core surrounded by flexible peripheral strands (Bork *et al.*, 1994). The four strands, B, C, E and F, which are present in all Ig fold structures, are conformationally and sequentially invariant by comparison to the other β-strands (Williams, 1987). Strands B and E form an antiparallel β-ladder in the centre of one sheet that packs against the antiparallel β-ladder formed by strands C and F in the centre of the other sheet, and this forms the structural core of the Ig fold (Figure 1.4). The structural core is stabilised by intersheet loops

Figure 1.4. (Overleaf) The structural core of the immunoglobulin fold. The structurally conserved β-strands (B, C, E and F) are shown for the V-set domain from the light chain of mouse IgA Fab J539 (PDB code: 2fbj). Strand B is residues L19 to L25, strand C is L31 to L37, strand E is L69 to L74 and strand F is L83 to L91. The four strands and the N- and C-terminal residues of each strand are labelled, and the mainchain of each strand is shown as a black ribbon representation. Atoms are coloured according to type; green is carbon, blue is nitrogen, red is oxygen and yellow is sulphur. Black broken lines illustrate mainchain hydrogen bonds. The structure is orientated so that the N-terminal residue (not shown) would be at the top and so that the sandwich is seen side- on. Strands E and B form an antiparallel β-ladder in one sheet. The sheets pack together to bury mainly hydrophobic residues in the core of the fold (e.g. LeuL32, TrpL34, LeuL72, IleL74, IleL21 and TyrL85) and expose polar residues on the surface (SerL69, SerL71, ThrL73, ThrL20 and ThrL22). The intersheet disulphide bridge formed between CysL23 on strand B and CysL87 on strand F stabilizes the fold.

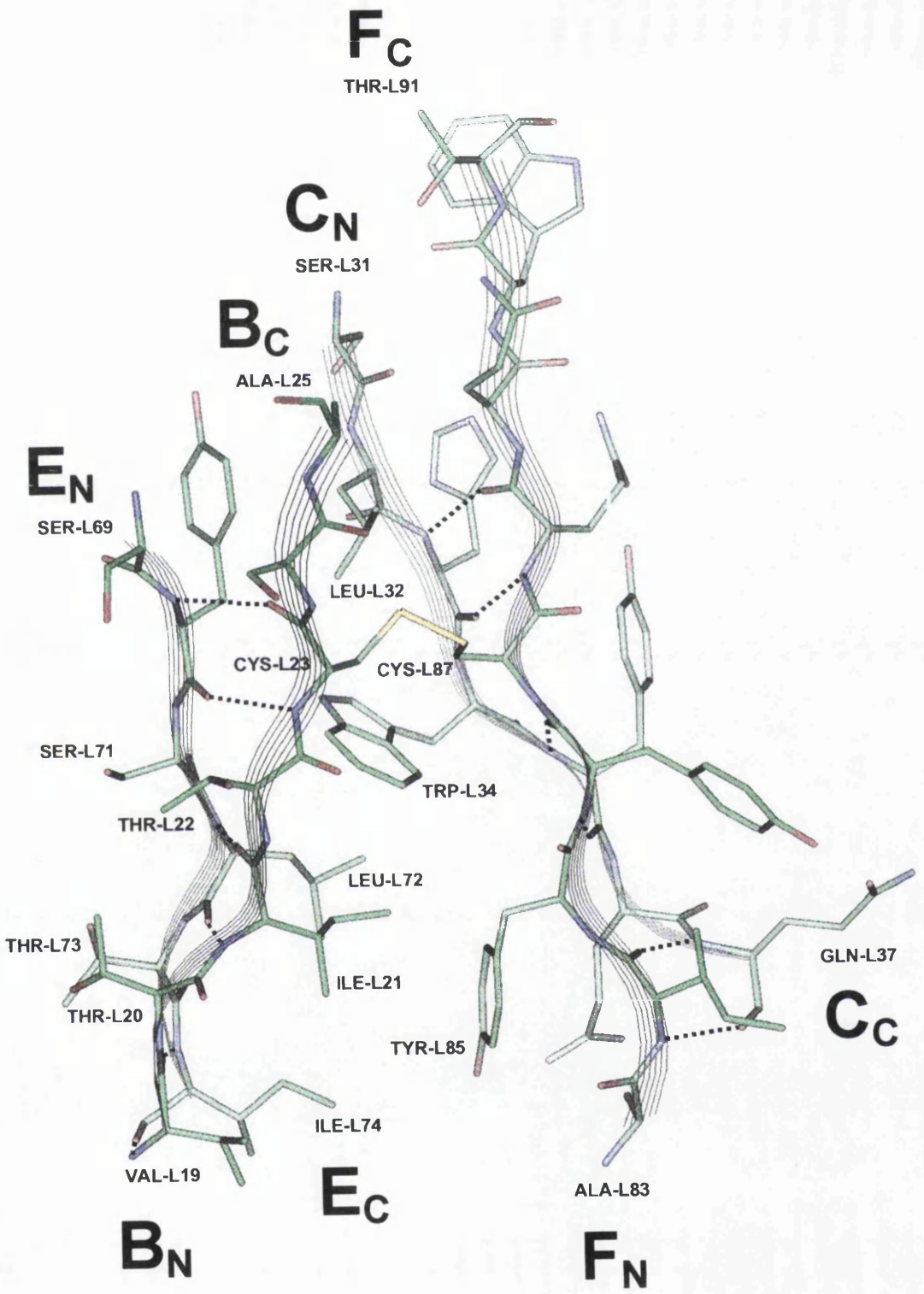


Figure 1.4. The structural core of the immunoglobulin fold (legend on page 19).

between strands B and C, and strands E and F, and by the aligned packing of mainly hydrophobic sidechains (Figure 1.4). In a recent folding study, Parker *et al.* (1998) verified the importance of this structural core by showing that the mainchain hydrogen bonds which are crucial to the folding of the V-set domain of CD2 are formed by a cluster of residues on strands B, C, E and F. The four sets of Ig fold result from differences in five of the other strands A, A', C', C" and D, all of which occur at the periphery of β-sheets. Within each Ig fold type, these strands and the other peripheral β-strand (G) exhibit conformational flexibility.

The perception that Ig folds consist of a rigid structural core surrounded by flexible peripheral strands raises the possibility that there are other topologies in addition to the four Ig fold sets. By analogy to the loss of strand A in some V-set fold structures and the switching of the "C'/D" strand between sheets in the C1 and C2 sets, it has been proposed that variants on the C1- and C2-sets that have strand A switched to the GFC(C') sheet could also exist (Bork *et al.*, 1994). The first two domains of the IL-1 receptor have been described as I-set fold structures but, while both domains have an A' strand, they lack A strands (Vigers *et al.*, 1997). It will be interesting to see whether these domains represent an unusual variation on the I-set fold, or whether they become adopted as a new Ig-set topology. The tolerance of conformational variability in the loops and edge strands the Ig fold, as well as the ability to withstand modifications in surface-exposed β-sheet residues, provides the means for the Ig fold to achieve structural and hence functional diversity.

Immunoglobulin-like folds have been observed in proteins that do not belong to the immunoglobulin superfamily and the Ig fold was one of the nine "superfold" structures identified by Orengo *et al.* (1994). Domain types which share the Ig-fold topology include fibronectin type III (FnIII) domains (de Vos *et al.*, 1992), the PapD chaperone protein from *Escherichia coli* (Holmgren & Bränden, 1989) and cadherin domains (Shapiro *et al.*, 1995). These domain types all possess a β-sandwich structure and exhibit the same arrangement and connectivity of β-strands that is possessed by the Ig-fold, but lack similarity to Ig fold sequences. The existence of these analogous domain structures has led to the proposition that the Ig-fold, like other "superfold"

structures, is highly stable and tolerant of sequence mutations. This proposition is independent of whether the analogous domains arose by divergent or convergent evolution and reveals how a single, stable fold structure can give rise to multiple protein functions. It is also worth noting that proteins containing FnIII or cadherin domains are often cell-surface proteins that perform adhesion functions analogous to those performed by certain IgSF proteins.

1.2.5. Intradomain disulphide bridges

In the V- and C1-set folds of immunoglobulin proteins, the core structure is 'pinned' by an intradomain disulphide bridge, formed by cysteines on strands B and F, and a conserved tryptophan residue on strand C, which it packs against (Figure 1.4; Lesk & Chothia, 1982). This disulphide bridge was thought to be essential for stabilizing the Ig fold structure. While the B-F disulphide bridge is conserved in many structures of all four Ig sets (Figure 1.3), it is not found in all Ig-fold structures (Williams & Barclay; 1988). Of the known V-set domain structures, the first domain of human (Withka *et al.*, 1993) and rat CD2 (Driscoll *et al.*, 1991), and the third domain of rat CD4 (Brady *et al.*, 1993) do not contain the B and F cysteine residues. In muscle IgSF proteins, I-set domains fold without the formation of a B-F disulphide bridge. In telokin (Holden *et al.*, 1992) and the 27th Ig domain of titin (Improta *et al.*, 1996) there are four and two buried cysteines respectively, however disulphide bridges are not formed because they are intracellular proteins. In the second domain of human CD4 (Ryu *et al.*, 1990; Wang *et al.*, 1990), which has the C2-set structure, the B-F disulphide bridge is replaced by a buried disulphide bridge between cysteines on strands C and F. This disulphide bridge is unusual because it is formed between adjacent cysteines on neighbouring antiparallel β-strands: a mainchain conformation which is unfavourable for the required geometry of the bond (Richardson & Richardson, 1989).

Some Ig fold structures have other intradomain disulphide bridges in addition to the B-F disulphide bond. For example, in the second domain of CD2, an additional disulphide bridge is formed between a cysteine on the A-B loop and a cysteine at the C-terminal end of strand G (Jones *et al.*, 1992). In the I-set domains of ICAM-2 (Casasnovas *et al.*, 1997) and VCAM-1 (Jones *et al.*, 1995; Wang *et al.*, 1995) an extra

disulphide, which is located between the B-C and F-G loops, produces a narrow N-terminal tip to these molecules and this could influence their interactions with viruses (Wang *et al.*, 1995). A different situation is observed in the structure of the CD8 α_2 homodimer that has a third buried cysteine on strand C, as well as the B-F disulphide (Leahy *et al.*, 1992). This third cysteine has the appropriate geometry to form an alternative disulphide bridge with the B strand cysteine, and the B-C disulphide bridge has been detected biochemically in the α-chain of CD8 αβ heterodimers (Kirszbaum *et al.*, 1989). It has been suggested that formation of the B-C disulphide bridge could induce small structural changes in the V-set domain of CD8 that thereby change its properties (Leahy *et al.*, 1992).

1.2.6. The IgSF structures

The known structures of IgSF proteins are summarized in Figure 1.3 and Table 1.1. Close evolutionary relationships mean that many of these structures can be loosely fitted into one of several functional groups. The immunoglobulins, the MHC proteins and the TcR proteins are all polymorphic immune recognition proteins. CD1 and the rat neonatal Fc receptor are similar to MHC class I molecules but are not polymorphic. CD1, like the MHC molecules, has an antigen-presenting, while the neonatal Fc receptor mediates the transfer of maternal IgG into the newborn. CD4 and CD8 are also immune system proteins and act as accessory molecules in the complex formed between an MHC molecule on an antigen-presenting cell, its associated peptide, and a TcR molecule on a T-lymphocyte. CD2, P_0, ICAM-2, NCAM and VCAM-1 are all cell-adhesion molecules (Chothia & Jones, 1997). Telokin, titin and twitchin are muscle proteins. The IL-1 receptor is a cytokine receptor and initiates a cascade of inflammatory responses upon binding of an agonist. The functions of these proteins are consistent with the view that IgSF proteins perform general adhesion functions.

An obvious source of variability between IgSF proteins is the utilization of different Ig fold sets. The distribution of Ig domains throughout the known IgSF structures implies conserved evolutionary, structural and functional roles for the four Ig fold sets (Figure 1.3). V-set structures have been determined for immune system proteins and cell adhesion molecules and they are typically adjacent to a C1- or C2-set

domain, with the exception of CD8 and P_0. In multi-chain immune system proteins, the V-set domains from two different polypeptide chains associate by means of reciprocal interactions. In the IgSF structures, V-set domains always occur at the first (N-terminal) domain position, apart from CD4 which has a V-set domain at the third domain position. The presence of a V-set domain at the head of many IgSF structures suggests a fundamental role for this domain in these proteins: a view which has been shaped by the polymorphic, binding regions on the V-set domains of antibodies (Wu & Kabat, 1970). C1-set domains have only been observed in immune recognition proteins and a C1-set domain is found in association with a C1-set domain from a second polypeptide chain. In these structures, C1-set domains do not occur at the N-terminal domain position. C2-set domains have only been observed in the structures of cell surface molecules and no structures currently exist which have an N-terminal C2-set domain. I-set domains have also been observed in cell adhesion molecule structures and only I-set fold structures have been determined for intracellular muscle proteins. In the known structures I-set domains occur at the N-terminal domain in VCAM-1 and ICAM-2, at middle domain positions in titin and telokin, while telokin corresponds to the C-terminal domain of myosin light chain kinase.

1.2.7. Multidomain IgSF structures

The number of component domains can differ greatly between IgSF proteins. This is strikingly evident upon comparison of the major membrane protein of peripheral nerve myelin, P_0, whose extracellular region consists of a single Ig domain, and the muscle proteins twitchin and titin, which respectively contain 30 and 112 Ig domains in addition to other domain types. IgSF proteins also differ in the number of polypeptide chains that they contain. For example, the immune system IgSF proteins tend to have more than one chain; immunoglobulin monomers have four polypeptide chains, while MHC molecules, TcR molecules and CD8 all comprise two chains. Varying the numbers of domains and chains is a fundamental means of generating structural and hence functional variability in the IgSF superfamily.

The concatenation of multiple Ig domains is favoured by the general structure of the Ig fold, which has its N- and C-termini on opposite sides of the fold. It is evident

from the existing IgSF structures that the joining of Ig domains presents another means of producing structural variability. Several different types of polypeptide linker have been used to join domains and seemingly small differences between these linkers can have profound effects on the gross morphology of the protein.

In the two types of CD4 structure - which contain either domains 1 and 2 (Ryu *et al.*, 1990; Wang *et al.*, 1990) or domains 3 and 4 (Brady *et al.*, 1993) - the domains are joined directly without a polypeptide linker between domains, which results in the G strand of the N-terminal domain (domain 1 or 3) running continuously into the A strand of the C-terminal domain (2 or 4). This type of linking results in rigid, rod-like structures in which the two domains are twisted relative to each other about the long axis of the structure. Despite the similarity of the connections between the two types of CD4 structure, there is a 30° difference in the orientation of domains 3 and 4 compared to domains 1 and 2 (Brady *et al.*, 1993). These structures are stabilized by the burial of a large surface area, e.g. 880 $Å^2$ is buried between domains 1 and 2 of CD4 (Jones *et al.*, 1995).

A different method of linking Ig domains is observed in the two domain structures of CD2, VCAM-1 and ICAM-2. These cell adhesion molecules all have a short polypeptide linker (approximately six amino acids) between their two domains. In CD2 the linker is flexible, as evidenced by domain re-orientations between separate CD2 structures (Jones *et al.*, 1992; Bodian *et al.*, 1994), and it produces a 13° to 41° bend between the two domains, as calculated by Casasnovas *et al.* (1997). The linker separates the domains relative to CD4, which reduces the surface area buried at their interface to approximately 400 $Å^2$ (Jones *et al.*, 1995). Domain re-orientations are observed between independent VCAM-1 structures, which indicates that it also has a flexible linker (Jones *et al.*, 1995; Wang *et al.*, 1995). This linker produces a 5° to 35° bend between the two domains (Casasnovas *et al.*, 1997). However the buried surface area is 850 $Å^2$, which is more similar to CD4 than CD2 (Jones *et al.*, 1995), and this is because the loops in the C-terminal C2-set domain are much longer in VCAM-1 than in CD2, and produce a flexible, hydrophobic "cradle" for the N-terminal domain (Wang *et al.*, 1996). The structure of ICAM-2 resembles VCAM-1 in that long loops on its C-

terminal domain produce a large buried interface between its two domains (Casasnovas *et al.*, 1997). A 35° bend is observed between the two domains, but it is not known whether it has a flexible linker because there is only one molecule in the crystal asymmetric unit. The linkers discussed so far have all been from cell surface molecules and the linkers produce extended, twisted structures. It seems likely that these linkers are important for enabling molecules to extend away from the plasma membrane and thereby expose functional regions. The different degrees of bend and twist that these linkers produce between tandem domains create widely variable surfaces at the interfaces of domains. Flexible linkers could be important for correctly aligning the adhesive surface of a molecule with its ligand.

Short polypeptide linkers also connect domain pairs on immunoglobulin chains and TcR molecules. For example, in IgG such linkers occur between the two domains on the light chain, and between the first and second, and the third and fourth domains on the heavy chain. These linkers produce the necessary bends between tandem domains to allow the domain associations between different polypeptide chains (Padlan, 1994).

The immunoglobulins also demonstrate a third means of linking domains; a long polypeptide hinge that separates neighbouring domains thereby preventing interactions between them. In the two known crystal structures of intact IgG, mouse anti-canine lymphoma IgG2a (Harris *et al.*, 1992, 1997) and mouse anti-phenobarbital IgG1 (Harris *et al.*, 1998a), the orientation of the Fab fragments relative to the Fc is variable within each structure and between the two structures, and this clearly shows that the hinge is flexible. The two Fab and the Fc fragments are arranged to have a "distorted T shape" in the IgG2a structure and a "distorted Y shape" in the IgG1 structure (Harris *et al.*, 1998a). The conformation of hinge peptides may be restricted by cysteines forming interchain disulphide bridges and proline residues, which typically occur in the middle region of the hinge (Burton, 1990), or by interactions between loops on the third heavy chain domain with hinge residues proximal to the Fc fragment (Harris *et al.*, 1998a). Consequently, the upper part of the hinge (that proximal to the Fab fragment) exhibits the greatest degree of flexibility (Burton, 1990). The flexing of the Fab fragments is

presumed to be important for efficient binding to antigens.

In the three-domain structure of the IL-1 receptor there are two different types of interdomain connection, which are unlike those seen in the other IgSF structures. The unusual structure of this molecule is due to all three Ig domains forming the ligand binding site and in complex with an agonist or antagonist it resembles a "question mark" (Schreuder *et al.*, 1997; Vigers *et al.*, 1997). The first and second domains have a linker of approximately eleven residues but they are closely associated due to an interdomain disulphide bridge. This long domain connection enables the binding residues on the two domains to be orientated correctly for ligand binding. The connection between the second and third domains is approximately nine residues and is long and flexible like the antibody hinge. It keeps the third domain separate from the first and second domains, and it appears to function in closing the IL-1 receptor around the ligand upon binding.

1.2.8. Ig domain interactions

Immunoglobulin superfamily proteins have general adhesive functions. The known structures reveal that the Ig fold has evolved to utilize different regions of its structure to perform the specific interactions of these proteins. These interactions may be non-covalent interactions that stabilize tertiary structure, or they may be the homophilic or heterophilic interactions that convey protein functionality. Interactions may involve either the GFCC'C" or DEBA faces as these present flat surfaces suitable for forming multiple contacts or they may involve specific loop.

One type of interaction is the reciprocal interaction between the GFCC'C" faces of two V-set domains, which was first observed in the antibody Fv fragment. In this interaction, the orientation of the two GFCC'C" sheets is approximately parallel, i.e. there is only an acute angle between the directions of the strands in the two sheets relative to each other (not to be confused with the directions of neighbouring strands within the β-sheets). The packing of the two GFCC'C" sheets resembles a β-sandwich except that the main contacts are formed by the sidechains of edge strand residues (Chothia *et al.*, 1985). In both GFCC'C" sheets, β-bulges in stands G and C' are

important in forming the interaction (Chothia *et al.*, 1985; Colman, 1988). The association of two V-set domains can vary significantly between Ig Fv structures and in some structures this mobility of the GFCC'C" interface is important for reshaping the antigen-binding site upon ligand binding (Davies & Padlan, 1992; Padlan, 1994). An analogous interaction is observed between the two V-set domains in the TcR (Garcia *et al.*, 1996). In CD8 dimers, the two V-set domains also interact through their GFCC'C" faces, which have the characteristic G and C' bulges (Leahy *et al.*, 1992). In the known multi-chain IgSF structures, the reciprocal packing of GFCC'C" faces is the only means by which V-set domains dimerize to stabilize protein tertiary structure.

The GFCC'C" face of the Ig fold has also been implicated in homophilic interactions between cell adhesion molecules. In the crystal structures of rat and human CD2, two symmetry-related CD2 molecules pack together through interactions between the GFCC'C" faces of their V-set domains (Jones *et al.*, 1992; Bodian *et al.*, 1994). The packing of these faces can be regarded as being antiparallel because there is an 180° rotation between the directions of the strands in the two GFCC'C" sheets relative to each other. Although these homophilic interactions are not directly significant *in vivo* (van der Merwe *et al.*, 1993), the natural ligands of rat and human CD2 are the homologous IgSF molecules CD48 and CD58 respectively (Arulanandam *et al.*, 1993) and it is proposed that their adhesion complexes are formed by interactions similar to those observed between two CD2 molecules. Significantly, these GFCC'C" faces are very flat in both forms of CD2 and their binding specificities therefore seem likely to be determined by the distributions of charged residues (Bodian *et al.*, 1994). Symmetric packing is also observed between the GFCC'C" faces of molecule pairs in the P_0 crystal structure and the C' strand is particularly important for this interaction (Shapiro *et al.*, 1996). The reciprocal interactions between GFCC'C" faces of V-set domains provides an important model of homophilic adhesion between two IgSF molecules.

The opposite type of interaction on the reverse side of the Ig-fold, i.e. the reciprocal packing of the DEBA faces of two domains, is observed to stabilize the association of two C1-set domains. This type of packing has been observed between the C_H1 and C_L domains in the Fab fragments of immunoglobulins (Padlan, 1994) and

between the two C_H3 domains in the IgG Fc structure (Deisenhofer, 1981), which is approximately parallel in both cases. Analogous dimerization of C1-set domains occurs in the structures of MHC molecules (Bjorkman *et al.*, 1987; Brown *et al.*, 1993), CD1 (Zeng *et al.*, 1997) and rat neonatal Fc receptor (Burmeister *et al.*, 1994a). The interaction of the two C1-set domains of the TcR are is like the C_H3-C_H3 interaction, but the C1-set domain on the TcR α-chain has an unusual structure and an Asn-linked oligosaccharide on this domain makes interactions at the interface of the two C1-set domains. In the known IgSF structures, the reciprocal packing of DEBA faces is the only means by which two C1-set domains directly associate. Dimers of C1-set domains seem to be important for stabilizing the tertiary structures of larger proteins.

An important type of functional interaction is antigen binding by the immunoglobulins. The reciprocal GFCC'C" sheet packing of two Ig V-set domains is essential for the correct formation of the antigen-binding site, which is made up of three CDRs from each domain. The CDRs correspond to the B-C, C'-C" and F-G loops and in the V_H domain they are termed CDR-*H1*, CDR-*H2* and CDR-*H3* respectively, while in the V_L domain they are termed CDR-*L1*, CDR-*L2* and CDR-*L3*. These six loops form a continuous surface that is exposed at the N-terminal tip of the antibody molecule. The sequences and hence the structures of these loops are modified to mediate specific antigen binding. In the structures of antibody-antigen complexes the interactions made by the six CDRs vary although CDR-*H3*, which is the most variable and is located at the centre of the binding site, plays a prominent role (Padlan, 1994). A similar binding site is observed in the TcR. It also has six polymorphic CDRs but CDR-*1* and CDR-*2* of each V-set domain are less variable than their counterparts in antibodies. In the only known structure of a TcR molecule bound to a peptide-presenting MHC molecule, CDR-*1* and CDR-*3* of both V-set domains make contact with the peptide, while all three CDRs from the α-chain V-set domain and CDR-*3* from the β-chain V-set domain make contact with the MHC molecule (Garboczi *et al.*, 1996).

In the structures of certain non-polymorphic IgSF molecules, loops that correspond to the immunoglobulin CDRs have also been implicated in adhesive functions. P_0 crystallizes under near physiological conditions with four-fold symmetry

and it is proposed that P_0 molecules on the same cell surface associate to form tetramers (Shapiro *et al.*, 1996). Within these tetramers, the B-C (CDR-*1*) loop on one molecule interacts with the C"-D and E-F loops of the next.

Another distinct region of the Ig fold has been adopted for binding integrins. VCAM-1, ICAMS-1, -2 and -3, and mucosal addressing cell adhesion molecule-1 (MAdCAM-1) form an IgSF subclass of integrin-binding molecules. Structures are known for integrin-binding motifs in the N-terminal I-set domains of VCAM-1 and ICAM-2 (Jones *et al.*, 1995; Wang *et al.*, 1995 & 1996; Casasnovas *et al.*, 1997). In both molecules, the integrin-binding motif contains an acidic residue and occurs on the C-D loop. This loop connects the two sheets at the edge of the fold and, despite being situated at the C-terminal end of the fold, it is highly exposed because it is distal from the interdomain contacts. Site-directed mutagenesis experiments showed that residues on the C-D edge and the GFC face of both proteins were important for integrin binding (Osborn *et al.*, 1994; Clements *et al.*, 1994; Staunton *et al.*, 1990; Klickstein *et al.*, 1996). A four-residue deletion in the C-D loop causes the C' strand to be lost in the ICAM-2 I-set domain and consequently the functional acidic residues in VCAM-1 and ICAM-2 were shown to be presented differently (Casasnovas *et al.*, 1997).

Recently, a novel type of IgSF interaction has been observed for the IL-1 receptor in complex either with IL-1β (Vigers *et al.*, 1997) or with its antagonist (Schreuder *et al.*, 1997). In both structures, even though the agonist and antagonist sites are partially different, all three I-set Ig fold domains make contact with ligand and as such the binding site is dependent upon the correct association of these domains, rather than through well-defined regions of loops or β-sheet. Residues that make contact with ligand in either one or both of the structures are located on; strand A and the B-C loop of the first domain; strands A and B, and loops B-C and D-E of the second domain; strands D, G, F and C, and loops B-C, D-E and G-F of the third domain; and the two inter-domain linkers. Therefore the mode of binding employed by the IL-1 receptor has been dependent upon the evolution of its gross structure, which has a characteristic "question mark" shape that wraps around a ligand by means of the flexible hinge between its second and third domains. The close association of the first two domains

of the IL-receptor is instrumental in forming an interaction site that involves the A'-B edge of the Ig fold. Although there are only a few examples of interactions performed by the Ig fold for which structures are known, it is evident that different modes of interaction have developed through the use of different regions of the fold.

1.2.9. Glycosylation of IgSF structures

With the exception of the muscle protein structures, the known IgSF structures are from extracellular proteins and accordingly many of them have glycosylation sites. This is also true for many of the IgSF proteins whose structures are not known (e.g. Brümmendorf & Rathjen, 1995). A few structures exist for IgSF proteins with Asn-linked oligosaccharide chains and these structures indicate that carbohydrate plays varied yet influential roles in IgSF proteins.

In both of the C_H2 domains of the IgG Fc structure, the DEBA face of the Ig fold is covered by a carbohydrate chain, which is attached to an asparagine residue in the D-E loop of each C_H2 domain (Deisenhofer, 1981). The carbohydrates pack the cavity between the two C_H2 domains. It has been remarked that even in the absence of carbohydrate the two C_H2 domains would be sterically-restricted from interacting directly through their DEBA faces (Padlan, 1994), and the carbohydrates therefore seem to be important for mediating weak interactions between the two domains.

In the structure of an αβ T-cell receptor complexed with MHC, four of seven possible Asn-linked glycosylation sites have ordered carbohydrates (Garcia *et al.*, 1996). While three of these carbohydrates are small or project into the solvent, the fourth and largest carbohydrate is involved in crystal contacts and interactions between the two C1-set domains. This carbohydrate is attached to the α-chain C1-set domain and possibly strengthens the association through its DEBA face with β-chain C1-set domain.

The two domain ICAM-2 protein could only be crystallized with the high-mannose glycans at its six Asn-linked sites intact and it is one of the most heavily glycosylated proteins to have been crystallized (Casasnovas *et al.*, 1997). All six oligosaccharides extend into solution. The three carbohydrates on the integrin-binding

I-set domain are distal to the binding motif on the C-D loop. The three glycans on the C2-set domain are evenly distributed around the membrane proximal region and it is thought that they could function in orienting the ICAM-2 molecule so that it stands perpendicular to the membrane.

The structure of the V-set domain of human CD2 with a single intact Asn-linked glycan has been studied by NMR (Withka *et al.*, 1993; Wyss *et al.*, 1995). This glycan is of the high-mannose type and the glycosylation site occurs on the opposite side of the fold to the GFCC'C" sheet, which has been implicated in the cell adhesion function of CD2 (Jones *et al.*, 1992). This glycosylation site is absent in rat CD2 but human CD2 fully-lacking this glycan loses the ability to bind CD58 and monoclonal antibodies (Recny *et al.*, 1992; Wyss *et al.*, 1995). In the NMR structure, this glycan is shown to interact with a lysine residue that is at the centre of five surface-exposed lysine residues on the DEB face of the fold. This glycan therefore plays an important role in stabilizing the V-set fold of human CD2 by means of specific interactions with polypeptide.

1.2.10. Summary and conclusions

Stable domain folds are the fundamental subunits of protein evolution and structure. The immunoglobulin fold is an important type of domain and sequence analyses have shown that it has evolved into a widespread superfamily of proteins. The known IgSF structures reveal that the Ig fold is a highly adaptable one. There are four recognised sets of Ig fold, which are classified according to their β-strand content. All Ig fold structures have a conserved four strand core (B, C, E and F) that is surrounded by conformationally variant edge strands A, A', C', C", D and G. Structural and hence functional diversity arises because the stability of the Ig fold enables it to withstand highly variable loop and edge strand structures, as well as modifications in its surface-exposed β-sheet residues. Different disulphide bridging patterns are also used to modify Ig fold structures. The IgSF structures vary in the number of domains that they contain and the manner in which domains are linked produces different types of structure: Ig domains may be connected directly to create rigid rod-like structures, or a short linker may be used to produce elongated structures that have a degree of flexibility, or a long hinge-like linker may be used to produce highly flexible connections between domains.

The Ig fold has evolved to utilize different regions for interactions. In the known structures, specific regions are often conserved in closely-related proteins to perform similar functions: the GFCC'C" face is used for interactions between two V-set domains in multichain proteins and in homophilic adhesion between two molecules, the DEBA face is used for interactions between C1-set domains in multichain proteins, the CDR loops are polymorphic in the immunoglobulins and TcR molecules and are used to bind antigens, the C-D loop in the VCAM-1 and ICAM-2 structures are used for binding integrins, and in the IL-1 receptor the ligand binding site is formed by the correct association of its three Ig fold domains. Glycosylation is also variable between IgSF proteins and is observed to play different roles in the known structures. Although the known immunoglobulin superfamily structures as a whole exhibit a wide range of variable features, the observation that closely-related proteins conserve specific structural and characteristics enables valid predictions to be made about unknown IgSF structures.

1.3. Structural techniques

The experimentally-determined atomic structures in the Brookhaven databank have been solved by either X-ray crystallography or NMR methods. Certain proteins are more amenable than others for study by either one or both of these techniques and a cursory examination of the known IgSF structures highlights some of the common difficulties associated with these techniques.

X-ray crystallography is the more powerful technique and is obviously dependent upon first obtaining a well-ordered crystal. Aside from the challenge in finding the conditions necessary for crystal growth, it is unlikely that some protein molecules could be crystallized at all. Glycosylation, for reason of its heterogeneity, and a high degree of flexibility are two characteristics often associated with multidomain extracellular proteins such as many members of the IgSF, and can prevent crystal formation. In order to overcome the problem of glycosylation, strategies have been used to produce unglycosylated protein variants or to enzymatically cleave significant amounts of the carbohydrate chains. For example, the crystal structure of P_0, which has a single Asn-linked glycosylation site, lacked carbohydrate because it was expressed in

Escherichia coli (Shapiro *et al.*, 1996), while human CD2 was endoglycosidase H treated so that only a single GlcNAc residue occupied each of its three Asn-linked sites (Bodian *et al.*, 1994). The problem of interdomain flexibility can be avoided by dividing the structure into smaller fragments and the study of antibody and CD4 structures exemplify this. There are more than one hundred crystal structures based on antibodies, but only two of these are full structures of intact antibodies (Harris *et al.*, 1997, 1998a). Crystals of intact, four domain CD4 have been obtained but they diffract poorly (Davis *et al.*, 1990; Kwong *et al.*, 1990). This is attributed to a long flexible linker separating the second and third domains and consequently only the tandem domain structures of domains one and two (Ryu *et al.*, 1990; Wang *et al.*, 1990) or domains three and four (Brady *et al.*, 1993) have been determined. An advantage of dividing large multidomain structures into smaller fragments is that this strategy is also used to determine the specific functions of domains (Campbell & Downing, 1994). Some general problems associated with protein structure determination by X-ray crystallography include the need to obtain heavy atom derivatives if homologous structures are not known, the use of non-physiological buffers to promote crystal growth and the effect of the crystal packing on the conformations of surface residues and even the domain arrangement within multidomain proteins.

The conventional application of NMR methods to the study of protein structure is limited by the size of the protein, and the upper limit is in the region of 20,000 Da (MacArthur *et al.*, 1994). Of the known IgSF structures, only structures of isolated Ig fold domains (approximately 10,000 Da) have been determined by NMR. These include the V-set domains of rat and human CD2 (Driscoll *et al.*, 1991; Withka *et al.*, 1993), and the I-set domains from titin (Pfuhl *et al.*, 1995; Improta *et al.*, 1996; Pfuhl *et al.*, 1997) and twitchin (Fong *et al.*, 1996). Other problems associated with NMR structure determination include the high protein concentrations required for experiments which may cause aggregation, and ambiguities that can arise in the structure from insufficient NMR distance constraints and consequently multiple structures may agree with the data.

Although X-ray crystallography and NMR represent the best means of obtaining high resolution (atomic coordinate) protein structures, low resolution techniques, which

include electron microscopy, small-angle solution scattering and hydrodynamic measurements, can be used to validate high resolution methods, or more significantly for obtaining structural information when high resolution techniques are not applicable. Electron microscopy can be used to directly visualize the whole structure of a protein in semi-crystalline forms or when flattened or stained on a template, but the experimental conditions can potentially perturb the structure. Small-angle scattering measurements are performed in solution so they are advantageous because experimental conditions can be selected and varied, as desired. A major application of solution scattering is in modelling the domain arrangements of multidomain proteins. First, atomic coordinate structures are obtained for individual domains, either directly from crystal or NMR structures, or produced by homology or analogy modelling techniques. The domain orientations are varied to test different interdomain connections and domain interactions and the models are assessed by comparison of their theoretical scattering curves to the experimental data. Hydrodynamic measurements can be used to complement the results from solution scattering studies. Applications of this approach have been reviewed recently (Perkins *et al.*, 1998a, 1998b).

Chapter 2

Protein Structure Determination Methods

2.1. Introduction

The structures of proteins are commonly studied using spectroscopic methods. Spectroscopy is defined as the study of the interaction of electromagnetic radiation with matter, excluding chemical effects (Campbell & Dwek, 1984). The interactions of neutrons and electrons with matter are commonly included in a discussion of spectroscopy because, like electromagnetic radiation, they can be described by a wave function. Different spectroscopic methods provide different information on the structure of a protein, and three examples are now summarized.

Circular dichroism (CD) and Fourier transform infrared (FT-IR) spectroscopy can be used to obtain information on the secondary structure content of a protein. In CD spectroscopy, the effect of an asymmetric molecule (such as a protein) on plane polarized light is investigated (Campbell & Dwek, 1984). A plane polarized wave is produced from two circularly polarized light beams that have equal amplitudes but rotate in opposite directions. Upon interaction with an asymmetric molecule, the amplitude of one beam changes relative to the other to produce an elliptically polarized beam. The degree of ellipticity is measured as a function of wavelength and characteristic values are observed for α-helix and β-sheet conformations at wavelengths between 180 and 250 nm. In FT-IR spectroscopy, the absorption of a protein sample is measured over the infra-red spectrum (Haris & Chapman, 1994). Of particular importance is the frequency range from 1690 to 1600 cm^{-1}, which results mainly from C=O stretching although N-H bending and C-N stretching also contribute. The precise frequencies of maxima or bands are determined by the nature of the hydrogen bonding schemes involving C=O and N-H groups. Thus the secondary structure content of a protein can be determined by analysing the band frequencies in this region. The bands are resolved using mathematical second derivative and deconvolution procedures and it is possible to assign different mainchain conformations according to these bands. Bands between 1620 and 1640 cm^{-1} and between 1648 and 1657 cm^{-1} are usually assigned to β-sheet and α-helix structures respectively. Both CD and FT-IR spectroscopy enable an estimate of the relative proportions of secondary structure types to be obtained, and this is useful in many applications.

Low resolution spectroscopic techniques that provide information on the overall dimensions of a molecule include small angle X-ray and neutron solution scattering methods and electron microscopy. Small angle X-ray and neutron methods examine the scattered waves produced from the diffraction of an incident beam by protein molecules in solution. A detailed account of small angle solution scattering is given in Section 2.2. Electron microscopy uses the same principle as light microscopy and enables the shape of a protein molecule to be visualized directly. In a transmission electron microscope, protein sample is studied using an electron beam that is focussed by a series of lenses to produce magnification factors in the region of ×100,000. Information on the gross structure of a protein molecule can also be obtained from hydrodynamic analyses. The basis of hydrodynamic experiments is that, for a solution subjected to a centrifugal force, the movement of protein molecules is dependent on the size and shape of the molecule. An overview of the application of sedimentation coefficients to protein structure analyses is given in Section 2.2.6.

The experimentally-determined atomic coordinate protein structures that are currently deposited in the PDB have been determined by either X-ray crystallography or multi-dimensional NMR. X-ray crystallography involves the interaction of X-rays with a crystallized molecule and a detailed account of this method is given in Section 2.3. In NMR spectroscopy, the effect of an applied magnetic field on nuclei that possess a magnetic moment or spin is determined. Nuclei that possess spin include ^{1}H, ^{13}C and ^{15}N. When a strong external magnetic field is applied to a protein solution, these spins reach an equilibrium in which the net spin is aligned in the direction of the magnetic field. Radiofrequency pulses are used to apply a second magnetic field perpendicular to the first, which converts the equilibrium alignment to an excited state. Radiofrequencies are emitted by the nuclei as they revert to the equilibrium state, and the exact frequency of radiation emitted from a nucleus depends on its molecular environment. This forms the basis of structure determination by NMR. Two-dimensional NMR experiments are used for structure determination, and different two-dimensional experiments reveal different interactions between hydrogens that are spatially close to each other. The major two-dimensional experiments are the COSY experiment, which gives signals that correspond to hydrogen atoms that are covalently

connected through one or two other atoms (known as J-couplings), and the NOESY experiment, which gives signals that correspond to hydrogen atoms that are close together in space but may be far apart in the sequence (known as nuclear Overhauser effects). The signals from these experiments are assigned to the protein sequence in order to obtain distance constraints from specific hydrogen atoms in one residue to hydrogen atoms in a second residue. A series of models are constructed that are consistent with these distance constraints.

2.2. Small angle solution scattering

In the following section, an overview of small angle X-ray and neutron scattering is given. This describes the basic theory behind the solution scattering experiment, and concludes with some of the useful measurements of macromolecular dimensions that can be obtained from solution scattering data. For a more comprehensive description of the theory of solution scattering, the reader is referred to the texts of Glatter & Kratky (1982) and Perkins (1988).

2.2.1. X-ray scattering theory

Solution scattering is a diffraction technique that can be used to study the overall structure of proteins. An X-ray scattering experiment is performed by irradiating a sample with a highly collimated beam of monochromatic X-rays, and measuring the intensity of scattering I against Q, where Q is a function of the scattering angle 2θ, from which the scattering curve $I(Q)$ is obtained (Figure 2.1). Scattering is due to interaction of X-rays with electrons in the sample: upon irradiation each electron oscillates and emits electromagnetic waves of the same wavelength λ in all directions, but phase-shifted by π with respect to the incident X-ray beam. This type of scattering is known as coherent scattering. The intensity of scattering by an electron is proportional to the X-ray scattering length f of the electron, which has a value of 2.81 fm. For an atom, the X-ray scattering length is the atomic number (the number of electrons it contains) multiplied by the electron scattering length. Accordingly, the hydrogen isotopes, ^{1}H and ^{2}H, both have the scattering length of an electron of 2.81 fm, while the atoms ^{12}C, ^{14}N, ^{16}O and ^{32}S, which are also relevant to proteins, have f values of 16.9 fm, 19.7 fm, 22.5 fm and 45.0 fm respectively.

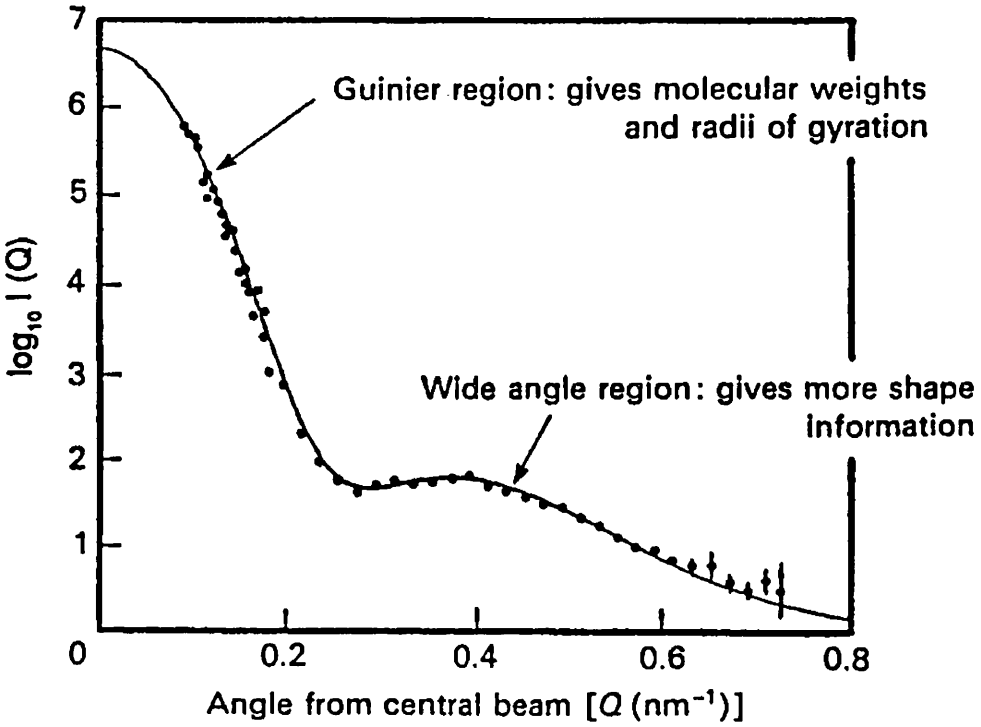


Figure 2.1. General features of a solution scattering curve $I(Q)$ measured over a Q range. The scattering curve is analysed in two regions, that at low Q giving the Guinier plot from which the overall radius of gyration R_G and the forward scattering intensity $I(0)$ values are calculated, and that at larger Q from which more structural information is obtained. At low Q, the scattering curve is truncated for reason of the beamstop (Adapted from Perkins, 1994).

2.2.1.1. The Debye equation

The form of the scattering curve $I(Q)$ is described by the Debye equation. Diffraction phenomena arise from interference between scattered waves and consequently the scattering curve $I(Q)$ is determined by the spatial arrangements of electrons in the protein. X-ray scattering from two points in a protein is depicted in Figure 2.2. The incident X-ray beam is defined by the unit vector s_0 and is scattered from an origin point O in a direction denoted by the unit vector s_1. Since scattering is elastic, s_0 and s_1 have the same amplitude, and for convenience this is set as $2\pi/\lambda$. Q is the scattering vector $(s_1 - s_0)$, and its amplitude is $4\pi\sin\theta/\lambda$ (Figure 2.2). When the scattering angle is zero, waves scattered from all points in the protein will be in phase and the intensity of scattering is the sum of all scatterers. When the scattering angle is non-zero, interference is produced by phase differences between scattered waves, and this can be considered using the path difference between scattering points. In Figure 2.2, the incident X-ray beam is scattered by a second point P, and the path difference between waves scattered by points O and P is $AO + OB$. This path difference corresponds to a phase difference of $2\pi(AO + OB)/\lambda$. If the vector between O and P is r, then $AO = -rs_0$ and $OB = rs_1$ and the phase difference is $r(s_1 - s_0)$, or more simply rQ. The phase difference rQ between each individual scatterer in a macromolecule determines its scattering curve $I(Q)$, and this relationship is contained within the Debye equation:

$$\overline{F^2(Q)} = \sum_p \sum_q f_p f_q \frac{\sin(rQ)}{rQ} \qquad \text{Eq. 2.1.}$$

where $F^2(Q)$ is the square of the structure factor $F(Q)$ of the macromolecule, $F(Q)$ is a regularisation of the scattering curve $I(Q)$ to the intensity of scattering by one electron at zero scattering angle I_e $[I(Q) = I_e F^2(Q)]$, and the summations are performed over all points p and q, where f_p and f_q are the respective X-ray atomic scattering lengths. The Debye equation gives the average of $F^2(Q)$ for all possible orientations of the molecule, and this would therefore require that scattering is spherically symmetrical about the incident beam. It can be seen that the Debye equation is only dependent on the magnitude of Q. For this to be true in a scattering experiment, the scattering sample would have to consist of a monodisperse solution of identical molecules that occupy all

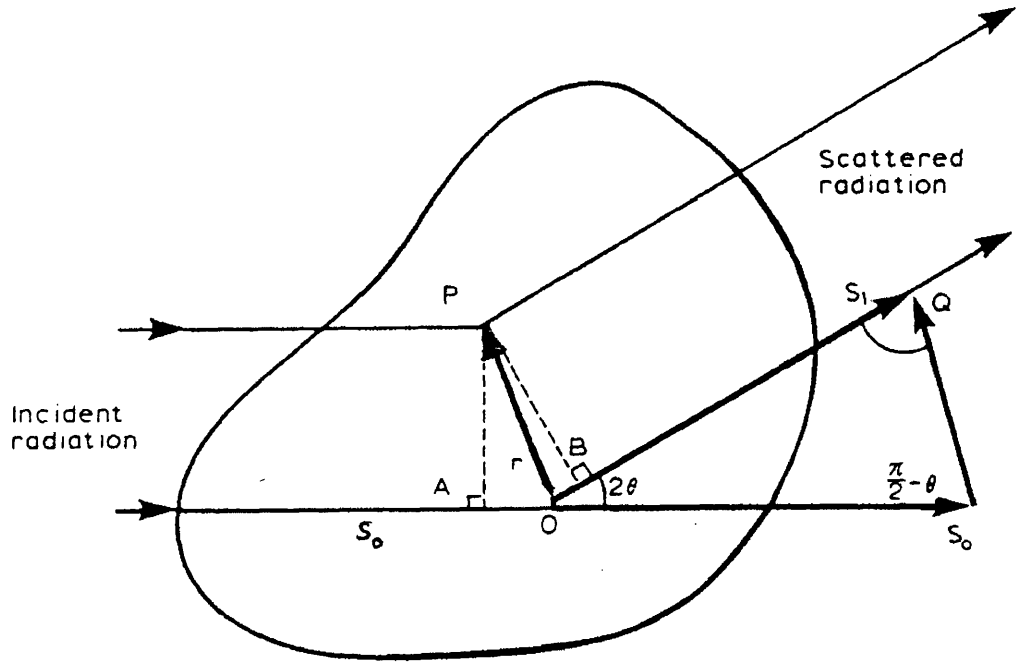


Figure 2.2. Schematic representation of X-ray scattering from two points in a protein molecule. Diffraction is by two points *O* and *P* separated by a distance *r* within a single particle in a solution scattering experiment. *A* and *B* correspond to the perpendiculars to the incident and scattered beams. The unit vectors s_0 and s_1 define the incident and scattered radiation, and *Q* defines the scattering vector $(s_1\text{-}s_0)$ (Adapted from Perkins, 1988).

Note: For solution scattering studies it is convenient to set the amplitudes of the vectors s_0 and s_1 to $2\pi/\lambda$. However, it is more common for the amplitudes of these vectors to be set to $1/\lambda$ when discussing the theory of X-ray crystallography (see Section 2.3).

orientations equally, and act as independent scattering entities. A limitation of the Debye equation is that it describes particles in *vacuo* and it therefore needs to be modified in order to consider aqueous protein solutions.

2.2.1.2. Two-phase model of solution scattering

The simplest consideration of an aqueous protein solution is as a two-phase system of solute and solvent. The Debye equation is modified to consider the contrast $\Delta\rho$ between the scattering density of the macromolecule in vacuo ρ_v and the scattering density of the solvent ρ_s, where $\Delta\rho = \rho_v - \rho_s$. The Debye equation can also be revised according to the resolution limits of the solution scattering experiment. In an X-ray solution scattering experiment, data were typically collected to a maximum usable Q of approximately 2.0 nm^{-1}. The maximum Q value can be used to calculate the structural resolution of scattering, i.e. the smallest distance d measured in the experiment. Bragg's Law ($\lambda = 2d\sin\theta$) is used for this calculation. A Q value of 2.0 nm^{-1} corresponds to a resolution of around 3 nm. At this resolution, a few large volume elements dv_p of scattering density $\rho(r_p)$ can be used to replace individual X-ray atomic scattering lengths f_p in the Debye equation ($f_p = \rho(r_p)dv_p = \rho(r_p)d^3r_p$). Thus the intensity of scattering is:

$$\overline{F^2(Q)} = \int_V\int_V (\rho(r_p) - \rho_s)(\rho(r_q) - \rho_s) \frac{\sin(rQ)}{rQ} d^3r_p d^3r_q \qquad \text{Eq. 2.2.}$$

$$\overline{F^2(Q)} = \int_V\int_V (\rho(r_p) - \rho_s)(\rho(r_q) - \rho_s) \frac{\sin(rQ)}{rQ} d^3r_p d^3r_q$$

where the integration is taken over the macromolecular volume V.

2.2.2. X-ray solution scattering

2.2.2.1. Sample preparation

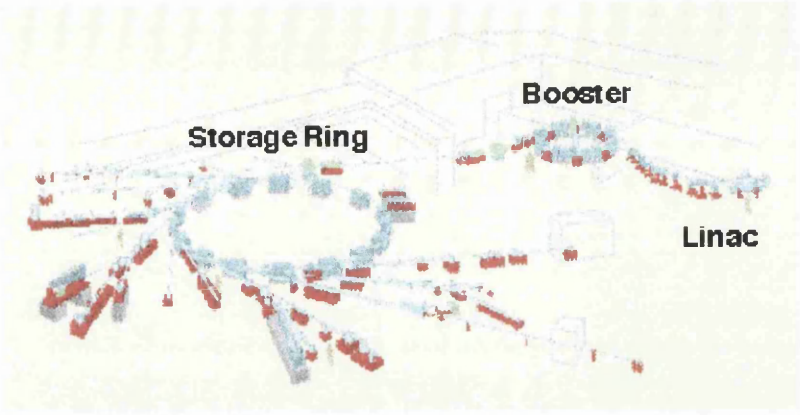
X-ray solution scattering experiments were designed according to the two-phase model of solution scattering expressed in Equation 2.2. The first requirement was for a pure, monodisperse solution of protein at a concentration that was high enough for its scattering curve to be measured. The tendency of protein samples to aggregate necessitated that each sample was subject to gel filtration to remove non-specific aggregates, then reconcentrated as shortly before data collection as possible. The choice of buffer was also important. In X-ray scattering experiments, the closer the buffer is

to pure water, the higher the sample transmission becomes, and hence better counting statistics can be obtained. Phosphate buffered saline (PBS; 12 mM phosphate, 140 mM NaCl, pH 7.4) was used for X-ray scattering experiments. To ensure appropriate corrections were made for solvent scattering, the protein sample was dialysed against its buffer, and the scattering of the buffer was subtracted from the scattering of the protein solution. Ideally, a sample of volume 0.5 ml and concentrated to 10 mg/ml would be used for X-ray experiments. The concentration of a protein solution can be determined from its tryptophan content by measuring its absorbance at 280 nm, and using an absorption coefficient (1%, 1cm) calculated from its amino acid and carbohydrate composition by the corrected Wetlaufer procedure (Perkins, 1986).

2.2.2.2. X-ray scattering at SRS Daresbury

X-ray scattering experiments were performed at the Synchrotron Radiation Source (SRS) at Daresbury, Warrington, U.K. Synchrotron radiation is emitted from electrons that are accelerated while moving at speeds close to the speed of light. Several stages are used in the used in the production of synchrotron X-rays at Daresbury (Figure 2.3a). Electrons are produced by a hot cathode source and then accelerated to almost the speed of light in a linear accelerator (Linac). The energy of the electrons leaving the Linac is increased in a booster synchrotron from 12 to 600 million electron volts (MeV). The electrons are then injected into the storage ring, where a high power radiofrequency accelerating system increases their energy to 2,000 MeV. In the storage ring, 16 dipole magnets force the electrons to follow a circular path, and they travel around the 96 m circumference 3.12 million times a second. The beam current at the start of its "life-time" is typically between 200 and 300 mA, but the current continually decreases as electrons are lost. The electrons are usually kept in orbit for up to 24 hours before the beam has to be regenerated. As the electrons are deflected by the magnetic field, they emit "white" X-rays of all wavelengths down tangential beamlines where experimental stations are set up to use the X-ray beams for dedicated experiments (Figure 2.3a). Small angle X-ray solution scattering experiments were performed on stations 2.1 and 8.2. A perfect Ge or Si crystal is used to horizontally focus and monochromate the X-ray beam to a wavelength of 0.154 nm. The X-ray beam is then vertically focused by a curved mirror, and collimated by sets of slits (Figure 2.3b; Towns-Andrew *et al.*,

(a)



(b)

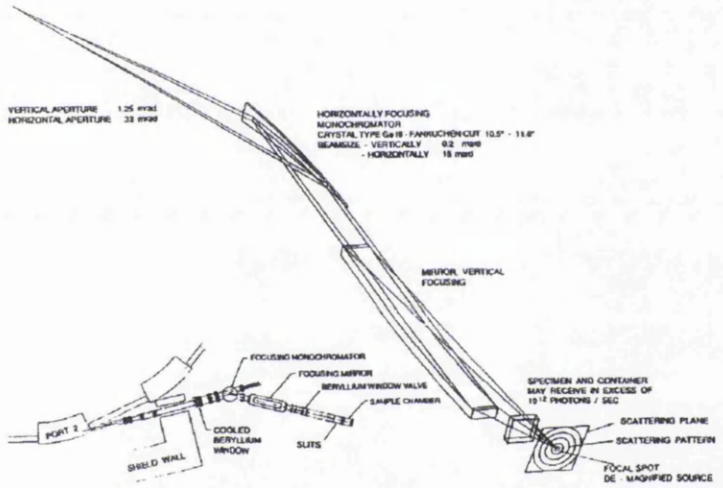


Figure 2.3. X-ray solution scattering at the SRS Daresbury. (*a*) Schematic representation of the synchrotron. The linear accelerator (LINAC), booster synchrotron and the storage ring are shown. Synchrotron radiation is emitted by electrons in the storage ring which is transmitted down beamlines to the different experimental stations. The small-angle scattering instruments are located on beamlines 2 and 8 (Adapted from the Daresbury Laboratory World Wide Web Site *http://www.dl.ac.uk*). (b) Layout of the X-ray solution scattering camera at Station 2.1. This operates at 0.154 nm using a monochromator-mirror optical system. A focal spot of size 0.3 × 2.2 mm^2 is produced, with a beam cross-section of 1 × 5 mm^2 at the sample position. The optics are in vacuum and built on a vibration-isolation system. Between the sample and the detector (not shown) are sections of vacuum tubing of length between 0.5 and 5 m mounted on an optical bench. The scattering pattern is measured using an area detector that is interfaced to a minicomputer. Inset at the lower left is an overall view showing how the X-ray beam is taken from the synchrotron storage ring (Adapted from Perkins, 1994).

1989). This monochromatisation method produces an X-ray beam which has negligible wavelength spread.

Samples were placed in a specially designed perspex cell with 10 to 20 μm thick ruby mica windows. The mica windows were held in place with Teflon plugs, and the windows were regularly changed during data collection sessions. The sample cell had a 1 mm path length, a surface area of 2 mm (vertical) by 8 mm (horizontal), and could hold a maximum volume of 25 μl. The sample cell was held in the X-ray beam by a brass sample holder, which was maintained at 15°C using a water bath. Prior to data collection, the sample holder was aligned in the beam using X-ray sensitive "green paper" that turns red on exposure. The scattering intensities of the sample were measured using a 500-channel quadrant detector (Worgan *et al.*, 1990). A quadrant detector measures intensities in a 70° angular sector of a circle, and gives good counting statistics at large Q values where the intensities are weaker. The nominal position of the main beam is located at the centre of the circle, and a beam stop made from lead was used to protect the detector. The response of the 500 detector channels is not uniform, so for each data collection session the detector response was measured for several hours using a uniform ^{55}Fe radioactive source, and this was used to correct the experimental X-ray scattering measurements. The distance between the detector and the sample was set so that intensities were measured to a maximum Q value of 2.0 to 2.2 nm^{-1}. On station 2.1 sample-to-detector distances of 3.36 m to 3.79 m were used, while on station 8.2 sample-to-detector distances of 3.26 m to 3.54 m were used. Scattering intensities were measured as a function of detector channel number, so for each data collection session the X-ray diffraction pattern of fresh, wet, slightly stretched rat tail collagen was measured for calibrating the Q range of the detector. X-rays produce free radicals that can be destructive to proteins, causing them to aggregate quickly on exposure. To guard against this, the scattering intensity of each sample was measured for 10 minutes in 10 equal time frames, and the time frames were used to check for radiation-damage effects. The protein samples were measured in alternation with their respective buffers in order to minimize buffer subtraction errors as the incident beam decreased in intensity. For each sample, an ion chamber monitor positioned before the sample holder was used for monitoring of the X-ray beam intensity, while measurements from a second ion chamber

monitor positioned after the sample holder automatically allowed for the sample transmission and incident flux in data reduction.

2.2.2.3. Reduction of SRS scattering data

SRS scattering data were reduced to obtain scattering curves *I(Q)*. The SRS scattering data were written as binary files, which contained ten time frames of scattering intensity as a function of detector channel number. Data reduction was performed using the OTOKO software (Figure 2.4; Bendell, P., Bordas, J., Koch, M.H.C., and Mant, G.R., EMBL Hamburg and CLRC Daresbury Laboratory, unpublished software). All scattering data were normalised to the counts measured by the ion chamber positioned after the sample holder using the .DIN procedure. This corrected for beam flux, transmission of the sample and exposure times. Each protein sample spectra had the scattering of the appropriate buffer subtracted from it using .ADD to give the scattering curve of the protein only. The resulting spectra were normalised against the detector response using the .DIV procedure. For each protein sample, the spectra from the ten individual time frames were plotted with .PL3 to determine whether there were any obvious time-dependent effects of the X-rays on the protein. If the protein spectra did not exhibit any radiation-damage effects, the ten time frames were averaged using .AVE, and the averaged intensities were plotted using .PLO. Prior to July 1997, the spectra exhibited a gap in the X-axis, which was due to the timing of the electronics of the detector (Figure 2.5). This was removed using .XSH by specifying the channel number at the beginning of the gap and the size of the gap. From July 1997 onwards, this gap was automatically removed by an improved electronic configuration. The diffraction pattern of collagen was used to calculate the Q values of the detector channels. The collagen diffraction pattern consists of a series of peaks (Figure 2.5). The major peaks are the 1st, 3rd, 5th and 9th order peaks, and the regular spacing between peaks enabled the beam centre (zero order peak) to be detemined. The spacing between successive peaks corresponds to a diffraction spacing d of 67 nm in the collagen fibre, and this enabled the Q values of the detector channels to be calculated ($Q = 2\pi/d$) for merging with the intensities. The .XAX command was used to produce a Q axis file which related the detector channel number to its Q value. The binary OTOKO data files were converted to ASCII text files using .PRT, and DOTKO was

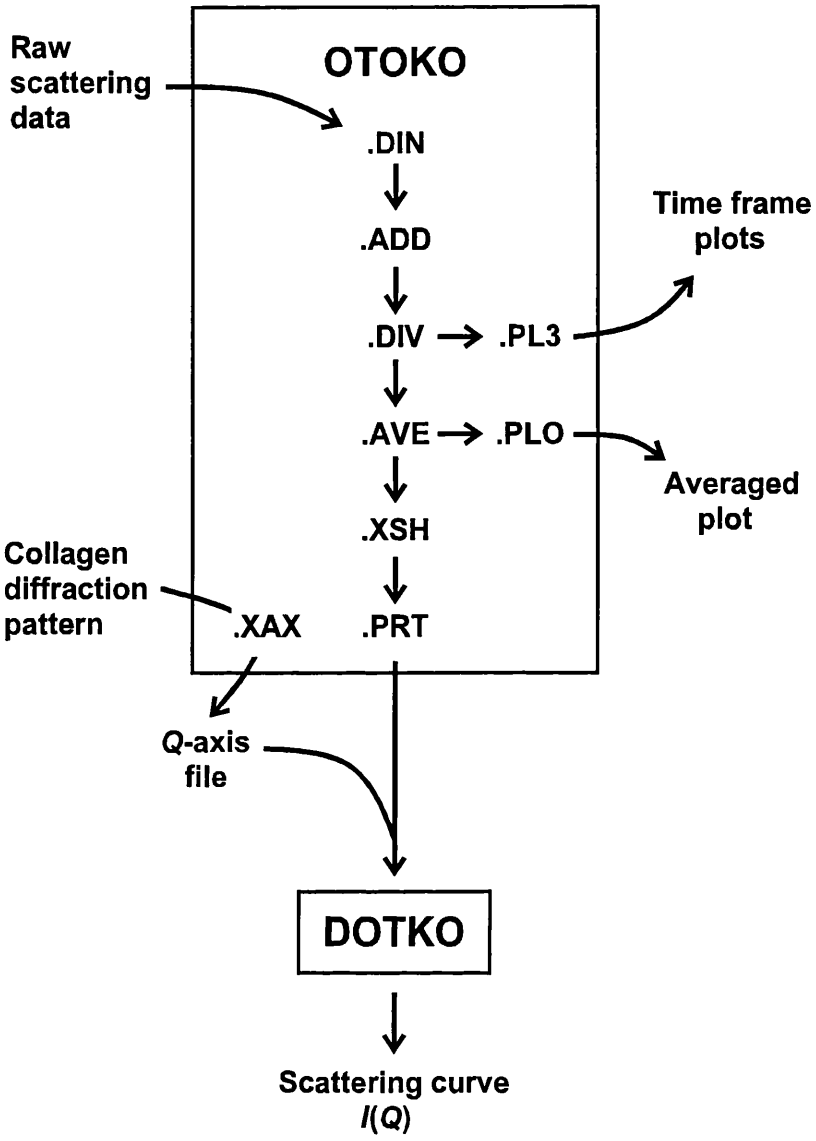


Figure 2.4. Flow diagram of the reduction procedure for SRS Daresbury X-ray scattering data. The majority of data processing was performed using OTOKO. The spectra were normalised against the counts measured at the back ion chamber using .DIN procedure. The buffer background was subtracted from the sample plus buffer spectra using .ADD. The resultant spectra were normalised to the detector response using .DIV and finally the 10 individual time-frames were averaged together with .AVE. The spectra could be plotted either with .PLO (plots out the averaged spectrum) or .PL3 (plots out the individual time-frames). Prior to July 1997, a gap had to be removed from the spectra using the program .XSH. The resulting spectra were converted to ASCII text format using the .PRT command. The Q axis is calculated using .XAX from the diffraction pattern of wet, slightly stretched rat tail collagen. DOTKO combines the Q axis file with the spectrum intensity files for final analyses using SCTPL5.

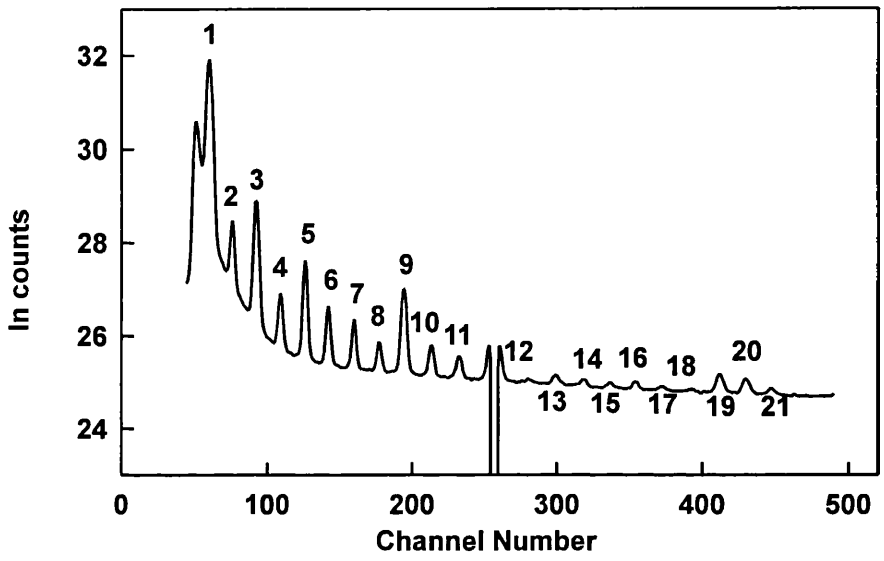


Figure 2.5. The X-ray diffraction pattern of collagen. A total of 21 diffraction peaks are visible here. The spacing between successive peaks corresponds to a real-space diffraction spacing of 67 nm, which is used to calibrate the Q range of X-ray scattering data. The gap between peaks 11 and 12 is an artefact produced by the detector electronics, and is removed during data reduction by determining the position of the beginning of the gap and the size of the gap in channel numbers. This gap did not exist for data collected after June 1997 (Adapted from Meyer, 1994).

used to combine the resulting intensity data files with the Q axis file to produce scattering curve files.

2.2.3. Neutron scattering: a comparison to X-ray scattering

2.2.3.1. Neutrons are scattered by atomic nuclei

The principles of coherent neutron scattering are the same as for coherent X-ray scattering and consequently neutron coherent scattering is also described by the Debye equation (Equation 2.1). However, there are some significant differences between neutron and X-ray scattering. Neutrons are scattered by the nuclei of atoms, unlike X-rays which are scattered by electrons. The neutron atomic scattering length b, which replaces the X-ray scattering length f in the Debye equation, does not exhibit a simple relationship to the atomic number. The atoms ^{12}C, ^{14}N, ^{16}O and ^{32}S, which are important in proteins, have similar b values of 6.651 fm, 9.400 fm, 5.804 fm and 2.847 fm respectively. A similar neutron scattering length is observed for the ^{2}H isotope (6.671 fm), however the ^{1}H isotope has a large negative b value (-3.742 fm). A negative scattering length value implies that there is no phase-shift of the scattered waves relative to the incident wave. The large difference of the neutron scattering of these two hydrogen isotopes is very important in neutron solution scattering experiments. In small angle X-ray scattering experiments, a second type of scattering known as incoherent scattering is negligible, but incoherent neutron scattering exists for nuclei with spin. In terms of protein atoms, incoherent scattering is only significant for ^{1}H nuclei where it becomes very large.

2.2.3.2 The hydration shell

The neutron scattering properties of ^{1}H atoms necessitate that neutron experiments are performed in 100% $^{2}H_2O$ buffers in order to obtain scattering curves that are comparable to those measured in X-ray experiments. However, even under these conditions, small differences are often observed between measurements from neutron and X-ray experiments and these are attributed to the hydration shell that surrounds protein molecules. The hydration shell refers to solvent water molecules that are closely associated with the protein by means of hydrogen bonds. In X-ray solution scattering experiments this hydration shell is detectable around the protein, but it is

invisible in neutron experiments. The model of scattering by an aqueous protein solution that is described by Equation 2.2 assumes that the solute and solvent are discrete entities, but in actuality water molecules hydrogen-bond to the surface of the protein. A water molecule hydrogen-bonded to the surface of a protein is electrostricted and has a smaller volume (0.0245 nm^3) than a water molecule in the bulk solvent (0.0299 nm^3) (Perkins, 1986). Consequently, the X-ray scattering density of a water molecule hydrogen-bonded to the surface of the protein is increased relative to that of a bulk solvent water molecule, so that it has a value similar to the X-ray scattering density of protein. The hydration shell detected by X-rays is typically 0.3 g of H_2O for every gram of protein (Perkins, 1986). In neutron experiments, the hydration shell contains ^1H and ^2H which exchange freely and rapidly with bulk solvent, and consequently there is almost no difference between the scattering density of bulk water molecules and those in the hydration shell, i.e. this means that the hydration shell is not detected by neutron scattering.

2.2.3.3. Contrast difference $\Delta\rho$

In the two-phase model of solution scattering (Equation 2.2) it is assumed that the macromolecule has a uniform scattering density, but in reality biological macromolecules can exhibit distinct regions of different scattering density. The four major classes of biological macromolecules, protein, carbohydrate, lipid and nucleic acids, all exhibit distinct X-ray and neutron scattering densities. For complex macromolecules, which consist of more than one class of macromolecule, e.g glycoproteins, lipoproteins and ribosomes, it is possible to alter the contrast of the solution scattering experiment in order to study the component macromolecules individually. Methods for altering the contrast include adding sucrose to X-ray scattering buffers, chemical labelling of the solute and *in vivo* deuteration of non-labile solute hydrogen atoms (Worcester, 1988), but the most convenient method is to alter the scattering density of neutron buffers using different ratios of ^2H_2O to ^1H_2O. The work presented in this thesis did not involve contrast variation experiments as such. Instead complementary studies were performed in a high positive solute-solvent contrast $\Delta\rho$ by X-ray scattering, and in a high negative solute-solvent contrast by neutron scattering in 100% ^2H_2O buffers. The combination of these approaches provided controls for the

presence of significant inhomogeneity within the glycoprotein protein and carbohydrate scattering.

2.2.4. Neutron solution scattering

2.2.4.1 Sample preparation

Samples for neutron scattering experiments were prepared in the same way as X-ray samples, except that PBS buffer in 100% ^2H_2O was used. After gel filtration and reconcentration, each sample was dialysed with stirring against this ^2H_2O PBS buffer for a minimum of 36 hours, in which the buffer was changed four times to ensure full ^1H-^2H exchange.

2.2.4.2. Neutron scattering on LOQ at the RAL

Neutron scattering experiments were performed on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory (RAL), Didcot, UK (Figure 2.6a) (Heenan and King, 1993). At ISIS, neutrons are produced by a spallation process. An ion source produces H^- ions, which are initially accelerated to 665 keV. In the second stage, the H^- ions pass through a linear accelerator (Linac), where they reach an energy of 70 MeV. Protons are then produced by stripping the electrons from the H^- ions using a very thin alumina foil. Pulses of protons are accelerated to 800 MeV in a 52 m diameter synchrotron, extracted and then sent to the heavy metal target station, at a rate of 50 times a second (50 Hz). Bombardment of the uranium or tantalum heavy metal target with high energy protons produces neutrons by chipping these from its nuclei. The neutrons are slowed by moderators to increase wavelengths. On LOQ the moderator is liquid hydrogen (25 K). The layout of the LOQ instrument is shown in Figure 2.6b. A supermirror bender removes neutrons with wavelengths less than 0.2 nm, while a frame overlap mirror removes neutrons with wavelength greater than 1.2 nm. Every other neutron pulse is cut out by a chopper, producing an operational frequency of 25 Hz, and the phase of the chopper is set so that wavelengths in the range 0.22 to 1.00 nm are used in scattering experiments. The neutron beam is collimated by two apertures and the beam used in solution scattering experiments had a diameter of 8 mm. Monochromatization of the neutron beam is achieved by time-of-flight techniques, based on a distance of 15.15 m from the heavy metal target to the detector.

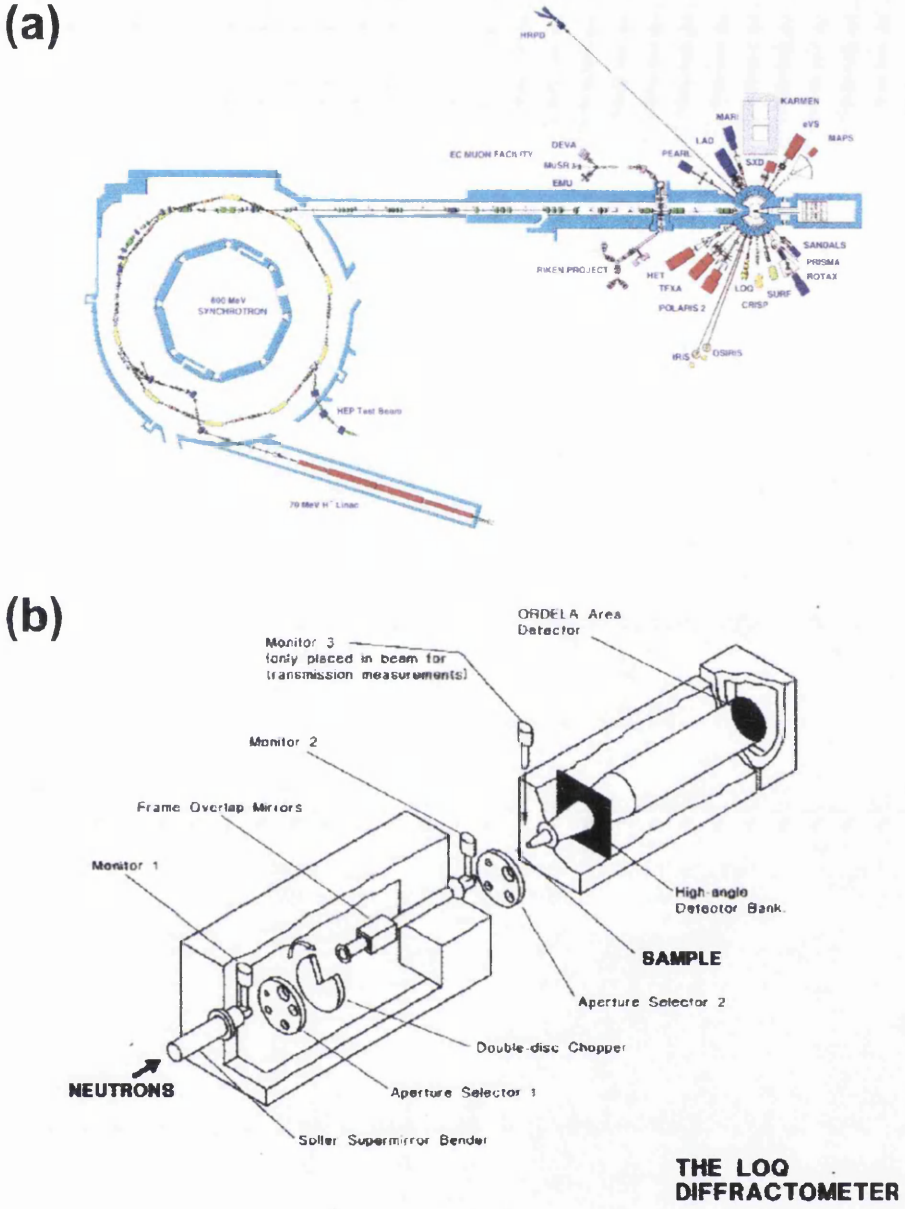


Figure 2.6. Neutron solution scattering on LOQ. (*a*) Layout of the LINAC, synchrotron and target station at ISIS located at the Rutherford-Appleton Laboratory, Didcot. Pulses of protons are accelerated to 800 MeV in the synchrotron, extracted and then sent to the heavy metal target station. As the high energy protons hit the uranium or tantalum target, fragments are chipped from its nuclei, producing neutrons. Neutron pulses are used at the experimental stations which emanate from the heavy metal target. (*b*) Schematic diagram of the LOQ diffractometer. A chopper is set so that wavelengths in the range 0.22 to 1.00 nm are used in scattering experiments. The neutron beam is collimated by two apertures, and the beam used in solution scattering experiments had a diameter of 8 mm. Monochromatization of the neutron beam is achieved by time-of-flight techniques. Scattering intensities were measured using a ^{3}He ORDELA detector. (Taken from the Rutherford-Appleton Laboratory Web Site *http://www.rl.ac.uk*).

The time-of-flight techniques are advantageous because they enable neutron scattering data to be measured over a Q range of 0.08 to 2.5 nm^{-1} in one experiment, and utilize fully the abundance of shorter wavelength neutrons, which contribute most to the scattered curve at large Q where the intensities are the lowest.

Samples were placed in 2 mm thick rectangular quartz Hellma cells, which were held in a motorised rack at a constant temperature of 15°C. Neutron scattering intensities were measured using a ^{3}He ORDELA detector, which had an active area of 64 cm × 64 cm, and was positioned at a fixed distance of 4.3 m from the sample. The sample rack held multiple samples, enabling the data collection procedure to be automated. Scattering intensities were measured for proteins and their corresponding buffers. The scattering from a partially deuterated polystyrene standard was also measured, and this was used to normalise the spectral intensities (Wignall & Bates, 1987). Neutron transmissions were measured for each protein sample and buffer, and for the polymer standard and an empty beam position, and these were also used to normalise the scattering data.

2.2.4.3. Reduction of LOQ scattering data

Raw data were collected in 100 time frames of 64 × 64 cells, and reduced using the standard ISIS software package COLETTE (Figure 2.7; Heenan *et al.*, 1989). The MASK file is routinely updated by the LOQ instrument scientists, and controls essential information on the properties of the detector. After ensuring that the beam size, the sample thickness, the moderator-to-sample distance and sample-to-detector distance had been assigned correctly, the raw scattering data were processed. The data were corrected for detector efficiency, transmission of the direct beam, and transmission of the sample. Absolute scattering intensities were calculated by normalisation of the spectra against a "LOQ standard sample" of partially deuterated polystyrene (Wignall & Bates, 1987), and protein spectra were corrected by buffer subtraction. The time-of-flight techniques were used to bin the scattering intensities into individual diffraction patterns, either for linear wavelength steps of 0.02 nm, or for logarithmic wavelength steps of 0.08%, over the wavelength range of 0.22 to 1.00 nm. These individual diffraction patterns were merged, using 0.04% logarithmic binning increments for Q, to give the full scattering

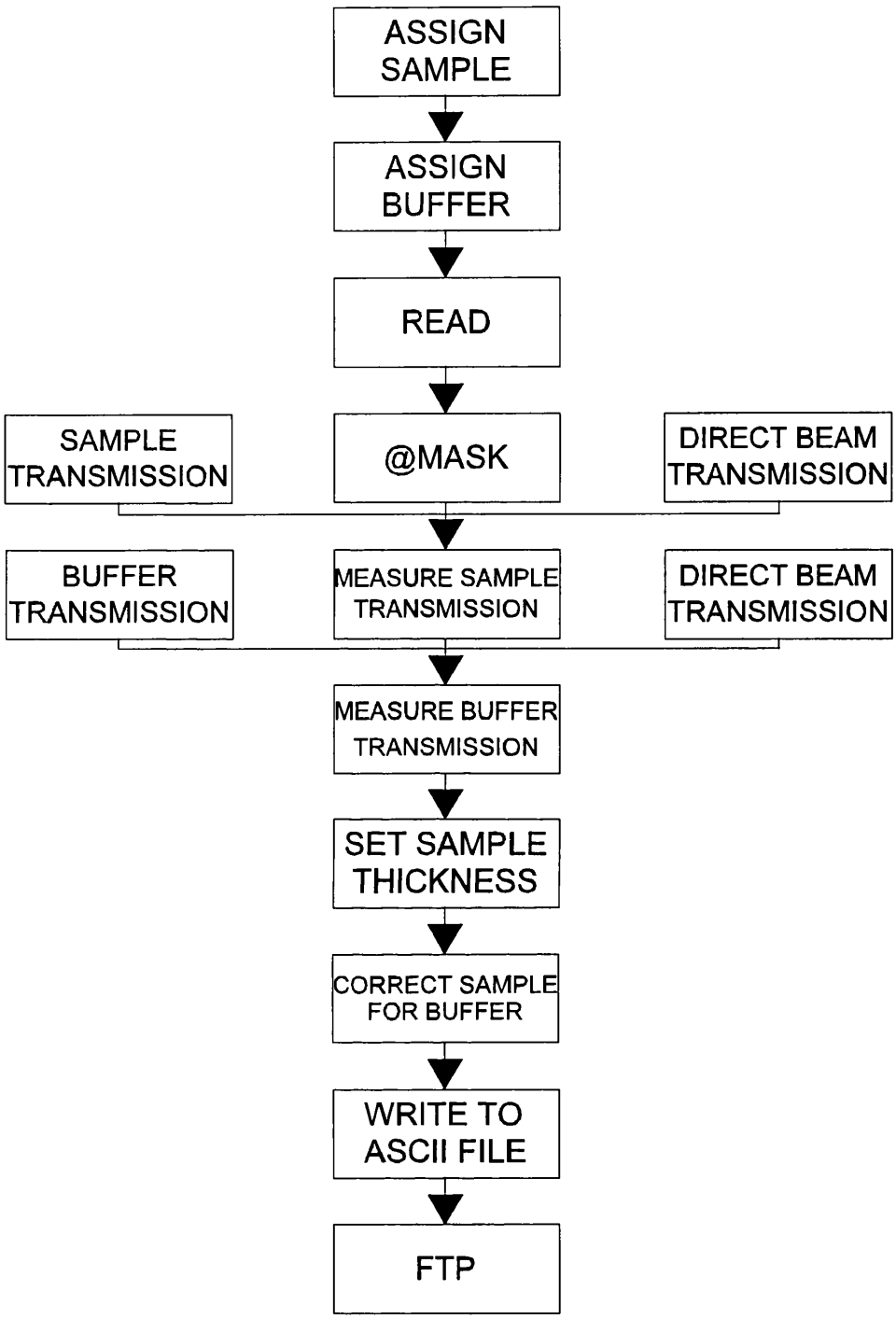


Figure 2.7. Flow diagram of LOQ data reduction using COLETTE. The diagram follows the stages of data reduction from the raw data files to the final transfer of the scattering curve spectra as ASCII text files to London. @MASK executes a file (mask.com), which is updated by the instrument scientists to account for fluctuations in the behaviour of the detector and changes in instrument configuration (Adapted from Meyer, 1994).

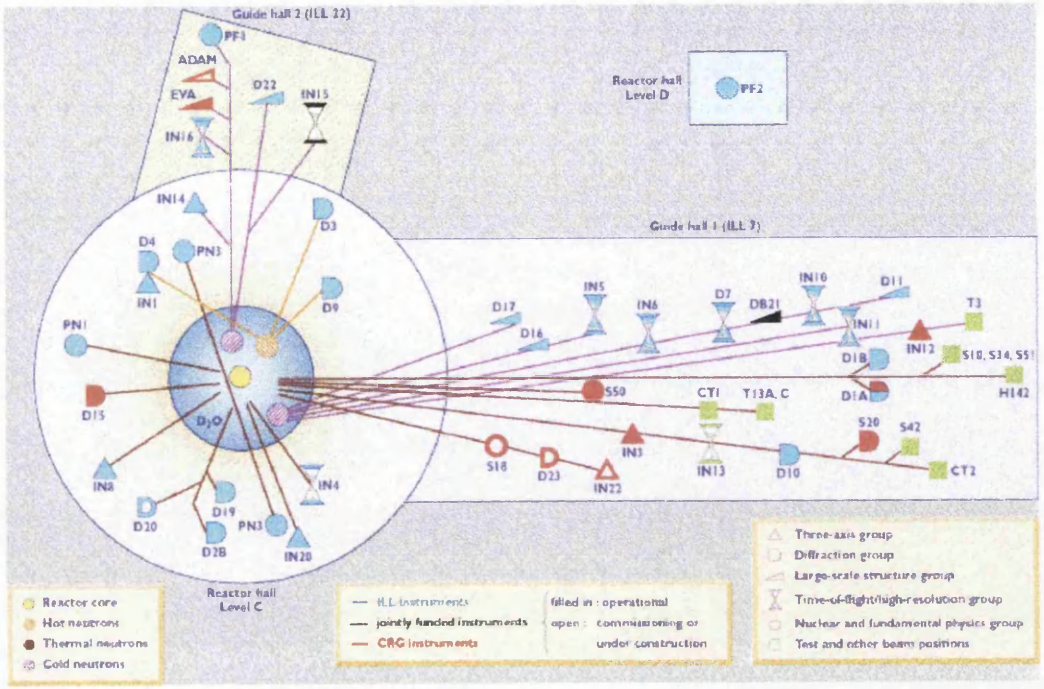
curve $I(Q)$ in an approximate Q range of 0.08 to 2.5 nm^{-1}. Logarithmic steps were optimal for analyses of the Guinier and wide angle regions of the scattering curves.

2.2.4.4. Neutron scattering on D22 at the ILL

Neutron scattering experiments were also performed using the high-flux reactor at the Institute Laue-Langevin (ILL) in Grenoble, France (Figure 2.8a). High energy neutrons are produced in the reactor core by the fission of U^{235}. The thermal neutron flux from the reactor core is in equilibrium with a ^2H_2O moderator (300K), and has a distribution of neutron wavelengths that peaks at 0.12 nm. Neutron flux wavelengths greater than 0.3 nm are enhanced by cold liquid ^2H moderators (25K). Instruments are set up to perform specific experiments using neutron beams as shown in Figure 2.8a. Small angle scattering experiments were performed using a large wavelength neutron beam from a cold liquid ^2H moderator on Instrument D22 (Figure 2.8b). On D22, the incident beam is monochromatized using a rotating drum velocity selector. The velocity selector has a helical slot, through which only neutrons of a specified wavelength can pass. A wavelength of 1 nm was selected, but the velocity selector typically produces a spread $\Delta\lambda/\lambda$ of 10% around the selected wavelength. In order to maximize the incident flux, the neutron beam was collimated using movable sections of guide tube, which had a rectangular beam aperture of 7 × 10 mm.

Samples were placed in 2 mm thick rectangular quartz Hellma cells, which were held in a remotely positioned rack at a constant temperature of 15°C. Scattering intensities were measured using a ^3He detector which had an active area of 96 cm × 96 cm, composed of 16,000 cells of size 0.5 cm × 0.5 cm. In order to measure the whole scattering curve $I(Q)$, intensities were measured for each protein and buffer at a sample-to-detector distance of 5.6 m to obtain intensities in the Q range between 0.07 and 0.76 nm^{-1}, and a sample-to-detector distance of 1.4 m to obtain intensities in the Q range between 0.28 and 2.80 nm^{-1}. Data acquisition times were typically 6 minutes for samples in ^2H_2O buffers. The background neutron counts were determined by blocking the neutron beam with a cadmium sample. In order to convert the scattering intensities to an absolute scale, the incoherent scattering of H_2O and the scattering of an empty sample cell were also measured.

(a)



(b)

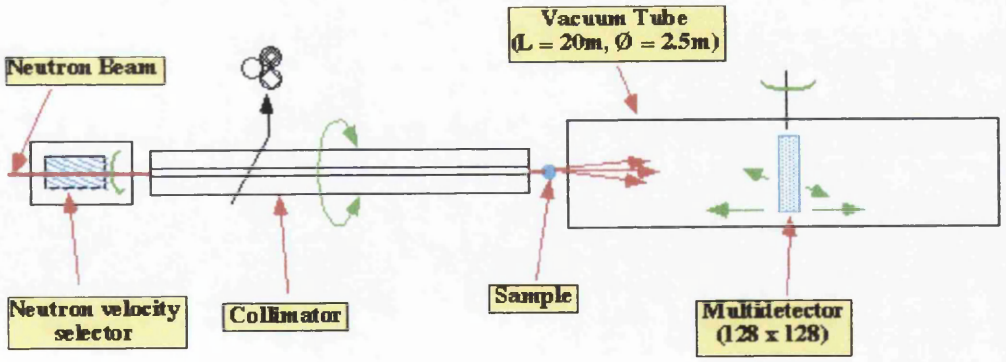


Figure 2.8. Neutron solution scattering at the ILL, Grenoble. (*a*) Layout of beam-tubes and instruments at the High Flux Reactor, ILL. High energy neutrons are produced by the fission of U^{235}. Neutron guides transfer the neutrons from the reactor core to the external instruments. (*b*) Schematic representation of station D22, on which small-angle solution scattering experiments were performed. (Taken from the ILL Web Site *http://www.ill.fr*).

2.2.4.5. Reduction of D22 scattering data

A number of calibration and normalisation measurements were performed during the processing of any ILL data run (Figure 2.9; Ghosh, 1989). The raw counts for each detector cell were listed using DETEC, from which the position of the beam stop and the beam centre was calculated. RNILS was used to calculate Q values for the detector cells, and the intensity I was calculated as a function of Q by taking the average values for detector cells over a given radial step length spacing of 1 cm. SPOLLY was used to combine the individual spectra for sample, buffer, water, empty cell and cadmium to obtain the protein scattering curves $I(Q)$. All spectra were normalised against the number of monitor counts. In SPOLLY, spectra were corrected for their neutron transmissions, the scattering intensities were converted to an absolute scale using the incoherent scattering of H_2O and an empty sample cell, and the scattering of the buffer was subtracted from each protein spectrum. Preliminary Guinier analyses (Section 2.2.5.1) were performed using RGUIM to ensure that samples were measured for long enough during data acquisition. RPLOT was used to plot spectra measured from the two sample-to-detector distances to check that their intensities overlapped sufficiently in the two Q ranges during measurements. RCARD produced the scattering curves $I(Q)$ in ASCII text format for further analysis in London. The two scattering curves for each protein sample were combined to obtain the full scattering curve $I(Q)$.

2.2.5. Analyses of reduced scattering curves $I(Q)$

Measurements of the gross structural details of a protein can be determined from its X-ray and neutron scattering curves.

2.2.5.1 Guinier analyses

At low Q values, the Debye equation (Equation 2.1) is reduced to a Gaussian curve, which becomes the Guinier approximation:

$$\ln I(Q) = \ln I(0) - \frac{R_G^2 Q^2}{3} \qquad \text{Eq. 2.3.}$$

where R_G is the radius of gyration and is defined as the root-mean-square distance of all scattering centres in the macromolecule from its centre of gravity, and $I(0)$ is the

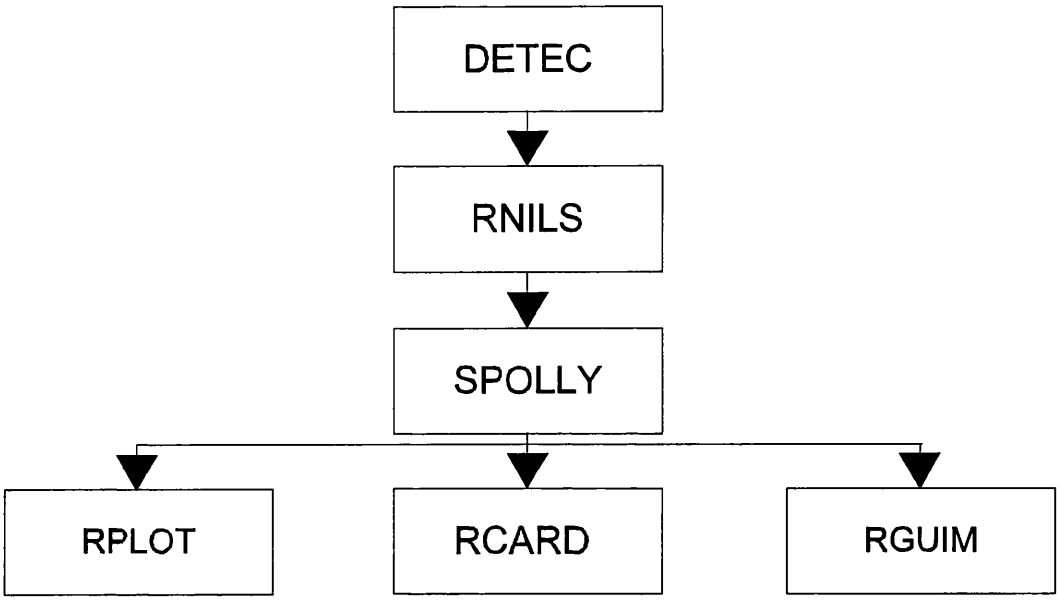


Figure 2.9. Flow diagram of D22 data reduction procedures. DETEC lists the raw counts from the detector cell by cell, RNILS lists and stores the radial distribution function $I(Q)$ of the detector, SPOLLY corrects and normalizes the sample spectrum $I(Q)$, RPLOT plots out 2 spectra on the same axes, RCARD produces a card image disk file of formatted data and RGUIM is used to calculate the Guinier and cross-sectional parameters (Adapted from Meyer, 1994).

intensity at zero scattering angle. This expression is valid in a $Q.R_G$ range up to 0.7-1.3 depending on the macromolecular shape. R_G and $I(0)$ are determined from an experimental scattering curve by plotting ln $I(Q)$ against Q^2 to give a straight line of slope $-R_G^2/3$ and intercept ln $I(0)$. R_G is a measure of elongation if the internal inhomogeneity is negligible, and a useful application of R_G is in the calculation of the anisotropy ratio of a macromolecule. The anisotropy is R_G/R_0, where R_0 is the R_G of the sphere of volume equal to that of the macromolecule (the R_G of a sphere = $\sqrt{(3/5)r^2}$; where r is the radius of the sphere). For typical globular proteins, the anisotropy ratio is approximately 1.28 (Perkins, 1988). The intensity at zero scattering angle $I(0)$ is proportional to M_r^2. The scattering intensities of neutron data measured on LOQ are normalised relative to a partially deuterated polystyrene standard, and consequently LOQ $I(0)/c$ values (c = sample concentration) are used to calculate protein M_r values. There is a linear relationship between the M_r determined from composition analyses, and $I(0)/c$ (Figure 2.10), which was used to determine the M_r and hence the oligomeric state of a newly-measured protein. The neutron data measured on D22 are normalised relative to the incoherent scattering of H_2O, and the protein M_r values can be calculated by reference to the instrument geometry (Jacrot & Zaccai, 1981).

Guinier plots were performed using the FORTRAN program SCTPL5 (A. S. Nealis & S. J. Perkins, unpublished software). The Q range selected for line-fitting went from the smallest possible Q to a maximum $Q.R_G$ of approximately 1.5. Nonspecifically aggregated proteins were identified by steeply curved Guinier plots at low Q values and were discarded.

2.2.5.2. Cross-sectional radius of gyration

If the protein has an elongated structure, Guinier-type analyses of the scattering curve at larger Q will give the mean radius of gyration of the cross-section R_{XS} and the mean cross-sectional intensity at zero scattering angle $[I(Q)Q]_{Q\to 0}$ (Hjelm, 1985):

$$\ln[I(Q)Q] = [\ln(I(Q)Q)]_{Q\to 0} - \frac{R_{XS}^2 Q^2}{2} \qquad \text{Eq. 2.4.}$$

R_{XS} and $[I(Q)Q]_{Q\to 0}$ are determined from an experimental scattering curve by plotting

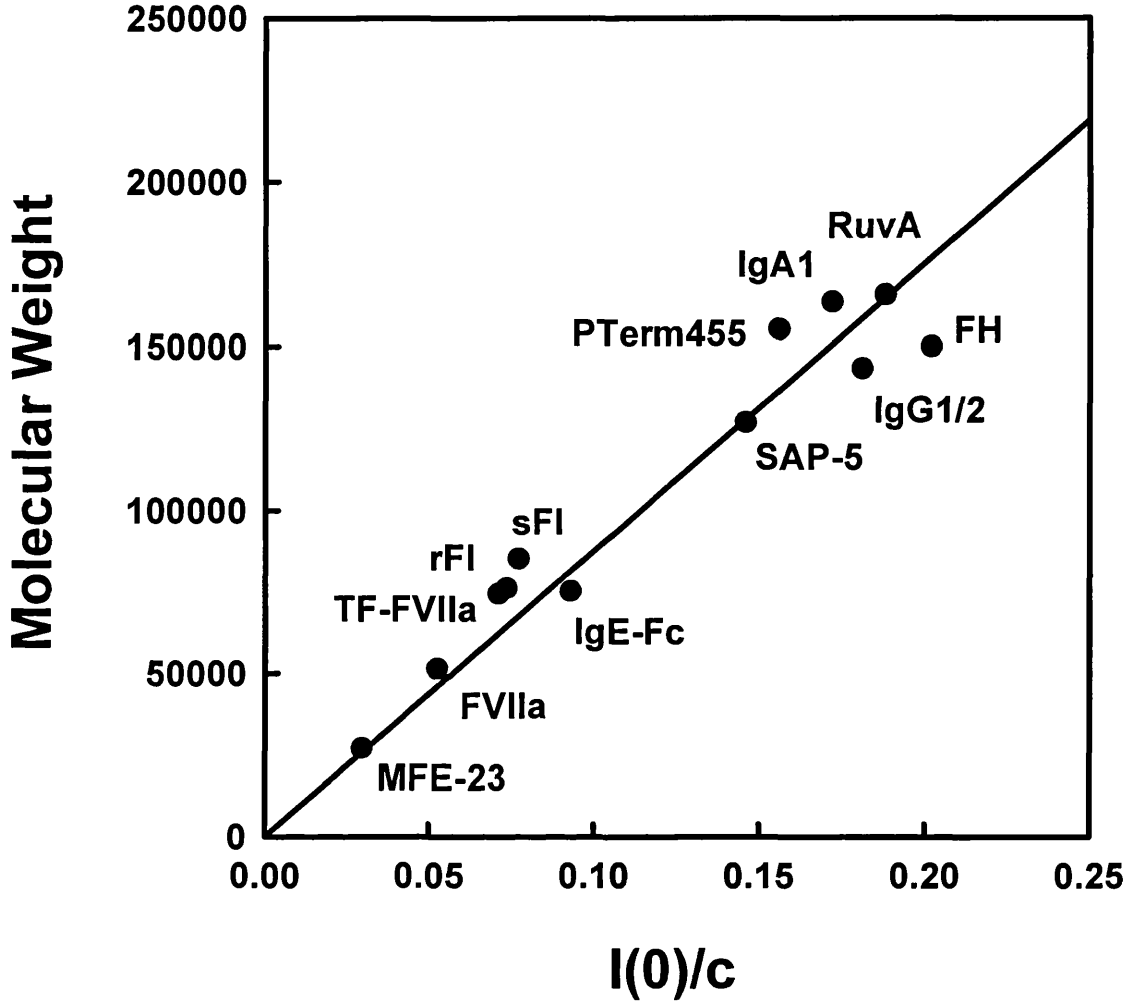


Figure 2.10. Linear relationship between the molecular weight and the neutron $I(0)/c$ values for glycoproteins in 100% ^2H_2O buffer measured on LOQ. In order of increasing molecular weight, the data correspond to the following proteins: MFE-23, factor VIIa, the factor VIIa-tissue factor complex, the IgE-Fc fragment, recombinant and serum factor I, pentameric serum amyloid P component, bovine IgG1/2, factor H, PTerm455, serum IgA1 and octameric RuvA (Perkins *et al.*, 1998a, 1998b; Chamberlain *et al.*, 1998).

$\ln[I(Q)Q]$ against Q^2 to give a straight line of slope $-R_{XS}^2/2$ and intercept $[\ln(I(Q)Q)]_{Q\to 0}$. For "T-shaped" protein structures such as the immunoglobulins, the cross-sectional plot exhibits two linear regions, a steeper innermost one and a flatter outermost one (Pilz *et al.*, 1973), and in this type of analysis the two regions are referred to as $R_{XS\text{-}1}$ and $R_{XS\text{-}2}$ respectively.

Cross-sectional plots were performed using SCTPL5. The Q range selected for line-fitting went from a minimum Q value which was greater than the Q values used in the Guinier fits to a maximum $Q.R_{XS}$ of 1.5. If the cross-sectional plot exhibited two R_{XS} regions, the Q range selected for calculating $R_{XS\text{-}2}$ went from a minimum Q value that was greater than the Q values used in the $R_{XS\text{-}1}$ fit to a maximum $Q.R_{XS\text{-}2}$ of 1.5.

2.2.5.3. Estimations of macromolecular dimensions

Assuming a protein is shaped like an elongated elliptical cylinder, the R_G and R_{XS} analyses can be combined to determine its length L (Glatter & Kratky, 1982):

$$L = \sqrt{12(R_G^2 - R_{XS}^2)} \qquad \text{Eq. 2.5.}$$

Alternatively, L is given by (Perkins *et al.*, 1986):

$$L = \frac{\pi I(0)}{[I(Q)Q]_{Q\to 0}} \qquad \text{Eq. 2.6.}$$

The consistency of these two calculations of L is a useful control of a correct cross-sectional analysis for a rod-like particle. For an assumed elliptical cylinder with length L, the two shorter dimensions A and B can be calculated from the unhydrated protein volume for neutron analyses, or from the hydrated protein volume for X-ray analyses:

$$V = \pi ABL \; ; \; R_{XS} = \sqrt{(A^2 + B^2) / 4} \qquad \text{Eq. 2.7.}$$

2.2.5.4. Real space distance distribution function

The scattering curve $I(Q)$ represents the macromolecular structure in reciprocal

space. The experimental $I(Q)$ curve can be converted into real space by a Fourier transform over $0 \leq Q \leq \infty$:

$$P(r) = \frac{1}{2\pi^2} \int_0^\infty I(Q)Qr \sin(Qr)dQ \qquad \text{Eq. 2.8.}$$

where $P(r)$ is the distance distribution function and corresponds to the number of distances r between any two volume elements within the macromolecule weighted by the product of their respective scattering densities. The maximum in $P(r)$ is the most frequently occurring intramolecular distance. $P(r)$ offers an alternative calculation of R_G:

$$R_G^2 = \frac{\int_0^\infty P(r)r^2dr}{2\int_0^\infty P(r)dr} \qquad \text{Eq. 2.9}$$

$P(r)$ termination errors occur because the experimental $I(Q)$ curve cannot be measured at zero angle or at very large Q, and in addition high signal to noise ratios are associated with measurements at large Q. The Indirect Transform Procedure (ITP) minimizes these termination errors, and was performed using the ITP-91 program (Glatter & Kratky, 1982). Another program was GNOM which is more automated and easier to use (Svergun *et al.*, 1988; Semenyuk & Svergun, 1991; Svergun, 1992). D_{max} is the maximum macromolecular dimension, and therefore corresponds to the length L. In GNOM, the $P(r)$ curve was calculated from $I(Q)$ for a range of estimated D_{max} values, and the $P(r)$ curve was selected according to several criteria: (1) $P(r)$ should exhibit positive values; (2) the R_G from ITP and Guinier analyses should agree; (3) $P(r)$ should be zero when r is zero; (4) $P(r)$ should be stable and reproducible for different experimental $I(Q)$ curves when D_{max} is varied over a reasonable range. The length L was determined from $P(r)$ when this became zero at large r, however errors in L can be significant as a result of the low intensity of $P(r)$ in this region.

2.2.6. Hydrodynamic analyses

Hydrodynamic analyses can be used to relate the effect of a centrifugal force on a protein in solution to the size and shape of the protein. Therefore the sedimentation coefficient $s^{o}_{20,w}$ of a protein provides structural information that is complementary to measurements determined by solution scattering (Perkins, 1988). The sedimentation coefficient is determined experimentally by analytical centrifugation of the glycoprotein in aqueous solution, by measuring the movement of sedimenting particles in a given time:

$$s^{\circ}_{20,w} = \frac{1}{t\omega^2} \ln \frac{r}{r_0} \qquad \text{Eq. 2.10}$$

where t is time; ω is the angular velocity of the rotor in radians per second; $r - r_0$ is the distance travelled by the particles. Published values of $s^{o}_{20,w}$ were used when available. If literature $s^{o}_{20,w}$ values were not available or if they were ambiguous, the $s^{o}_{20,w}$ was measured on a Beckman XL-A analytical centrifuge by Dr O. Byron at the National Centre for Macromolecular Hydrodynamics, Leicester.

The sedimentation coefficient results from three forces acting on a sedimenting particle. The centrifugal force is the product of the mass of the particle and the centrifugal acceleration of the particle and causes the particle to sediment. The centrifugal force is opposed by a buoyancy force that results from the displacement of solvent molecules by the particle. The third force is the frictional force of the solvent acting on the particle, and it is the product of the frictional coefficient and the sedimentation velocity. In terms of these three forces, the sedimentation coefficient is expressed as:

$$s^{\circ}_{20,w} = \frac{M_r(1 - \bar{v}\rho_{20,w})}{N_a f} \qquad \text{Eq. 2.11.}$$

where M_r is the molecular weight of the macromolecule; $\bar{v}$ is the partial specific volume of the particle; $\rho_{20,w}$ is the density of water at 20°C; N_a is Avogadro's constant (6.023×10^{23}); and f is the frictional coefficient. The sedimentation coefficient of a glycoprotein

is useful as this leads to the calculation of its frictional coefficient f which can then be compared with f calculated for a molecular model. M_r and $\bar{v}$ values can be calculated from the molecular composition using SLUV (Section 3.2.1; Perkins, 1986), and $\rho_{20,w}$ has an assumed value of 0.9982 ml/g. The frictional coefficient can then be used to calculate the frictional ratio f/f_o of the glycoprotein. f_o is the frictional coefficient of the sphere with the same volume as the glycoprotein, which is given by Stokes law:

$$f_o = 6\pi\eta r_s \qquad \text{Eq. 2.12.}$$

where η is the viscosity of the solvent, which is assumed to be 0.001002 N s m^{-2} for aqueous protein solutions; and r_s is the radius of the sphere. The frictional ratio is a measure of the elongation of the glycoprotein, and it is analogous to the solution scattering anisotropy ratio R_G/R_O.

2.3. X-ray crystallography

The following section gives a short outline of the X-ray crystallographic method with respect to obtaining electron density maps for a protein structure by means of molecular replacement. For a more thorough description, the reader is referred to one of the many texts on this broad subject, e.g. Glusker & Trueblood (1985), Rhodes (1993) and Drenth (1994).

2.3.1. Crystals

2.3.1.1. Crystal growth

A protein crystal is grown by slow, controlled precipitation from aqueous solution under conditions that do not denature the protein. In the most common methods for growing protein crystals, the protein is dissolved in an aqueous buffer containing a precipitant at a concentration just below that necessary to induce precipitation of the protein. Water is slowly removed from the solution by controlled evaporation, and when the solution reaches supersaturation small regular protein aggregates are formed that act as the nuclei for crystal growth. The initial crystallization of a protein is principally a trial-and-error process, in which numerous conditions are tested. It is

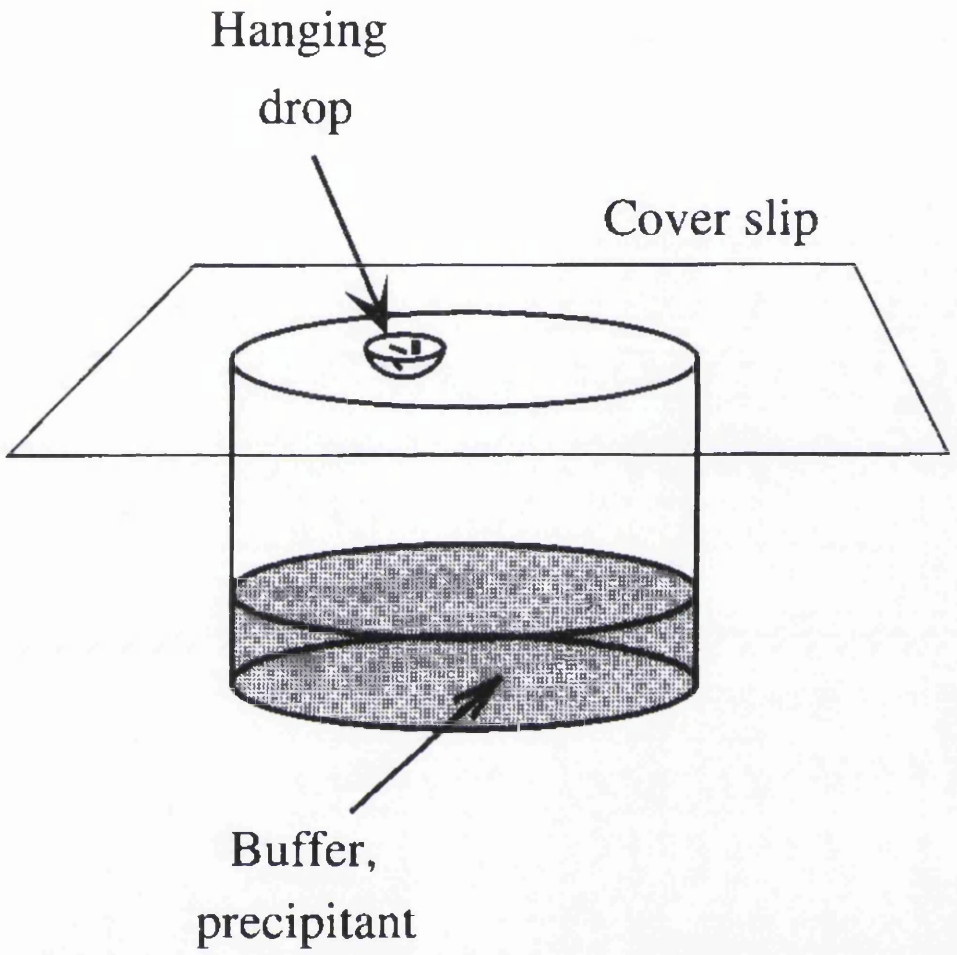


Figure 2.11. The hanging-drop vapour diffusion method. The droplet hanging under the coverslip contains a buffered protein solution mixed with precipitant at approximately half of the concentration required for crystal growth. As water evaporates from the droplet, its precipitant concentration increases and crystallization occurs in the droplet.

common to vary the protein concentration, the nature and concentration of the precipitant, pH and temperature in crystallization trials. MFE-23 (Chapter 6) was crystallized by Mr Jeremy Thornton from the Clinical Oncology Department at the Royal Free and University College Medical School and Dr Maninder Sohi from Dr Brian Sutton's group at the Randall Institute, Kings College, London. This crystallization utilized the hanging-drop vapour diffusion method (Figure 2.11). In this method, up to 25 μl of purified protein solution is mixed with an equal volume of precipitant solution, resulting in a precipitant concentration that is approximately 50% of that required for crystallization. This solution is suspended as a droplet from the underside of a microscope cover slip. The cover slip is placed on top of a reservoir containing approximately 0.5 to 1.0 ml of the precipitant solution, and an air-tight seal is formed using silicone grease. In this system, the major component is the reservoir of precipitant. Vapour diffusion results in a net transfer of water from the protein solution to the reservoir, and this proceeds until their precipitant concentrations are the same. At this equilibrium, the protein solution is maintained at the optimal precipitant concentration and thereby promotes crystal growth.

2.3.1.2. Basic crystal structure theory

A crystal is a regular, three-dimensional arrangement of molecules. It is constructed from the repetitive translation in three dimensions of a basic structural motif containing one or several protein molecules. The imaginary parallelepiped that contains one unit of the repeating structural pattern is termed the unit cell. A unit cell is defined by three vectors **a**, **b** and **c**, that correspond to the three edges of the unit cell and have lengths *a*, *b*, and *c* respectively, and the three unique angles between them are designated α, β, and γ. A crystal corresponds to a three-dimensional stack of unit cells and, for simplicity, each unit cell can be represented by a point. These points have a regular arrangement; a three-dimensional array or lattice. The crystal lattice is an important concept because it enables the interpretation of several crystal properties without the necessity of considering the content of the unit cell.

In the crystal lattice, the line in the **a** direction is termed the x-axis, the **b** direction is the y-axis, and the **c** direction is z-axis. Imaginary planes can be constructed

to connect the points in the lattice (Figure 2.12). As will be seen in Section 2.3.2.1, these lattice planes are important because diffracted X-ray beams are produced by the incident beam reflecting from them. A great many sets of lattice planes can be drawn and planes in the same set are parallel and equidistant. The perpendicular distance between planes in the same set is designated *d*. Each set of planes is identified by three indices *h*, *k*, and *l*, according to how they intersect the x-, y- and z-axes of the crystal lattice. The index *h* gives the number of planes in a given set that intersect the unit cell in the x-axis direction (i.e. the **a** edge of each unit cell), and the indices *k* and *l* give the number of planes that intersect the unit cell in the y-axis and z-axis directions respectively. The unit cell is therefore bounded by the planes ($\bar{1}$00), (010) and (001) (Figure 2.12a). Planes with indices ($\bar{2}$10) are shown in Figure 2.12b. A crystal lattice may present several choices of unit cell (Figure 2.12c). If this is the case then convention dictates the selection of the unit cell whose shape displays the full symmetry of the lattice and which is most convenient, e.g. the axial lengths may be the shortest ones possible and the interaxial angles may be as near as possible to 90°. The rotational symmetry of the lattice is of prime importance when choosing the correct unit cell. There are seven crystal lattice systems, which are defined by the minimum symmetry of the unit cell. These are triclinic, monoclinic, orthorhombic, tetragonal, rhombohedral, hexagonal and cubic. Ordinarily, a unit cell is represented by a single lattice point that is arbitrarily located at one of its corners. However, points may be added at the centres of one or all three unique unit cell faces (face-centering) or at the centre of the unit cell (body-centering). As a result, there are 14 crystal (Bravais) lattices which are listed in Table 2.1.

The molecules in a crystal are related by symmetry. An object is said to be symmetrical if after some movement, real or imaginary, it is indistinguishable from its initial state. Translations, rotations and mirror operations represent the simplest symmetry operations in crystal systems. The stacking of unit cells to form the crystal produces a three-dimensional translational symmetry. Certain crystal lattice systems have a rotational axis about one or more of their three axes (Table 2.1). For example, a monoclinic unit cell has a 2-fold rotation axis along **b** (an *n*-fold rotation corresponds to 360°/*n*). This means that if the unit cell is rotated 180° about the **b** axis the resulting

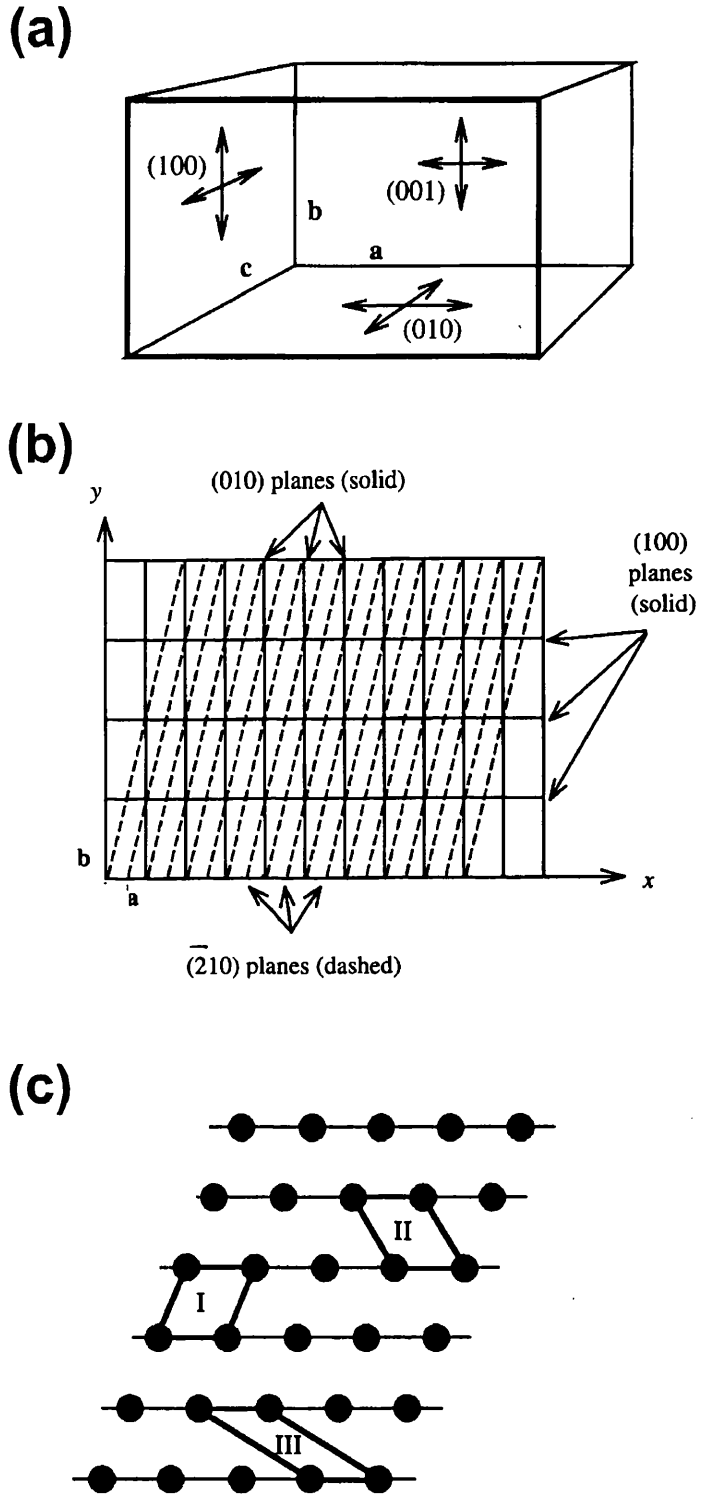


Figure 2.12. Crystal lattice planes. (*a*) A unit cell with edges **a**, **b** and **c** is bounded by the planes (100), (010) and (001). (*b*) $(\bar{2}10)$ planes in a two-dimensional section of lattice. (*c*) A two dimensional lattice, in which the unit cell can be chosen as either I, II or III. (Figures (*a*) and (*b*) adapted from Rhodes, 1993; Figure (*c*) adapted from Drenth, 1994).

Table 2.1. Crystal systems and Bravais lattice types. (Adapted from Glusker & Trueblood, 1985).

Crystal system	Rotational symmetry elements and cell geometry restrictions	Bravais lattices*
Triclinic	No rotational symmetry. No restrictions.	*P*
Monoclinic	**b** is a 2-fold rotation axis. $\alpha = \gamma = 90°$.	*P, C*
Orthorhombic	**a**, **b** and **c** are mutually perpendicular 2-fold rotation axes. $\alpha = \beta = \gamma = 90°$.	*P, C, F, I*
Tetragonal	**c** is a 4-fold rotation axis, **a** and **b** are both 2-fold rotation axes. $a = b$ and $\alpha = \beta = \gamma = 90°$.	*P, I*
Hexagonal	**c** is a 6-fold rotation axis. $a = b$ and $\alpha = \beta = 90°$ and $\gamma = 120°$.	*P*
Rhombohedral	3-fold rotation axis along one body-diagonal of the unit cell. $a = b = c$ and $\alpha = \beta = \gamma$.	*P*
Cubic	3-fold rotation axes along all four body-diagonals of the unit cell and four 4-fold axes parallel to each crystal axis. $a = b = c$ and $\alpha = \beta = \gamma = 90°$.	*P, F, I*

* A face-centred cell is designated *F* if all faces are centred and *A*, *B* or *C* if only one pair of faces is centred. A body-centred unit cell is designated *I*. A unit cell which does not have face- or body-centering is termed primitive and designated *P*. *C* in monoclinic can alternatively be *A* or *I*; *C* in orthorhombic can alternatively be *A* or *B*. *P* in rhombohedral is often called *R*.

unit cell is identical to the original cell.

It is also common for molecules within a unit cell to be related by rotation axes. A mirror plane acts to convert a left-handed molecule to a right-handed molecule. More complex symmetry operations are produced from combinations of these three basic symmetry operation-types: an *n*-fold inversion is a combination of an *n*-fold rotation and a mirror plane operation; a screw axis is an *n*-fold rotation combined with a translation parallel to the rotation axis; and a glide plane results from a mirror operation combined with a translation. The symmetry of a unit cell is described by its space group. There are 230 different ways in which the symmetry operations and the lattice systems can be combined in a crystal and thus there are 230 space groups. However, not all of the 230 space groups are applicable to protein crystals because naturally-occurring proteins contain only L-amino acids and are therefore incompatible with mirror plane operations, which would produce D-amino acids. The 230 space groups are described in the International Tables for Crystallography (Hahn, 1996). Each space group is represented by a capital letter and a number that indicates the lattice type and other symbols which represent its symmetry operations. A convenient description of unit cell symmetry is given by equivalent positions. These are positions in the unit cell that are superimposed on each other by the symmetry operations. For example, the space group $P2_1$ has a 2-fold screw axis along the unit cell axis **c** which means that for a molecule at position *x*, *y*, *z* there will be an identical molecule at position *-x*, *-y*, ½+*z*; where the ½ corresponds to a translation of *c*/2 along the z-axis. If a unit cell contains symmetry it will consist of two or more identical regions that are related by the symmetry operations. Such a region is termed the asymmetric unit and it represents the smallest part of the crystal from which the entire structure of the crystal can be reproduced when all symmetry operations are applied. For example, the $P2_1$ unit cell contains two asymmetric units.

2.3.2. X-ray diffraction experiment

2.3.2.1. Geometry of X-ray diffraction

The monochromatic incident beam produced by an X-ray source is diffracted by a crystal into many discrete X-ray beams (commonly referred to as reflections). The angle at which a diffracted beam will occur can be calculated by considering diffracted

beams as being reflections from the crystal lattice planes. Each set of equivalent, parallel planes in the crystal lattice is treated as an independent diffractor that produces a single reflection. Diffraction only occurs from a set of parallel planes, with indices *hkl* and interplanar spacing d_{hkl}, when Bragg's Law is satisfied:

$$2d_{hkl}\sin\theta = n\lambda \qquad \text{Eq. 2.13.}$$

where λ is the wavelength of the X-rays; and θ is both the angle between the plane and the incident beam and between the plane and the diffracted beam, and 2θ is therefore the angle of diffraction; and n is an integer. This model of diffraction is illustrated in Figure 2.13a and is used to determine the geometry of data collection. When Bragg's Law is satisfied, the X-rays reflected from successive planes are exactly in phase and thus they interfere constructively to produce a strong diffracted beam. If the conditions of Bragg's Law are not met, the reflected beams will be out of phase and, because a crystal is essentially an infinite array of unit cells, they will consist of equal positive and negative contributions that result in no diffracted beam.

In a diffraction experiment, the position and intensity of each diffracted beam are measured by a detector. The position of a reflection is designated by indices *hkl* and these are determined by counting reflections outwards from the central reflection; where the central reflection is taken as the origin with indices *hkl* = 000. The fact that the same convention is used to index the crystal lattice planes and the reflected X-ray beams reveals that there is an intimate relationship between them. The reason for this relationship is best appreciated by considering an imaginary second lattice that is reciprocal to the crystal lattice (it is common to refer to the crystal lattice as the real lattice). Figure 2.13b is used to demonstrate how a reciprocal lattice is produced from a crystal lattice. For example, in the reciprocal lattice the point (120) occurs along the line that is normal to the plane (120) in the crystal lattice and passes through the origin of the reciprocal lattice. The distance from the origin of the reciprocal lattice to point (120) to its origin is $1/d_{hkl}$; where d_{hkl} is the perpendicular spacing between (120) planes in the crystal lattice. All points in the reciprocal lattice are determined from their corresponding planes in the crystal lattice by the same formula. The unit cell of the

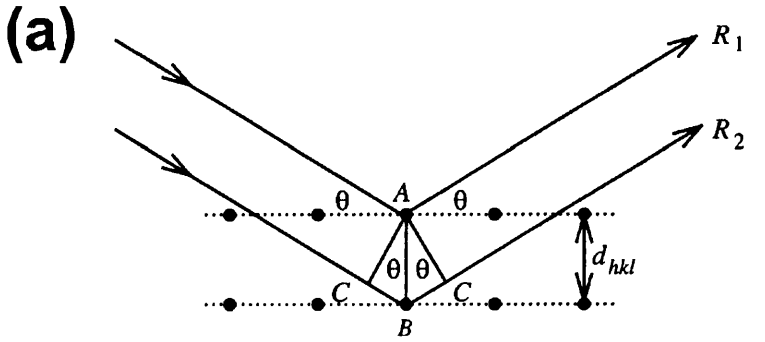
Note: In this discussion of X-ray diffraction from a crystal, for convenience the scattering vectors s_0 and s_1 have their amplitudes set at $1/\lambda$, (see Section 2.2.1.1).

Figure 2.13. (Overleaf) The geometry of X-ray diffraction.

(*a*) Conditions that produce strong diffracted X-rays. If the additional distance travelled by the more deeply penetrating ray R_2 is an integral multiple λ, then rays R_1 and R_2 interfere constructively.

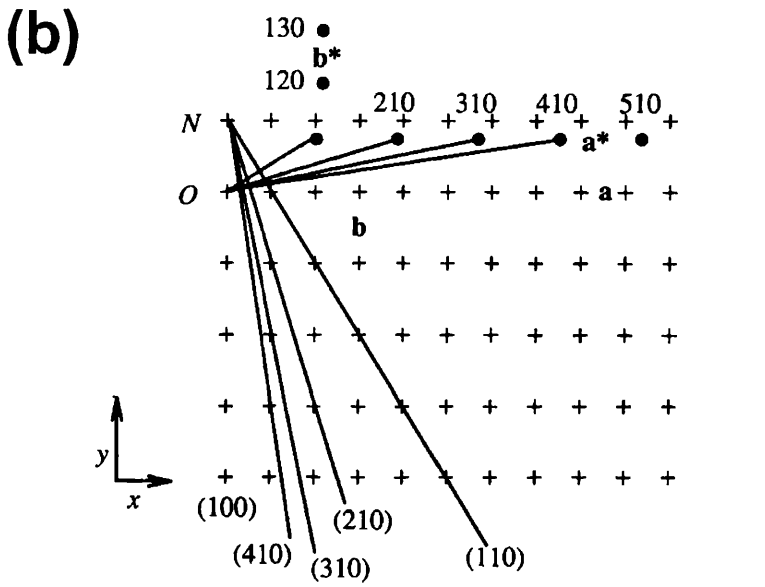
(*b*) Construction of the reciprocal lattice. Crystal lattice points are shown as plus signs (+) and reciprocal lattice points are shown as dots (•). The crystal lattice planes (110), (210), (310) and (410) are shown as solid lines emanating from the point *N*. The reciprocal lattice point (210) is located along the line from the origin *O* that is normal to the crystal lattice plane (210) and at a distance $1/d_{hkl}$ from *O*, where d_{hkl} is the perpendicular spacing between (210) lattice planes. The equivalent relationship is used to locate all reciprocal lattice points. The real unit cell edges **a** and **b** are shown, as well as the corresponding reciprocal unit cell edges **a*** and **b***.

(*c*) Diffraction in reciprocal space. An **a*b*** reciprocal lattice plane and a section of the Ewald sphere with radius $1/\lambda$ is shown. The ray *R* emerges from the crystal when the reciprocal lattice point *P* intersects the circle (Adapted from Rhodes, 1993).



1. $\sin\theta = BC/AB$

2. $BC = AB \sin\theta = d_{hkl} \sin\theta$



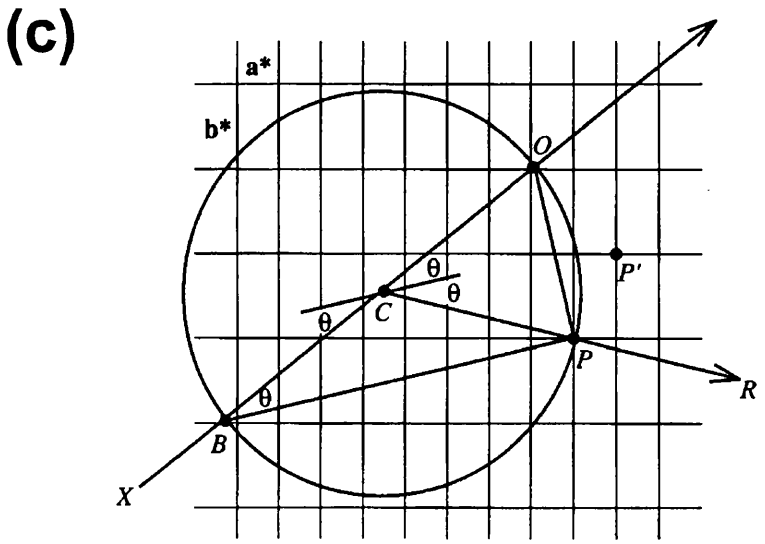


Figure 2.13. The geometry of X-ray diffraction (legend on page 73).

reciprocal lattice has axes **a***, **b*** and **c***, and **a*** is normal to **b** and **c**, **b*** is normal to **a** and **c**, and **c*** is normal to **a** and **b**. The length of **a*** is a^* and is equal to the reciprocal length of unit cell edge **a** ($a^* = 1/a$), and equivalent relationships are true for the lengths of **b*** and **c***. Therefore, a small real unit cell will have a large reciprocal unit cell, and *vice-versa*. Distances in the crystal lattice are expressed in Å or nm, whereas distances in the reciprocal lattice are expressed in $Å^{-1}$ or nm^{-1}. Although the Å and $Å^{-1}$ units are not S.I. units, they are more commonly used in X-ray crystallography because they give simpler numbers since 1 Å corresponds to the diameter of a hydrogen atom.

The reciprocal lattice occupies the same space as the crystal and a sphere of reflection (Ewald sphere) can be constructed to describe the diffraction with respect to the reciprocal lattice (Figure 2.13c). The Ewald sphere has a radius $1/\lambda$ and the incident beam lies along a diameter. The origin *O* of the reciprocal lattice is set as the point where the incident beam emerges from the sphere. Whenever the crystal is rotated so that a reciprocal lattice point touches the surface of the sphere, Bragg's Law is satisfied and a reflected beam can be measured. In Figure 2.13c, the **a*b*** plane of a reciprocal lattice is shown. The X-ray beam enters the sphere at the point *B* and emerges from it in the direction of the arrow *XO*. The reciprocal lattice point *P* is in contact with the circle. A right-angled triangle *PBO* can be constructed, in which the angle *PBO* is equal to θ. From this triangle it is possible to obtain Bragg's Law:

$$\sin\theta = \frac{OP}{BO} = \frac{OP}{2/\lambda} \qquad \text{Eq. 2.14.}$$

$$2\,\frac{1}{OP}\,\sin\theta = \lambda \qquad \text{Eq. 2.15.}$$

P is a reciprocal lattice point and will therefore have the indices *hkl*. Accordingly, the length of the line *OP* is $1/d_{hkl}$ (Figure 2.13b). Equation 2.15 now becomes $2d_{hkl}\sin\theta = \lambda$, which is Bragg's Law when $n = 1$ (Equation 2.13). In Figure 2.13c, the line OP between the origin and point *P* of the reciprocal lattice is perpendicular to the crystal lattice plane which has the same indices *hkl* as *P*. The line *BP* is perpendicular to *OP*

and is therefore parallel to the crystal lattice planes that produce the reflection *P*. The line parallel to *BP* that passes through *C* represents a crystal lattice plane that produces the reflected X-ray beam in the direction *CP* at an angle 2θ to the incident beam. In a diffraction experiment the crystal is rotated in the X-ray beam to bring various reciprocal lattice points into contact with the Ewald sphere. According to this model of diffraction, the positions of reflected beams and their number will only depend on the crystal unit cell dimensions and not the unit cell contents.

2.3.2.2. The diffractometer

The most common X-ray source used for studying protein crystals is the rotating anode tube, although a sealed X-ray tube source may be used for preliminary work and synchrotron radiation can be used if beam time is available. A Rigaku RU200 rotating anode source was used for MFE-23 X-ray crystallography experiments (Chapter 6).

A schematic representation of a rotating anode tube is shown in Figure 2.14a. In the rotating anode tube, a hot filament (cathode) emits electrons that are accelerated by electrically charged plates towards the anode. The anode is usually a copper metal plate. Although most of the electron energy is converted to heat, high intensity X-rays can be produced because the prohibitive build up of heat that is associated with sealed tube X-ray sources is dissipated by the metal anode rotating rapidly. A small part of the electron energy is emitted as X-rays. The X-ray emission spectrum of copper has two sharp peaks that correspond to electron transitions between inner orbitals in its atoms. When high energy electrons collide with and displace electrons from low lying orbitals in the anode atoms, the resulting vacancy is accommodated by electrons dropping from higher orbitals, and the excess energy is emitted as X-rays of a specific wavelength. The two peaks in the copper X-ray emission spectrum occur at wavelengths 1.54 Å which corresponds to an L-shell to K-shell transmission (K_α), and at 1.39 Å which corresponds to an M-shell to K-shell transmission (K_β). X-rays produced at low angles are selected, and these emerge from the tube through windows made of thin beryllium foil. A diffraction experiment is performed using a monochromatic X-ray beam and the K_α radiation is typically selected for copper anode X-ray sources. Monochromatization of

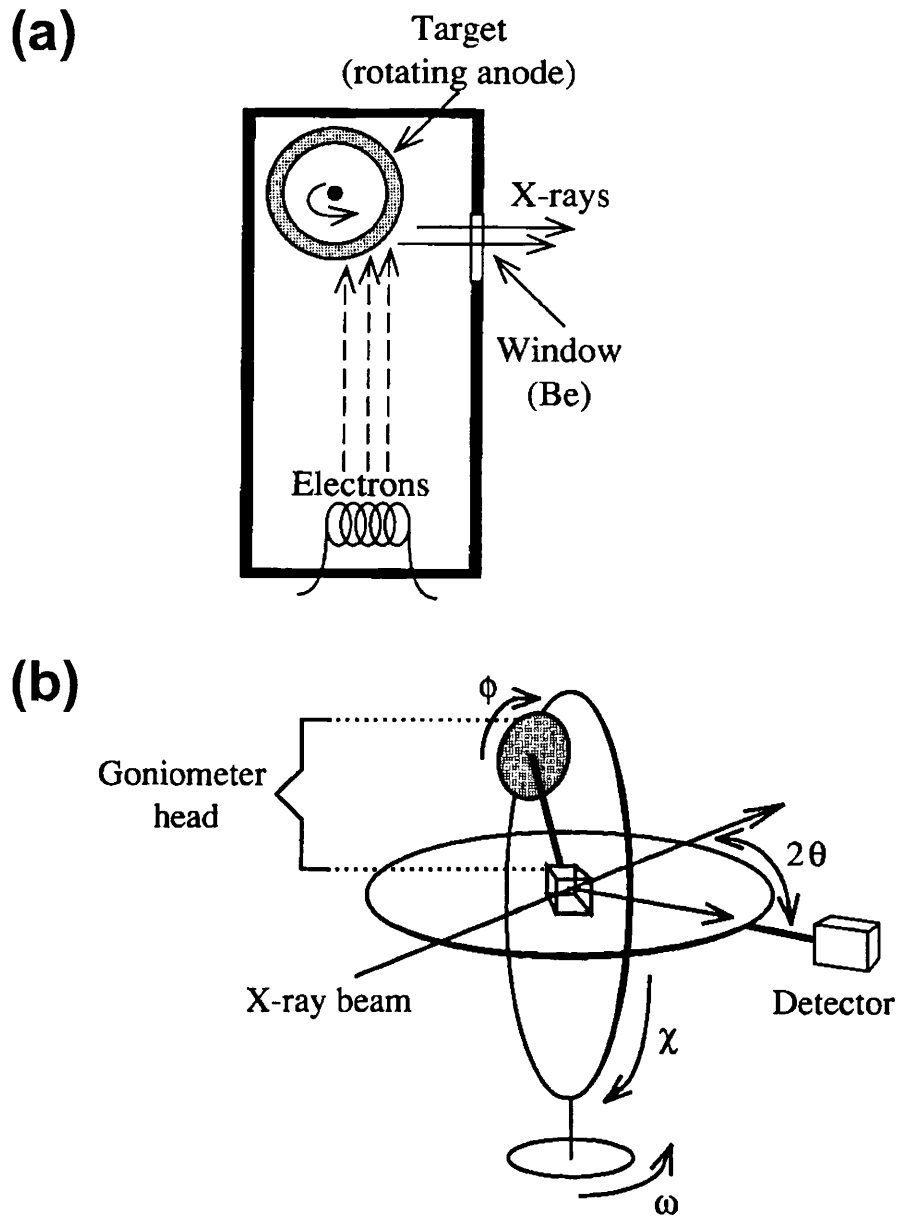


Figure 2.14. X-ray diffractometer. (*a*) A rotating anode tube X-ray source. X-rays are emitted from a rotating metal anode that is bombarded with electrons. (*b*) Representation of an X-ray diffractometer as a system of movable circles. The position of the X-ray beam is fixed. The crystal can be rotated about three angles ϕ, χ and ω, while the position of the detector with respect to the beam is denoted by the angle 2θ (Adapted from Rhodes, 1993).

the copper X-ray emission can be achieved by using either a nickel filter that absorbs strongly at the wavelength of the copper K_β radiation or a graphite monochromator that reflects X-rays of 1.54 Å wavelength.

Several types of detector can be used to measure reflections in an X-ray diffraction experiment. The older detection methods include photographic film and single photon scintillation counters, but they are too time-consuming to gain much use in modern protein X-ray crystallography. Instead, image plates and area detectors have been the detectors of choice in recent years. For MFE-23 data collection (Chapter 6), a Rigaku R-AXIS-II image plate system was used. An image plate consists of a flat base which has a thin layer of an inorganic storage phosphor deposited upon its surface. When an X-ray beam hits the image plate, phosphor electrons are excited to higher energy levels. A significant part of the energy from this interaction is retained by electrons trapped in colour centres. Normally, this stored energy will dissipate slowly over several days, but if the image plate is illuminated with light the stored energy can be released immediately. In an image plate instrument, the plate is scanned with a red laser which causes the energy to be emitted as blue light. If the red light is filtered away, the blue light can be measured and, under certain conditions, the light emitted from a region on the plate is proportional to the number of photons which that particular region was exposed to. An image plate system is advantageous because it can measure reflections over the entire intensity range in a single exposure.

Considerable care must be taken when performing a diffraction experiment with protein crystals. A crystal is held together by non-covalent interactions between molecules, and in protein crystals these interactions are primarily hydrogen bonds between hydrated protein surfaces, which are notably weak in nature. It is therefore essential that mechanical stresses upon protein crystals are avoided. In addition to this, protein crystals have large solvent-filled holes and channels, which commonly account for between 25% and 70% of the crystal volume. The solvent is an essential determinant of the correct protein structure and hence crystal integrity, and for this reason protein crystals are kept in their mother liquor. Thus, a diffraction experiment is typically performed on a crystal that is mounted in a capillary tube filled with its mother liquor

and sealed with resin.

In order to collect diffraction data, the capillary tube containing the crystal is mounted on a goniometer head. A goniometer head has two perpendicular axes, which allow rotation of the crystal in two planes, and two perpendicular sledges, for translations of the arcs. The goniometer head is itself mounted on a goniostat, and rotational and translational adjustments are made to the goniometer head to ensure that the crystal is centred in the X-ray beam. A goniostat is a system of moveable circles which allow automated rotation of the crystal with respect to the X-ray beam and the detector. A goniostat, a fixed X-ray source and a detector constitute what is termed the diffractometer (Figure 2.14b). In the diffractometer, the moveable circles of the goniostat enable the crystal to be independently rotated around three axes (χ, ϕ, and ω), while the detector can be rotated about a fourth angle (2θ which is concentric to ω). During data collection, the crystal is rotated in small oscillation steps around a single axis perpendicular to the X-ray beam. The crystal is oscillated back and forth at each step because normally, for any one stationary position, few or no diffracted beams will be obtained because the conditions of Bragg's law (Equation 2.13) may not apply for that particular orientation of the crystal. For an image plate diffractometer, oscillation steps of approximately 2° are used and this value is determined from the distance between reciprocal lattice points, the maximum resolution and the width of the spots. It is desired that the largest possible oscillation angle is used for reason of minimizing the number of required exposures. It is also possible to minimize the number of required exposures by orientating the shortest reciprocal distance (longest real unit cell distance) along the rotation axis. Symmetry in the crystal lattice also reduces the number of diffracted beams that need to be measured. X-ray diffraction experiments on MFE-23 were performed by Dr Tommy Wan in Dr Brian Sutton's group at the Randall Institute, Kings College, London (Chapter 6).

2.3.3. Obtaining electron density maps from diffraction data

The model of X-ray diffraction contained within Bragg's Law (Section 2.3.2.1) demonstrates that the directions and number of X-ray reflections are dependent on the unit cell dimensions, but not the content of the unit cell. Consequently, an alternative

model of diffraction is required in order to produce atomic coordinate models of a protein from crystallographic data. As has already been stated in Section 2.2.1, X-rays are scattered by electrons and the intensity of scattering at any angle is the combination of the X-rays scattered from different atoms, which give varying degrees of constructive and destructive interference. This means that the intensities of reflections are determined by the content of the unit cell or, to put this more specifically, the intensities of reflections are dependent upon the three-dimensional distribution of electrons in the unit cell. The X-radiation scattered by one unit cell of a crystal structure in any direction in which there is a diffraction maximum *hkl* is termed the structure factor F_{hkl}. The structure factor has both an amplitude $|F_{hkl}|$ and a phase α, where α is calculated relative to the origin of the unit cell. Typically, the structure factor amplitude is placed on an absolute scale, which is based on the amplitude of the radiation scattered by an electron at the origin of the unit cell under the same conditions. In crystallography, the scattering by an atom is conventionally referred to as the atomic scattering factor, and it is dependent on the angle of scattering. The structure factor F_{hkl} can be represented as the sum of the atomic scattering factors in the direction of reflection *hkl* for all atoms in the unit cell. However, rather than considering the diffraction from each atom in the unit cell it is more convenient to consider the electron density as being divided into many small elements of equal volume. The following Fourier series is then used to calculate the structure factor F_{hkl}:

$$F_{hkl} = \int_{h}\int_{k}\int_{l} \rho(x,y,z)\, e^{2\pi i(hx+ky+lz)}\, dx\, dy\, dz \qquad \text{Eq. 2.16.}$$

where the integrations are performed over h, k and l; $\rho(x,y,z)$ is the average electron density within the volume element centred on the point x, y, z; and i is the imaginary number $\sqrt{-1}$. While this expression demonstrates how the structure factors can be calculated from the electron density in the unit cell, in actuality it is desired that the electron density is calculated. The Fourier transform of Equation 2.16 enables the electron density at any point x, y, z to be calculated from the structure factors:

$$\rho(x,y,z) = \frac{1}{V_c} \sum_{h}\sum_{k}\sum_{l} F_{hkl}\, e^{-2\pi i(hx+ky+lz)} \qquad \text{Eq. 2.17.}$$

where V_c is the volume of the unit cell, and the triple summation is performed over all values of the indices *hkl*. By solving this expression, three-dimensional electron density maps can be calculated, into which a model of the protein structure can be built. Solving the structure of a protein by X-ray crystallography therefore requires the determination of its structure factors. It has already been stated that a structure factor has both an amplitude and a phase, and this is perhaps better depicted using the following variation of Equation 2.17:

$$\rho(x,y,z) = \frac{1}{V_c} \sum_{h} \sum_{k} \sum_{l} |F_{hkl}| \cos[2\pi(hx + ky + lz) - \alpha] \qquad \text{Eq. 2.18.}$$

where $|F_{hkl}|$ is the structure factor amplitude and α is the phase. The structure factor amplitude $|F_{hkl}|$ is proportional to the square root of the intensity I_{hkl}, which is measured in the X-ray diffraction experiment. However, the phase angles cannot be obtained directly from the experimental data, and this is commonly referred to as the phase problem. Initial approximations of the phase angles must therefore be determined by an alternative means. Examples of phase-determination methods include isomorphous replacement, anomalous dispersion, and molecular replacement. The following sections describe how the initial electron density maps were obtained from the raw MFE-23 diffraction data (Chapter 6).

2.3.3.1. Data processing

An X-ray crystallography experiment yields a series of two-dimensional images (frames), in which the intensities of reflected X-ray beams are measured against the angle of diffraction (2θ). Each image therefore represents a different slice of the three-dimensional reciprocal space. In order to solve the structure of a protein, these data must be converted to a suitable format. The aim of data processing is to extract the indices *hkl*, the intensity I_{hkl} and the variance of intensity σI_{hkl} for each reflection from the raw data frames. The MFE-23 data was processed using the DENZO program (Otwinowski, 1993). DENZO has a graphical-user-interface and, prior to processing, the data frames were visualized to check that reflections had been resolved. The first stage of data processing was the assignment of reflection indices *hkl* for the determination of the unit cell dimensions, and hence the types of Bravais lattice, that

were compatible with the data. Indexing was performed using between 100 and 300 reflections from a single data frame. An autoindexing routine is utilized in DENZO which performs a complete search of all possible indexing for all reflections simultaneously, but one index at a time. This complete search method allows indexing without the requirement of previous knowledge of the crystal unit cell. After the search, the program finds the three best linearly independent vectors that would index all reflections with the minimal unit cell volume, and definitions of the 14 Bravais lattices from the International Tables for Crystallography are used to determine the lattice types that fit this unit cell. DENZO requires the input of several parameters that describe the geometry of data collection. After autoindexing, a Bravais lattice was assigned to the crystal and certain detector and crystal dependent parameters could be refined. These parameters include the position of the direct X-ray beam, the unit cell dimensions, the orientation of the crystal, the orientation of the detector, and the spread of the beam. It is a fact that a protein crystal will not exhibit a perfect array of unit cells but instead it will possess a mosaic of many submicroscopic arrays in rough alignment with each other, which results in an X-ray reflection emerging from the crystal as a narrow cone rather than a linear beam. A parameter determining the mosaicity of the crystal was also refined in DENZO. Once parameters had been refined, the reflections on all of the data frames were indexed according to the selected crystal lattice type and their intensities were calculated. Data processing outputs files that list the indices *hkl*, the intensity I_{hkl} and the variance of intensity σI_{hkl} for each reflection, and each file contains a single data frame.

2.3.3.2. Data reduction

Due to the symmetry of crystals, there is a certain amount of redundant information in a crystallography data set. Crystallographic data is therefore reduced according to the symmetry operations defined by its space group. Data reduction was performed using the suite of programs from the CCP4 (CCP4, 1994; Version 2.4). In order for data reduction to be performed, the reflection data files must be in the correct format. ROTAPREP was used to convert DENZO files into the MTZ format used by CCP4 programs. The data was then sorted with SORTMTZ so that equivalent reflections were adjacent in the data file. These equivalent reflections will not be on the

same intensity scale, for reasons such as a change in the detector sensitivity during data collection, degradation of the crystal in the beam and a change in the intensity of the beam. Factors for scaling equivalent reflections were calculated from the internal redundancy of the data using SCALA (Evans, 1993). After equivalent reflections have been put on the same intensity scale, their mean intensity values were calculated using AGROVATA (Evans, 1993). It is unlikely that all observations of a reflection will be reliable, therefore AGROVATA weights intensities in its averaging calculation against estimates of the error for each observation. Error estimates are produced from the σI_{hkl} value produced by the DENZO, and from the standard deviation of intensity for multiple observations of the same reflection. The TRUNCATE program is used to convert averaged intensities to mean amplitudes, and place the resulting values on an approximately absolute scale.

2.3.3.3. Molecular replacement

The structure of a protein that is homologous to the protein whose structure is being determined can be used to model the unknown structure. If the homologous structure is correctly placed in the unit cell of the protein under study, phases can be obtained from it that will approximate to the phases of the unknown structure. This is known as molecular replacement, and for it to succeed both the proper orientation and the precise position of all molecules in the asymmetric unit must be determined. However, it is impractical to conduct a simultaneous search of all orientations and positions of the model because this would entail a computationally prohibitive number of model combinations. Molecular replacement is therefore divided into two steps. First, a rotation search is used to determine compatible orientations. Then a molecule in a compatible orientation is subjected to a translation search in order to identify its precise position.

Prior to carrying out molecular replacement it is useful to estimate the number of protein molecules that are present in the unit cell (and hence the number of expected molecular replacement solutions). Matthews (1968) determined that the solvent content of 116 different protein crystal forms is typically near to 43%. In order to quantify the relative protein and solvent content of crystals he developed a coefficient V_m, where V_m

is the unit cell volume per unit of protein molecular weight. V_m values generally occur between 1.68 Å^3/Da and 3.53 Å^3/Da. Knowing the unit cell dimensions, together with the volume of a single protein molecule which can be calculated using SLUV (Chapter 3; Section 3.2.1), the calculation of the number of protein molecules in the unit cell that produces a V_m value within the above range of values is straight-forward.

The rotation search uses the Patterson function. The Patterson function $P(u,v,w)$ is a Fourier summation without phase angles:

$$P(u,v,w) = \frac{1}{V_c} \sum_{h} \sum_{k} \sum_{l} |F_{hkl}|^2 \cos 2\pi(hu + kv + lw) \qquad \text{Eq. 2.19.}$$

where u, v and w are the relative coordinates in the Patterson unit cell; V_c is the volume of the unit cell; $|F_{hkl}|$ are the structure factor amplitudes; and the triple summation is performed over all values of the indices h, k and l. The u, v and w coordinates are used for the Patterson cell to avoid confusion with the coordinates x, y and z in the real cells. The Patterson function produces a contour map of $P(u,v,w)$ that can be calculated without any previous knowledge of the structure because it does not include phase angles. Vectors between atoms in the real structure show up as vectors from the origin to maxima in the Patterson map. Thus, if any two atoms in the real unit cell are separated by a vector (u,v,w) then there will be a peak in the Patterson map at (u,v,w). Vectors between atoms in the same molecule will have relatively short vectors in the Patterson map and their endpoints will be found close to the origin. These are termed self-Patterson vectors and they are independent of the position of the molecule in the unit cell, provided that its orientation is not altered. Consequently, if a model structure is sufficiently similar to the unknown protein structure, the region in the Patterson map containing the self-Patterson vectors will be similar for the model structure and the desired structure when they are in the same orientation. In a molecular replacement rotation search, the model is systematically rotated about three orthogonal axes x, y and z which have their origins at its centre, and the resulting Patterson maps are compared to the experimental Patterson map for the desired protein. This type of search, in which the model and the target structure are different proteins, is referred to as a cross-rotation function. A cross-rotation search will generally be performed to identify the orientations

of all molecules in the asymmetric unit of the crystal.

Once the correct orientation(s) of the model has been determined, a translation search is performed to identify its precise position(s) with respect to the origin of the unit cell. For a single molecule, the simplest way of performing a translation search would be to translate systematically the model along three orthogonal axes x, y and z in order to find the position where its structure factor amplitudes $|F_{hkl}|$ are most similar to those that have been determined for protein being studied. However there are usually more than one molecule per asymmetric unit and consequently more sophisticated translation functions have been developed to find their respective positions. One such translation function has been derived by Crowther & Blow (1967). Like the rotation search, this method is also based on the Patterson function, except that vectors in the Patterson map that correspond to inter-molecular vectors (cross-Patterson vectors) are used. The positions of the molecules in the unit cell are varied until the cross-Patterson vectors between them are in agreement with the Patterson map of the unknown structure.

MFE-23 is an antibody single-chain Fv fragment (Chapter 6), and because antibodies constitute one of the most studied types of protein structure it was convenient to use a molecular replacement strategy to determine approximations of the phase angles. Molecular replacement was performed using the AMORE suite of programs (Navaza, 1994). AMORE consists of five programs. First, the SORTING program sorts the experimental structure factor amplitude data into an appropriate format. Next, the TABLING program calculates the continuous Fourier coefficients corresponding to the search model. The ROTING program is then used to calculate the cross-rotation function according to the fast rotation method of Crowther (1972). The model's self-Patterson vectors are typically calculated over a radius close to the maximum intramolecular distance from its centre. The rotation function works best when low and high resolution data are excluded. This is because low resolution data are rather insensitive to rotation, while the high resolution data can be too sensitive to differences between the model and the unknown protein structure. Solutions to the rotation function are evaluated as a correlation coefficient and the standard output contains all solutions greater than 50% of the maximum value. The TRAING program is used to perform the

translation function, which is based on the Crowther & Blow method (1967). Solutions from the rotation search are input one at a time, and solutions are evaluated as a correlation coefficient. The top ten solutions greater than 50% of the maximum values are outputted. For an asymmetric unit containing more than one molecule, a potential solution can be fixed in subsequent searches for different rotation function solutions. Once the orientation and position of each molecule in the asymmetric unit have been determined, the six positional parameters are refined using the FITING program (Castellano *et al.*, 1992).

The precise space group and hence the lattice type of the crystal is determined by means of collective evidence that originates with the initial indexing of the crystal in DENZO and extends to the solutions from molecular replacement searches. It was therefore typical to first process data in the lowest symmetry space group of the highest symmetry Bravais lattice that is compatible with the data. If data reduction showed that this space group was incorrect, the data was re-processed using the lowest symmetry space group of the compatible lattice type with the next highest symmetry. Data processing and reduction was repeated until a Bravais lattice could be accepted. Once the data had been reduced to the lowest symmetry space group for a Bravais lattice-type, molecular replacement trials were performed using the different rotational symmetry space groups that were compatible with the lattice type. E.g. this would determine whether the asymmetric units within the unit cell were related by a screw axis. After the type of rotational symmetry had been determined, the data was reduced further in the highest possible symmetry space group and the molecular replacement was repeated for this space group. The model outputted by the AMORE FITING program was used to calculate phases for generation of electron density maps and to provide the initial model of MFE-23. The method used to convert this initial model to the MFE-23 structure is presented in Chapter 3 (Section 3.3).

Chapter 3

Biomolecular Modelling

3.1. Introduction

As the term intimates, biomolecular modelling is concerned with creating models that mimic the real structures of biological molecules. An atomic coordinate model represents the most detailed type of model for a protein molecule that can be determined experimentally by X-ray crystallography or NMR. The models deposited in the Brookhaven databank are examples of atomic coordinate models, in which Cartesian x, y, z coordinates specify the position of each atom. In X-ray crystallography models hydrogen atoms are usually omitted because these are not generally detectable at the resolution of the experiment. An important function that can be performed using an atomic coordinate model is the development of hypotheses regarding the function of a protein based on the spatial organization of amino acid residues. It is for this reason that, when a crystallography or NMR model is not available, an approximate atomic coordinate model can provide a useful substitute when interpreting a protein's function. This chapter describes three methods that have been used to develop atomic coordinate models of proteins in the course of this thesis. In Section 3.2, an account is given of an automated technique for modelling the structures of multi-domain proteins, which is based on constraints from X-ray and neutron solution scattering data and domain fold structures. In Section 3.3, the method by which a model can be built and refined against X-ray crystallography data is outlined. In Section 3.4, a description of the procedures used to build a model of a protein domain based on homology to a known structure is given. Finally, several means that can be used for analysing the properties of a protein model are reported in Section 3.5.

3.2. X-ray and neutron solution scattering curve modelling

Experimental scattering data present an important means of validating low resolution models of proteins and glycoproteins in solution. Full atomic models are converted into X-ray and neutron scattering curves, and these modelled scattering curves are compared to experimental scattering data. The models that show the best agreement to the experimental data are selected. Hydrodynamic analyses can be used to further validate solution scattering models. The methodology of X-ray and neutron scattering curve modelling has been reviewed recently (Perkins *et al.*, 1998a), and an outline of the modelling procedure is shown in Figure 3.1. This section provides a description of this

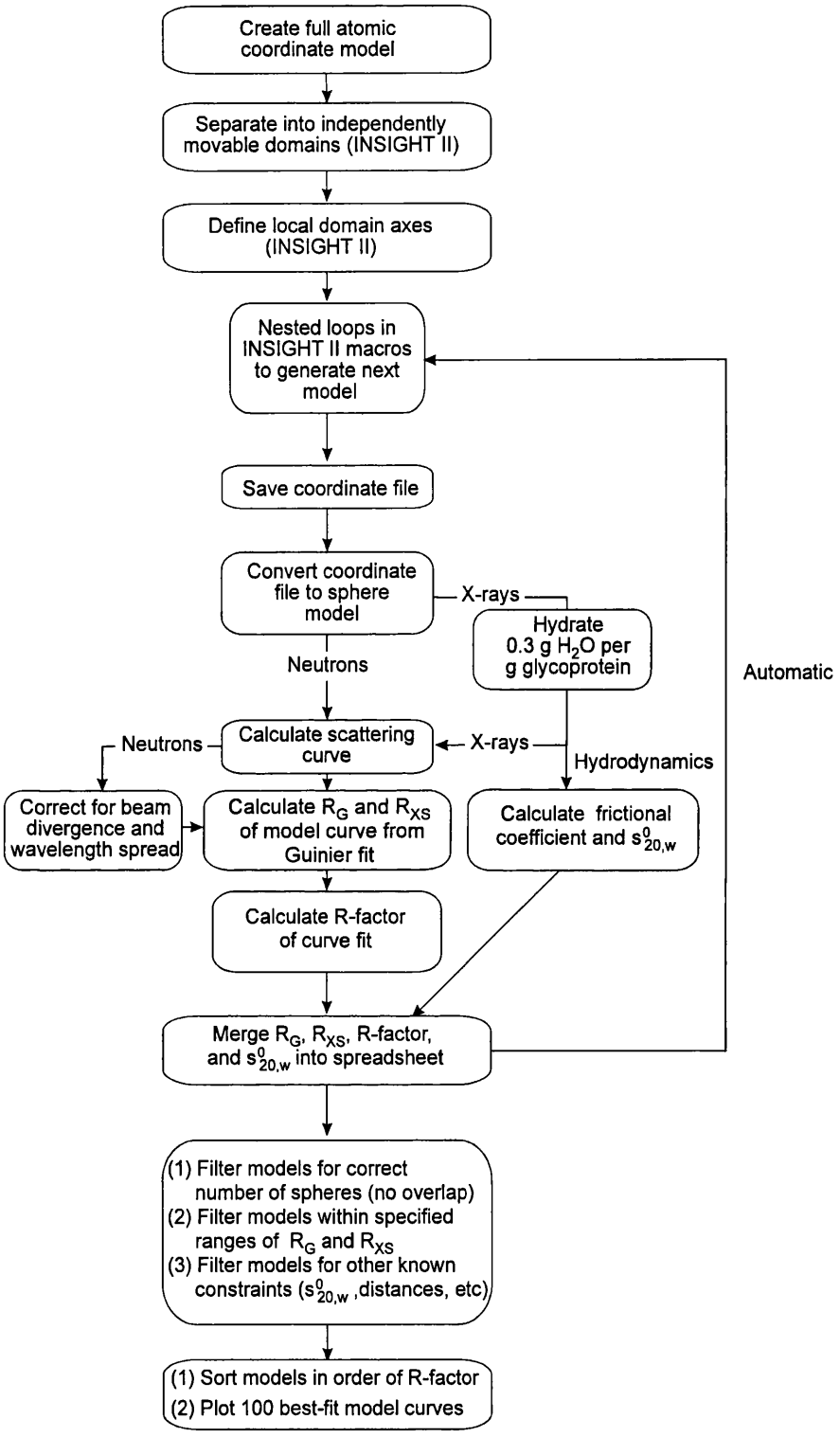


Figure 3.1. Flow chart of the procedure for the automated generation of multidomain models and analysis by scattering curve fits. Each box describes either a stage in the modelling procedure, or how additional information is included to evaluate the models. The automated procedure utilises INSIGHT II and Unix executable script files on Silicon Graphics workstations. The resulting parameters are filtered and sorted using Excel spreadsheets. (Adapted from Perkins *et al.*, 1998a).

methodology. It should be noted that, for convenience, the term protein is used loosely to describe the protein or glycoprotein under study, unless the protein and carbohydrate moieties of a glycoprotein are being specifically considered.

3.2.1. Analysis of glycoprotein composition

Prior to solution scattering modelling, essential information was extracted from the amino acid and monosaccharide composition of the protein using SLUV, which is a FORTRAN program (Perkins, 1986). Thus the volume of an unhydrated protein is calculated from standard crystal structure volumes for each of the amino acid and monosaccharide residues (Chothia, 1975; Perkins, 1986). The extinction coefficient, for a wavelength of 280 nm, is calculated from the tryptophan, tyrosine and cysteine content of a protein by a correction of the Wetlaufer procedure (Perkins, 1986). Electron densities and neutron matchpoints are calculated for the individual protein and carbohydrate components of a glycoprotein. Electron densities are calculated using the partial specific volume, which is derived from a consensus volumes data set that is based on the sum of the dry volume and the electrostricted water shell volume as calculated from the amino acid and carbohydrate composition (Perkins, 1986), and gives values close to the Cohn & Edsall (1943) densitometric calculation. Matchpoints are calculated using the neutron scattering length density, which is derived from the summation of scattering lengths divided by the dry volume of the macromolecule, assuming a 10% nonexchange of the protein mainchain amide protons with solvent.

3.2.2. Small sphere modelling

In order to calculate its X-ray and neutron scattering curves, an atomic coordinate model must first be converted to a sphere model consisting of many small, overlapping spheres of the same total volume as the atomic model. To do this, the atomic coordinate model is placed within a three dimensional array of cubes and every cube that contains a minimum number of atoms has a sphere, of volume equal to the cube, placed at its centre. The minimum number of atoms required to assign a sphere is tested empirically by trial and error against the length of the cube side so that the resulting sphere model has a volume close to the unhydrated volume calculated for the protein. As a rule, the unhydrated volume is calculated from the actual protein

composition rather than from the model composition, for reason of compensating for any discrepancies between these two values. Such discrepancies may occur if the crystal structure is incomplete due to disorder or if homologous domain structures are used in the model as opposed to the actual domain structure. The length of the cube side that is used to produce an unhydrated sphere model is typically about 0.6 nm, which is much less than the maximum resolution of a normal scattering curve. The resolution d is calculated from $2\pi/Q$; which is derived from $\lambda = 2d \sin\theta$ and $Q = 4\pi \sin\theta/\lambda$; and for a scattering curve with a maximum Q of 2.0 nm^{-1} the maximum structural resolution is 3.1 nm. Once a grid size has been determined, this is kept fixed during an automated curve fit search.

Neutron scattering experiments do not detect a hydration shell around proteins, so unhydrated sphere models can be used for neutron scattering curve calculations (Smith *et al.*, 1990; Perkins *et al.*, 1993; Asthon *et al.*, 1997). However, a hydration shell is detected around proteins in X-ray scattering experiments, and therefore calculation of an X-ray scattering curve requires that an unhydrated sphere model is modified according to the volume of the hydration shell. The hydrated volume of a protein is calculated assuming a hydration of 0.3 g of water/g protein and an electrostricted volume of 0.0245 nm^3 per bound water molecule (Perkins, 1986). Two methods have been used to produce a hydrated sphere model from an unhydrated sphere model. The simplest method is to increase the size of each sphere in the unhydrated model so that the resulting volume equalled the hydrated volume. Although this procedure is satisfactory for globular proteins of compact structure, it can significantly distort the protein structure if it contains a void space at its centre. To overcome this problem, an alternative method has been developed recently (Ashton *et al.*, 1997). In this second method, spheres corresponding to water molecules are added evenly over the surface of the unhydrated sphere model, so that the desired hydration volume is achieved. Hydrated sphere models are used for X-ray scattering curve calculations.

3.2.3. Debye scattering curve calculation

A sphere model containing spheres of a single scattering density is used to calculate a scattering curve $I(Q)$, by the following application of Debye's Law adapted

to spheres (Glatter & Kratky, 1982; Perkins & Weiss, 1983):

$$\frac{I(Q)}{I(0)} = g(Q)\ (n^{-1} + 2n^{-2} \sum_{j=1}^{m} A_j \frac{\sin Qr_j}{Qr_j}\ ) \qquad \text{Eq. 3.1.}$$

$$g(Q) = (3(\sin QR - QR \cos QR))^2 / Q^6R^6 \qquad \text{Eq. 3.2.}$$

where $g(Q)$ is the squared form factor for the sphere of radius R, n is the number of spheres filling the body, A_j is the number of distances r_j for that value of j, r_j is the distance between the spheres, and m is the number of different distances r_j. The single density scattering curve calculation is applicable to proteins, and to glycoproteins that have low carbohydrate contents provided that comparisons between the model scattering curves and the experimental data are consistent for the neutron and X-ray curves.

If a glycoprotein has a high carbohydrate content, differences between the scattering densities of its protein and carbohydrate moieties that may give rise to systematic deviations between the neutron and X-ray curves can be modelled using two-density sphere models. The relative scattering densities of the protein and carbohydrate spheres in an unhydrated sphere model are weighted by an approximation to the ratio of their calculated neutron matchpoints, compared to 100% ^2H_2O. Correspondingly, the relative scattering densities of the protein and carbohydrate spheres in a hydrated sphere model are weighted by an approximation to the ratio of their calculated electron densities compared to the electron density of bulk water. The different scattering densities of protein and carbohydrate are incorporated in two-density scattering curves calculated from (Glatter & Kratky, 1982):

$$\frac{I(Q)}{I(0)} = g(Q)\ [\ n_1\rho_1^2 + n_2\rho_2^2 + 2\rho_1^2 \sum_{j=1}^{m} A_j^{11} \frac{\sin Qr_j}{Qr_j} + 2\rho_2^2 \sum_{j=1}^{m} A_j^{22} \frac{\sin Qr_j}{Qr_j}$$

$$+ 2\rho_1\rho_2 \sum_{j=1}^{m} A_j^{12} \frac{\sin Qr_j}{Qr_j}\ ] \times (n_1\rho_1 + n_2\rho_2)^{-2} \qquad \text{Eq. 3.3.}$$

where $g(Q)$ is the squared form factor for the sphere of radius R (Equation 3.2); the model is constructed from n_1 and n_2 spheres of different densities ρ_1 and ρ_2; A_j^{11}, A_j^{22}, and A_j^{12} is the number of distances r_j for that increment of j between the spheres 1 and 1, 2 and 2, and 1 and 2 in that order; the summations Σ are performed for $j = 1$ to m, where m is the number of different distances r_j.

Synchrotron X-ray cameras utilise a pin-hole configuration that does not produce geometrical distortion of the beam, so a calculated X-ray curve does not have to be corrected in order for it to be compared to an experimental curve. Although neutron cameras such as those at the ILL and the RAL also use pin-hole geometries, their dimensions are larger than X-ray cameras and they also use longer wavelengths to maximise the available neutron flux, and these reasons necessitate that instrumental corrections are applied to the model neutron scattering curves. For the neutron scattering cameras D11 and D17 at the ILL in Grenoble, a Gaussian function based on a 16% wavelength spread $\Delta\lambda/\lambda$ (full-width-half-maximum) at λ of 1.0 or 1.1 nm and a beam divergence $\Delta\theta$ of 0.016 radians as an empirical correction, has been applied to model neutron curves. The theoretical values of $\Delta\lambda/\lambda$ are respectively 8% and 10% for these two cameras, while that for $\Delta\theta$ depends on both the beam aperture and the size of the detector cells and is approximately 0.01 radians. A re-evaluation of $\Delta\lambda/\lambda$ for D17 data for serum amyloid P component gave 10% in good agreement with theory, although $\Delta\theta$ was larger at 0.024 radians (Ashton *et al.*, 1997). The neutron fits deteriorate at large Q and this may indicate a small residual flat background that arises from incoherent scatter from the protons in the protein. The values used to correct model

neutron curves for comparison to D17 data are also used for comparisons to D22 data. On LOQ, wavelengths in the range from 0.2 to 1.0 nm are used simultaneously (where time-of-flight techniques provide the necessary monochromatisation), and although this complicates the beam corrections, a Gaussian function as for D17 data has been applied to model neutron curves to obtain reasonable comparisons to the experimental data. Values of 16% for $\Delta\lambda/\lambda$ for a putative λ of 1.0 nm and 0.016 radians for $\Delta\theta$ were initially used (Mayans *et al.*, 1995), but a later re-evaluation showed that values of 10% for $\Delta\lambda/\lambda$ for a putative λ of 0.6 nm and 0.016 radians for $\Delta\theta$ gave the best agreement between the model neutron curves and LOQ data (Ashton *et al.*, 1997).

The method of converting an atomic coordinate model into unhydrated and hydrated sphere models, followed by the calculation of neutron and X-ray scattering curves was initially tested using solution scattering data for β-trypsin and α_1-antitrypsin, for which crystal structures were known (Smith *et al.*, 1990; Perkins *et al.*, 1993). Recently the methodology has been evaluated more rigorously using the crystal structure for pentameric serum amyloid P component (Ashton *et al.*, 1997).

3.2.4. Automated modelling using domain fold structures

The procedure for calculating scattering curves from an atomic coordinate model is implemented by a suite of FORTRAN computer programs (S. J. Perkins, unpublished software) and these have been incorporated into automatic strategies for the generation of atomic coordinate models (Figure 3.1; Mayans *et al.*, 1995; Beavil *et al.*, 1995; Perkins *et al.*, 1998a). All solution scattering models are produced using INSIGHT II molecular graphics software and associated programs (Biosym/MSI, San Diego, USA). Typically, an initial model of the protein is generated from models of its constituent domain structures. After consideration of information relevant to the association of domains within the protein structure, moveable protein fragments are defined so as to consider the least flexible model. At its smallest, a moveable fragment could consist of a single domain fold, but if domains are known to associate into well-defined structures, e.g. the association of serum amyloid P component domains into a pentameric ring structure (Ashton *et al.*, 1997), the moveable fragments may be much larger. A set of axes is defined for each moveable fragment, and a Biosym Command Language macro

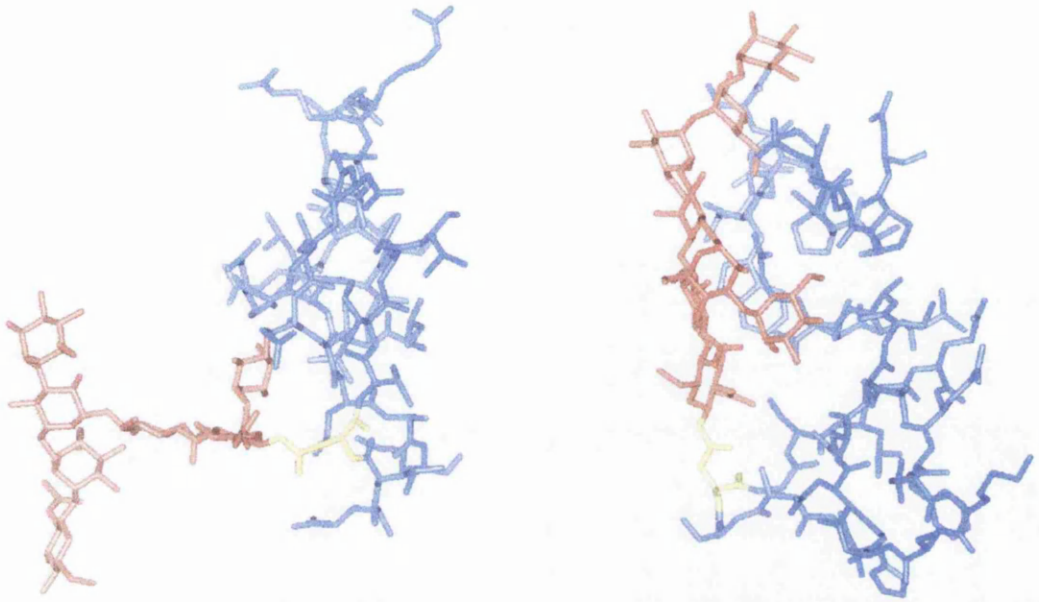


Figure 3.2. The two general conformations that N-linked carbohydrate chains adopt in glycoprotein structures. The carbohydrate chains which are coloured red are attached to an asparagine residue (yellow). The blue structure represents the region of the protein that is proximal to the glycosylation site. The example of an extended carbohydrate conformation is the oligosaccharide that binds to Asn159 in the human leukocyte elastase structure (PDB code: 1ppg; chain E; Bode *et al.*, 1989). In this conformation the carbohydrate extends away from the protein surface and into solution. The example of the compact carbohydrate conformation is the oligosaccharide that binds to Asn395 in the *Aspergillus awamori* glucoamylase structure (PDB code: 1agm; Aleshin *et al.*, 1992 & 1994). In this conformation the carbohydrate lies flat against the protein surface.

is written to systematically move the fragments, either by translations or rotations, and thereby generate a series of models. Such automated solution scattering modelling searches have been performed on a Silicon Graphics INDY R4400SC Workstation with 64 Mb of memory and a 4 Gb hard disk, and their run times were up to several weeks.

3.2.5. Carbohydrate structures

If the protein being studied contains glycosylation sites, an average three-dimensional structure is determined for the N- and O-linked oligosaccharide chains using carbohydrate structural composition analyses in the literature. These structures are modelled by adapting suitable oligosaccharide structures from the Brookhaven database, and the structures are positioned at putative N- and O-linked glycosylation sites on the protein. The structures of N-linked carbohydrates on glycoprotein crystal structures can be divided into two broad types; those which extend away from the protein surface into solution, and those which lie against the protein surface in a compact conformation (Figure 3.2). An example of an extended N-linked carbohydrate is bound to Asn159 on human leukocyte elastase (Brookhaven code: 1ppg; Bode *et al.*, 1989), while an example of a compact N-linked oligosaccharide is bound to Asn395 on *Aspergillus awamori* glucoamylase (Brookhaven code: 1agm; Aleshin *et al.*, 1992 & 1994). Both conformations of N-linked carbohydrate are usually tested in glycoprotein models to distinguish between extended and compact ones.

3.2.6. Model evaluation

Each model is evaluated using several criteria. The volume of each sphere model, indicated by the number of spheres generated from the coordinates during the grid transformation, is used to determine whether an automated modelling procedure has produced steric overlap of its domains. Each model is also evaluated by comparison of its calculated scattering curves to experimental data. Guinier fits are used to calculate R_G and R_{XS} values from each model scattering curve over the same Q ranges used for experimental measurements. An R-factor is calculated for a quantitative comparison of each model scattering curve $I(Q)_{cal}$ against experimental scattering data $I(Q)_{exp}$ over the whole Q range, where $I(0)_{cal}$ is set as 1000:

$$R = \frac{\sum |I(Q)_{exp} - I(Q)_{cal}|}{\sum |I(Q)_{exp}|} \times 100\% \qquad \text{Eq. 3.4.}$$

The use of R-factor values is analogous to the evaluation of crystallographic models, and it has been successfully applied to evaluation purposes for solution scattering modelling procedures (Smith *et al.*, 1990; Beavil *et al.*, 1995). In practice, to minimize the number of models that have to be considered from a large automated search, models that do not conform to defined ranges of model volume, R_G and R_{XS} are rejected using filters.

3.2.7. Hydrodynamic analyses

The theoretical sedimentation coefficient $s^o_{20,w}$ of each hydrated sphere model can also be determined and its comparison to the experimental value offers a further criterion for validating solution scattering models (Figure 3.1). In order to determine the theoretical $s^o_{20,w}$ for a hydrated sphere model, the frictional coefficient f of the model is calculated using the modified Oseen tensor procedure of Bloomfield (Garcia de la Torre & Bloomfield, 1977a, 1977b; Perkins, 1985). In this procedure, each sphere in a multi-sphere model is assumed to be a point source of friction. The i^{th} sphere has an associated frictional force F_i, which is calculated from its Stokes law frictional coefficient:

$$F_i = 6\pi\eta r_i(u_i - v_i)(i=1,n) \qquad \text{Eq. 3.5.}$$

where η is viscosity of the solvent; r_i is the hydrodynamic radius of the i^{th}sphere; u_i is the velocity of the i^{th} sphere; v_i is the velocity the solvent would have at the same point if the sphere was absent; and n is the number of spheres (also see Equation 2.12). The frictional coefficient of each sphere in a multi-sphere model is modulated by the hydrodynamic interactions of other spheres in the model, by what is referred to as a hydrodynamic interaction tensor. The hydrodynamic interaction tensor, T_{ij} for spheres of finite sizes and of different hydrodynamic radii of r_i and r_j is given by:

$$T_{ij} = \frac{1}{8\pi\eta R_{ij}} \left[I + \frac{R_{ij}R_{ij}}{R_{ij}^2} + \frac{(r_i^2 + r_j^2)}{R_{ij}^2}\left(\frac{I}{3} - \frac{R_{ij}R_{ij}}{R_{ij}^2}\right)\right] \qquad \text{Eq. 3.6.}$$

where I is the unit vector and R_{ij} is the vector between spheres i and j. The hydrodynamic interaction tensor and the frictional force terms are related by:

$$v_i = v_i^0 - \sum_{j=1}^{n} T_{ij} F_j \qquad \text{Eq. 3.7.}$$

(excluding terms with $i = j$), where v_i^0 is the unperturbed solvent velocity when the spheres are absent. The program GENDIA (Garcia de la Torre & Bloomfield, 1977a, 1977b; Perkins, *et al.*, 1993) is used to implement these equations, in which iterative procedures are used to calculate f for an array of spheres. The theoretical sedimentation coefficient can then be calculated from f using Equation 2.11 and compared to the experimental value.

3.3. X-ray crystallographic model building and refinement

By the method of molecular replacement, a model of the MFE-23 single chain Fv antibody fragment was built using X-ray crystallographic data (Chapter 6). The Fv coordinates from the J539 structure (PDB code: 2fbj; residues H1 to H120 and L1 to L106) provided the preliminary model and its calculated phases were used to generate initial electron density maps. The model was gradually improved by a cyclic process of manual rebuilding and refinement (Figure 3.3). In this section, the methods used in crystallographic model production are described.

3.3.1. Crystallographic refinement using X-PLOR

An X-ray crystallography model is refined to improve its quality. Refinement of the MFE-23 model was performed using the X-PLOR program system (Version 3.1; Brünger, 1992a). The primary use of X-PLOR has been the three-dimensional structure determination of macromolecules using crystallographic or NMR data. The program allows explorations of the conformational space of a macromolecule to be performed that are restricted against a combination of empirical energy functions and experimental data. Its application to the refinement of X-ray crystallographic models is enhanced by the provision of numerous example routines that can be adapted to solve specific problems.

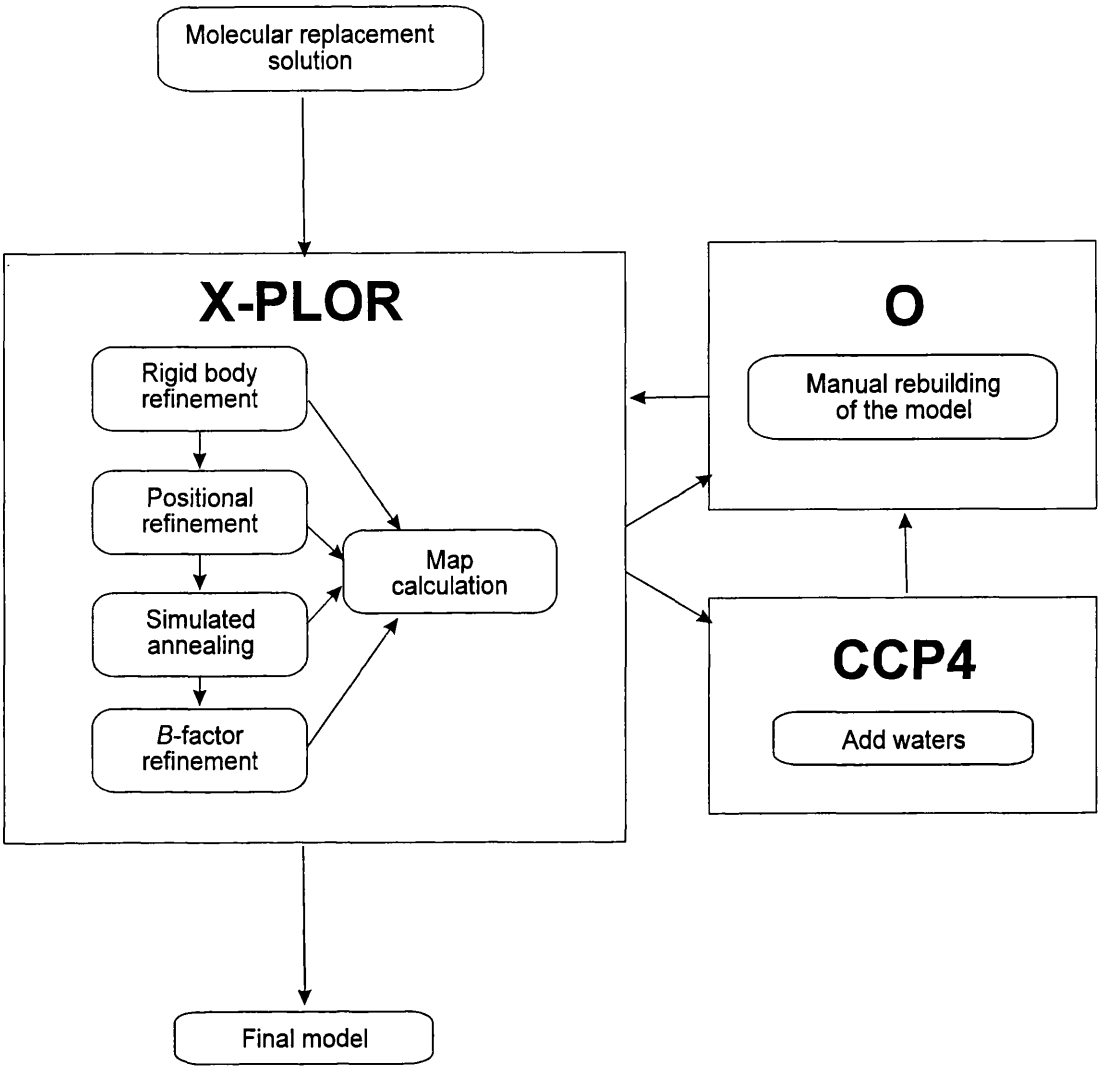


Figure 3.3. Flow chart of the procedures used to create an X-ray crystallography model from the initial phase solution by molecular replacement. The X-PLOR program contains numerous routines that are used for refining a model structure against experimental data and an empirical energy function. The model is manually rebuilt so that its atomic structure fits the features of electron density maps using the three-dimensional molecular graphics software O. After many cycles of model building and refinement, water molecules are added to the model using programs from CCP4.

There are two major aspects of model refinement. The first of these is the reduction of the differences between the experimental data (structure factor amplitudes) and the corresponding data calculated from the model structure, which has the aim of increasing the similarity between the model and the actual structure. The second function fulfilled by refinement is to ensure that the model is energetically acceptable, and thereby remove any bad stereochemistry that may have been introduced into the model during manual rebuilding, or to prevent the introduction of such bad stereochemistry upon convergence of the experimental and model structure factor amplitudes. In X-PLOR, the experimental crystallographic data and the empirical energy of the macromolecular system are incorporated in a single energy function:

$$E_{Total} = E_{Empirical} + E_{Xref} \qquad \text{Eq. 3.8.}$$

where E_{Total} is the total energy of the system, $E_{Empirical}$ is the energy of the molecule calculated from the empirical energy function, and E_{Xref} comprises a restraining target function based on the crystallographic diffraction data. Computing the total energy of a protein crystal system forms the basis of the refinement procedures in X-PLOR.

3.3.1.1. The empirical energy function

The empirical energy function contains conformational and nonbonded interaction energy terms, and the following expression contains those energy terms used for the refinement of MFE-23:

$$E_{Empirical} = \sum_{p=1}^{N} w^{p}_{Bond}E_{Bond} + w^{p}_{Angl}E_{Angl} + w^{p}_{Dihe}E_{Dihe} + w^{p}_{Impr}E_{Impr}$$

$$+ w^{p}_{Vdw}E_{Vdw} + w^{p}_{Elec}E_{Elec} + w^{p}_{Pvdw}E_{Pvdw} + w^{p}_{Pele}E_{Pele} \qquad \text{Eq. 3.9.}$$

Each energy term is computed between two selected sets of atoms, and the sum is carried out over all such sets of atoms with weights w_n^p. The term E_{Bond} describes the covalent bond energy, and the term E_{Angl} describes the bond angle energy. Bond torsion angles are defined by the terms E_{Dihe} and E_{Impr}, which describe the dihedral and improper

energies respectively. All of these energy terms describe the conformational energy of the macromolecular system. The remaining energy terms represent nonbonded energy terms. E_{Vdw} and E_{Elec} are the van der Waals and electrostatic energy terms respectively. These two terms along with the four conformational energy terms form the default empirical energy function in X-PLOR. For some crystallographic refinement routines, energy terms describing the effects of van der Waals interactions (E_{Pvdw}) and electrostatic interactions (E_{Pele}) from symmetry-related molecules are also included. X-PLOR calculates the empirical energy terms using parameters derived from an analysis of small molecules in the Cambridge Crystallographic Database (Engh & Huber, 1991).

At the start of each refinement cycle a molecular structure file containing information regarding the names, types, charges and masses of atoms; residue and segment names; and a list of the empirical energy terms, must be created. The *generate.inp* routine provided with X-PLOR is used to produce a molecular structure file from a model coordinate file. This routine is also used to add hydrogens to the atomic coordinate model, because these are essential for empirical energy calculations. In order to perform this routine, coordinate files for each of the individual segments of the model must be input separately. A segment is defined as a single polypeptide chain, a ligand, a substrate, a cofactor, a nucleic acid strand, or all the water molecules combined, and the model coordinate file must therefore be split into separate files containing the individual segments. The *generate.inp* routine also requires that the positions of disulphide bridges in each segment are defined. Two files are output by this routine: the molecular structure file and a second file that contains the atomic coordinates for the model, including those for hydrogens.

3.3.1.2. *R*-factor, R_{free} and the crystallographic target function

The preliminary model produces structure factor amplitudes that are in poor agreement with the observed experimental values. This agreement is usually represented by an *R*-factor:

$$R = \frac{\sum_{hkl} ||F_{obs}| - k|F_{calc}||}{\sum_{hkl} |F_{obs}|} \times 100\% \qquad \text{Eq. 3.10.}$$

where $|F_{obs}|$ and $|F_{calc}|$ are the structure factor amplitudes determined from the experimental data and the model respectively; and k is a scaling factor. It is not uncommon for the initial model to have an R-factor as high as 50%. During refinement the model is modified to find a closer agreement between the calculated and observed structure factor amplitudes and a well-refined model will generally have an R-factor value less than 25%. However, the R-factor alone is not suitable for quantifying the success of refinement. This is because modern refinement algorithms are so powerful that grossly incorrect structures can be fitted to experimental crystallographic data with a low R-factor (Kleywegt & Jones, 1997). It is for this reason that an independent R-factor, termed R_{free}, is used to validate the refinement (Brünger, 1992b). R_{free} is calculated from between 5 and 10% of the experimentally-determined reflections that have been selected at random, but evenly over the entire resolution range. The *setup_free_r.inp* routine from X-PLOR is used to randomly assign reflections to the R_{free} data set, and this data set is maintained independently throughout refinement. Once the R_{free} data set has been assigned, its constituent reflections are not used in standard R-factor calculations nor in computing E_{Xref}, the experimental crystallographic energy term. For an ideal refinement procedure, the R_{free} value will decrease at the same rate at which the R-factor is reduced, and the final R_{free} value should be less than 35% (Kleywegt & Jones, 1997).

E_{Xref} represents the target function for crystallographic refinement. For the refinement of MFE-23, E_{Xref} was calculated using a crystallographic residual function:

$$E_{Xref} = R' = \frac{W_A}{N_A} \sum_{hkl} w_{hkl} \left(|F_{obs}| - k|F_{calc}|\right)^2 \qquad \text{Eq. 3.11.}$$

where R' is the crystallographic residual, which is similar to the crystallographic R-factor; W_A is an overall weight; N_A is a normalization factor; w_{hkl} is the individual weights of the reflections; $|F_{obs}|$ and $|F_{calc}|$ are the structure factor amplitudes determined from the experimental data and the model respectively; and k is a scaling factor. The values of N_A, w_{hkl} and k are calculated within X-PLOR, but it is necessary for the user to define the value of W_A. The value of W_A is important because it governs the relative

influences that the E_{Xref} and the $E_{Empirical}$ energy functions have on the total energy of the system. Therefore, before the model can be refined against both energy functions, it is necessary to run the *check.inp* routine which calculates the "ideal" W_A value. This routine performs a brief molecular dynamics simulation that does not include the E_{Xref} term. In actuality, the W_A value calculated by *check.inp* will tend to weight the total energy function too much towards the E_{Xref} term and this can produce bad stereochemistry in the model. Consequently, a value in the region of half of the "ideal" is used. Kleywegt & Jones (1997) recommend that the behaviour of simulated annealing refinement is used to guide the selection of the W_A value; where this value should allow the *R*-factor and R_{free} values to decrease at similar rates. Once selected, the same W_A value is used throughout the refinement.

3.3.1.3. Rigid body refinement

The positions of the separate domains (or polypeptide chains) are refined before any manual rebuilding of the model is attempted. To do this, each domain is treated as a rigid body whose position is governed by three rotational parameters and three translational parameters. Refinement of the positions of these rigid bodies is performed using the *rigid.inp* routine. The empirical energy terms are excluded so that only the crystallographic residual is used to calculate the energy of the system. Refinement then proceeds by adjusting the six positional parameters until the *R*-factor reaches a minimum. At the end of rigid body refinement, the coordinates for the model are written to a file.

3.3.1.4. Positional refinement

The aim of refinement is to reduce the total energy of the molecular system by adjusting its atomic coordinates. It is convenient to visualize the energy function of a molecular system as a multi-dimensional surface, which undulates to produce numerous peaks (maxima of the energy function) and troughs (minima of the energy function). The model at the start of a refinement cycle will have a combination of atomic coordinates and energy that corresponds to a specific point on the energy surface. If an energy minimization algorithm is used it will aim to reduce the energy of the model by proceeding in a "downhill" direction until it locates a minimum close to the starting

point. To do this, the minimizer must determine both the direction towards the minimum and the distance to the minimum in that direction. For many procedures, the direction the minimization takes is driven by the local gradient or derivatives of the energy function. The gradient is calculated from the change in energy against the atomic coordinates of the model, and it indicates the direction towards of a minimum and the steepness of the local slope. Figure 3.4 illustrates a simple example of how derivatives can be used to find the minimum of an energy function. Starting from point **a**, the local derivative indicates that the minimum lies in the direction of **d**, and minimization proceeds in this direction until the minimum along this line, **c**, is reached. At **c** the largest derivative is perpendicular to the **a-d** direction and the minimization now proceeds in this direction until a new minimum is located. Thus, the whole minimization procedure is a series of searches to find the minimum along a one-dimensional line, and each one-dimensional line search is orthogonal to the one previous. Procedures that calculate the minimization direction from the gradient are commonly termed first-order minimization methods because they calculate the gradient from the first derivative. Two frequently used first-order algorithms are the steepest descents and conjugate gradient methods. In the steepest descents method, the minimization moves in the direction of the largest local "downhill" gradient as indicated in Figure 3.4. The direction of the gradient is determined by the largest interatomic forces and so the steepest descents method is a good one for relieving the highest energy features in an initial configuration. The method is generally robust even when the starting point is far from a minimum. However, the fact that each new direction undertaken by the minimizer is perpendicular to the previous one means that the directions the minimizer takes oscillate along the way to the minimum. This behaviour is particularly inefficient if the local energy surface resembles a long, narrow valley because the minimizer will have to take many steps. Also, as the steepest descents minimizer reaches the minimum, it is forced to make a right-angled turn at each point, even though that may not be the best route to the minimum, and oscillation of the path therefore means that it continually over corrects itself so that later steps reintroduce errors that were corrected by previous ones. This problem is overcome in the conjugate gradients method by preventing the next direction vector from undoing earlier progress. The conjugate gradients algorithm produces a complete basis set of mutually conjugate

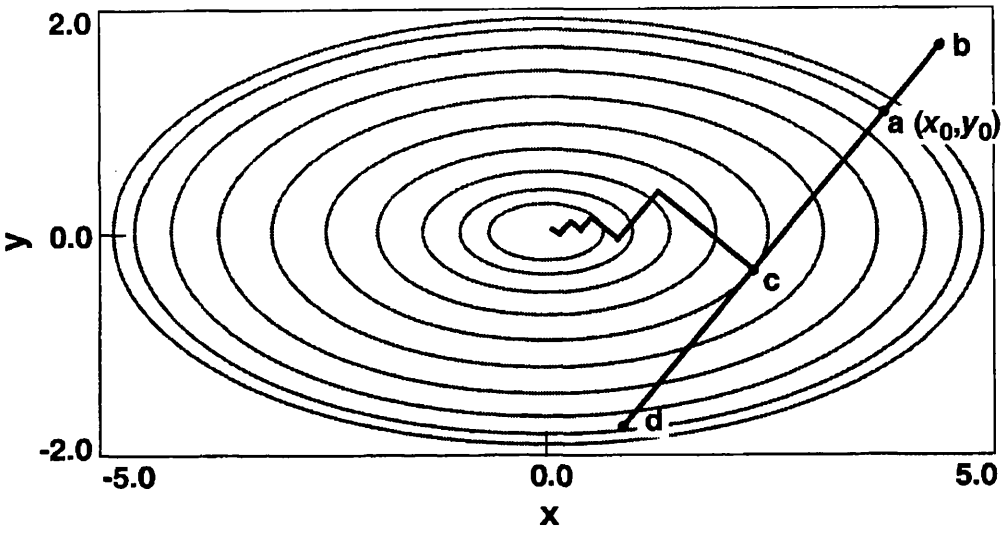


Figure 3.4. Minimization of the energy function $E(x,y) = x^2 + 5y^2$ by a steepest descent procedure. The derivative from the initial point **a** defines the line search direction, which is towards the point **d**. Point **c** corresponds the minimum of the energy function along this line. At point **c**, a line search orthogonal to the first is undertaken, which proceeds until its minimum is reached. Orthogonal lines searches are performed until the minimum of the energy function is reached (Adapted from DISCOVER 2.9.7/95.0/3.0.0 user guide).

gradient directions such that each successive step continually refines the direction toward the minimum. Thus, while the gradients at each point are orthogonal, the directions are conjugate. The conjugate gradient method is generally much better than the steepest descent method once the initial strain has been removed.

The total energy of a molecule can be minimized in X-PLOR using the conjugate gradient method of Powell (1977). During minimization, the coordinates of all atoms that have not been fixed by the user are allowed to change. This energy minimization is performed using the *positional.inp* procedure. Energy terms for the van der Waals and electrostatic interactions between symmetry-related molecules are added to the default empirical energy function. The crystallographic target function E_{Xref} is included in the total energy calculation so it is necessary to assign the value to W_A before performing positional refinement. Positional refinement is performed until the *R*-factor reaches a minimum and the coordinates for the model at the end of the refinement are written to a file.

3.3.1.5. Simulated annealing

The real power of X-PLOR for crystallographic refinement lies in its incorporation of a simulated annealing protocol. A conventional minimization procedure, such as the positional refinement method outlined above, only follows energetically "downhill" steps so the refinement is only able to find a local energy minimum. In contrast to this, simulated annealing includes a temperature parameter that allows "uphill" steps to be taken. During simulated annealing, the molecular system is initially subjected to high temperatures so that high energy barriers separating large regions of conformational space can be crossed. Then the system is slowly cooled to restrict access to successively smaller regions of conformational space until a conformation closer to the global minimum than the starting model is obtained. Simulated annealing has been routinely applied to crystallographic refinement since the success of this technique was demonstrated by Brünger and colleagues (Brünger *et al.*, 1987; Brünger, 1988). Before simulated annealing is performed in X-PLOR, several cycles of energy minimization are performed using the *prepstage.inp* routine for reason of relieving strain or bad contacts between the initial coordinates. The *slowcool.inp*

routine is then used to carry out simulated annealing. In both routines, energy terms for the van der Waals and electrostatic interactions between symmetry-related molecules are added to the default empirical energy function. The crystallographic target function E_{Xref} is included in the total energy calculation for both routines so it is necessary to assign a value to W_A. The molecular system is heated using molecular dynamics to its maximum temperature, which can be up to 4000K. Then the system is cooled in 25K steps to a minimum temperature of 300K. At each temperature reduction stage the system is allowed to reach equilibrium otherwise it would become trapped in a local energy minimum. Integrations of the equations that govern the motions of atoms during molecular dynamics simulations are performed in short time steps of 0.5 fs to keep the system stable. The *slowcool.inp* routine includes a final positional refinement stage to reduce the energy of the model, and the atomic coordinates of the final model are written to a file.

3.3.1.6. *B*-factor refinement

In a crystal, atoms vibrate around an equilibrium position and an atom will therefore occupy different positions in all unit cells. The effect this has on X-ray scattering is akin to an X-ray beam meeting a smeared atom on a fixed position, and as the thermal vibration increases so does the size of the smeared atom. Increasing the size of a smeared atom produces a decrease in the intensity of scattered X-rays, and to compensate for this in X-ray crystallography models the atomic scattering factor of its atoms must be multiplied by a temperature (*B*-) factor. In the MFE-23 refinement, the *B*-factors for all atoms in the initial model were set to 20. As modelling progressed, the *B*-factor for each individual atom was refined using the *brefinement.inp* routine, which employs a Powell conjugate minimization method. At the end of the refinement, the atomic coordinates with modified *B*-factors are written to a file.

3.3.1.7. Map calculation

At the end of a refinement cycle, electron density maps are calculated from the final model. These maps are used for manual rebuilding within the O molecular graphics software. Two types of electron density maps were used for the modelling of MFE-23 and these are $2F_{obs}$ - F_{calc} and F_{obs} - F_{calc} maps; where F_{obs} and F_{calc} are the

structure factor amplitudes determined from the experimental data and the model respectively. Both types of map can be calculated using the X-PLOR *map.inp* routine. It is necessary to convert the format of map files in order for them to be utilized by other programs such as O and CCP4. This can be done using the O-associated MAPMAN program (Kleywegt & Jones, 1995).

3.3.2. O macromolecular modelling software

O is a combination of graphical display program and versatile database system, that was produced primarily for protein crystallography model building. The software is available from Prof. T. A. Jones at the Uppsala University, Sweden, and version 5.10 was used to build the MFE-23 model. O was installed on a Silicon Graphics INDY R4000 workstation, which had a dials box and CrystalEyes three-dimensional viewing glasses (StereoGraphics Corporation, California).

Prior to rebuilding, in case there is an error in the molecular replacement solution, it is important to check the crystal packing does not produce bad contacts between molecules. O requires the input of the unit cell dimensions (Å) and the crystal space group for the calculation of symmetry-related molecules. A representation of the unit cell can be displayed to aid interpretation of the crystal packing. Contoured $2F_{obs}$ - F_{calc} and F_{obs} - F_{calc} electron density maps, which were scaled at 1σ and 3σ respectively were used for MFE-23 modelling. To permit clear visualisation of the model, maps are displayed as caged structures, and maps can be coloured differently for convenient discrimination. Modelling is carried out gradually: as the maps reveal recognizable structural features, the model is modified to fit into the molecular surface implied by the maps. O contains many commands that enable specific regions of the model to be altered and manipulated.

For a molecular replacement model, the sequence of the initial model differs from the actual protein. The *Mutate_replace*, *_delete* and *_insert* commands are used to modify the sequence. It is expected that 90 to 95% of sidechains will have a common rotamer conformation (Kleywegt & Jones, 1995b) and the *Lego_side_chain* command is used to check all residues against a standard rotamer database (Ponder & Richards,

1987) and to select a rotamer if it agrees with the density maps. More drastic structural changes can also be made. The *Tor_residue* command is used to make dihedral rotations around bonds. The *Move_atom*, *_zone* and *_fragment* commands are used to translate atoms or groups of atoms. The *Flip_peptide* command is used to perform a 180° rotation of a peptide group around the α-carbon to α-carbon direction. After a region has been rebuilt it is regularized using *Refi_zone*, which forces standard bond lengths and angles and fixed dihedral angles.

In each cycle of MFE-23 model rebuilding comparisons of the model to the electron density maps were generally performed in an N- to C-terminal direction. OOPS is an O-related program from Uppsala which is used to aid a model building cycle (Kleywegt & Jones, 1996). It generates a residue-by-residue profile of the model quality and highlights regions of the model that may need attention, such as unusual mainchain and sidechain geometries and strange *B*-factor and atom occupancy values. A real-space fit of the model against an electron density map can also be performed using OOPS. This utilizes the *RSR* commands in O (Jones & Liljas, 1984). For a chosen region of the model, electron density is calculated for all atoms within this region and this calculated density is scaled to the $2F_{obs}$ - F_{calc} electron density map. Evaluation of how well the calculated model density fits the observed $2F_{obs}$ - F_{calc} map is then performed using the *RS_fit* command, which computes a correlation coefficient between the two maps (Jones *et al.*, 1991).

3.3.3. Locating water peaks using CCP4

A protein crystal is typically between 25% and 70% water. The bulk of the solvent is assumed to be disordered and therefore it is commonly omitted from model building and refinement procedures. However, a significant number of water molecules are intimately associated with the surface of the protein molecule by means of hydrogen bonds and these water molecules constitute the so-called hydration shell. These water molecules influence the X-ray diffraction pattern of the protein and hence they are included in the modelling procedures. Once model building and refinement has proceeded to a point where no further decrease in the *R*-factor can be obtained, atomic coordinates for water molecules are added to the model. In MFE-23 modelling, water

molecules were identified using the PEAKMAX and WATPEAK programs from CCP4. PEAKMAX is used to locate the coordinates of all peaks in the $2F_{obs}$ - F_{calc} electron density map above a threshold of 3σ. The WATPEAK program is then used to select those peaks that occur within 4 Å of the surface of the protein molecule. The validity of the water molecules is determined by visual inspection in O.

3.4. Homology modelling

Homology modelling offers the best means of obtaining atomic coordinate domain models in the absence of X-ray crystallographic or NMR data (Lattman, 1995). This technique is based on the assumption that domains which appear to share the same ancestor gene will adopt a common fold structure. Needless to say, a homology model will not be as accurate as an experimentally determined structure and, in particular, loop structures are often unreliable. In its favour, homology modelling is able to provide reasonable representations of the core structures of proteins which are suitable for low resolution solution scattering modelling strategies, and the general localisation of residues may be good enough to enable interpretations and predictions of functionality. The approach that is used to obtain an atomic coordinate model from a sequence of interest is outlined in Figure 3.5. Obviously, the crucial stage in this approach is the identification of a structure that has the same fold as the target sequence. In the following description of homology modelling, the fold prediction methods that are used for analogy modelling will also be discussed because they can be beneficial for supporting predictions from sequence similarities alone, especially when sequence identity is in the ambiguous 20 to 30% range.

3.4.1. Sequence analysis

Sequence analysis is typically performed when a protein sequence is first determined by comparisons to homologous proteins, for reason of predicting its function and domain content. Comparisons are made using sequence alignment algorithms that score residue differences between equivalent positions in two sequences. The most commonly used scoring schemes have been derived by examining the substitution frequencies observed in sequence alignments, e.g. the Dayhoff matrices (Dayhoff *et al.*, 1978) and BLOSUM mutation matrices (Henikoff & Henikoff, 1992). Scoring schemes

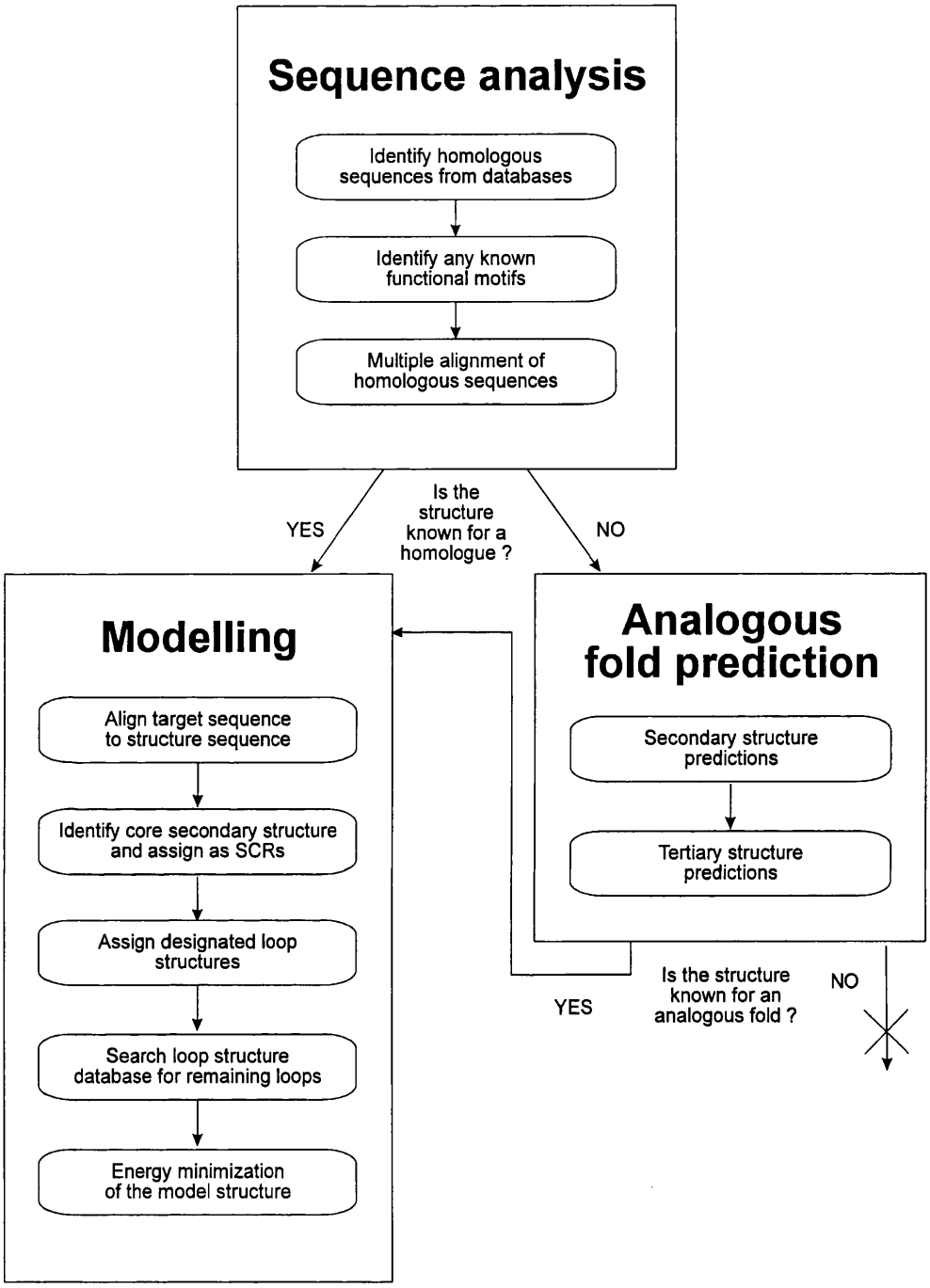


Figure 3.5. Flow chart of the procedures used to generate an atomic coordinate model for a target sequence based on a known structure that has the same fold. There are two alternative routes for identifying a structure with the same fold: either a fold structure is shown to have significant sequence similarity (homology) to the target sequence, or secondary and tertiary structure predictions for the target sequence are used to identify an analogous fold. Once the fold structure has been identified, an atomic coordinate model can be built for the target sequence by means of a rigid fragment assembly method.

based upon the nucleotide base changes required to interconvert the codons for the two residues, or the physico-chemical properties of amino acids have also been used (Barton, 1996). The identification of homologues can be performed using algorithms such as FASTA (Pearson & Lipman, 1988) and BLAST (Altschul *et al.*, 1990), which scan sequence databases for matches to a target sequence. Alternatively, databases such as Entrez store existing sequences with links to their close homologues. However, it may not be possible to find homologous sequences because the residues that are essential for the domain fold only represent a small fraction of the sequence. In such circumstances, short sequence motifs that are characteristic of a protein family or superfamily can be invaluable for predicting homologues. Known sequence motifs are compiled in the Prosite database (Bairoch, 1991).

Next the target sequence is aligned with its homologues. This may fulfil several purposes: for homology modelling the sequence must be correctly aligned against the homologous (template) structure; multiple sequence alignments can be used for secondary structure and tertiary structure predictions if the structure has not been solved for any of its homologues or if homology with a known structure is ambiguous; and multiple sequence alignments may be used to assess the functionality of the target protein. Multiple sequence alignments are conveniently performed using MULTAL (Taylor, 1988) or CLUSTALW (Thompson *et al.*, 1994). The multiple sequence alignments were generally refined manually, especially if the locations of the secondary structure elements had been determined from a homologous structure. MULTAL aligns sequences and alignments using a clustering method. Each sequence is first aligned to all other sequences to obtain pairwise alignments. Then, starting from the pairwise alignment with the highest degree of relatedness, the multiple alignment is constructed by adding to it, in order of decreasing relatedness, those pairwise alignments which have a sequence that overlaps with one of its "free ends". If a pairwise alignment cannot be linked to the multiple alignment for reason of it not containing an overlapping sequence, it is used to start a further alignment. Subsequent pairwise alignments are compared to the "free ends" of all alignments and when a pairwise alignment can join two multiple alignments, they are fused together. This alignment generally produces an ordered list of sequences, but a relatedness cutoff can be used to prevent subfamilies being linked

together. MULTAL is highly interactive and enables the user to alter many of the parameters that control the clustering and the final alignment stages in order to generate an acceptable alignment. CLUSTALW also starts by aligning each sequence to all others to obtain pairwise alignments. These are used to calculate a distance matrix giving the divergence of each pair of sequences and the matrix is used to calculate "guide trees" which describe the evolution of the sequences. The "trees" are then used to determine the progressive alignment of the sequences: starting from the sequences with the highest degree of relatedness at the tips of the "tree" the sequences are aligned in order of decreasing similarity to the roots of the "tree". At the progressive alignment stage, the choice of BLOSUM matrix is varied depending upon the divergence between sequences and gaps in the alignment are favoured more strongly in regions abundant in hydrophilic residues, which are predicted to correspond to loops. A significant advantage of CLUSTALW is that it creates an alignment without the requirement for significant user intervention.

3.4.2. Secondary structure predictions

Secondary structure arrangements (topology) define the folding pattern of a globular protein domain. If the fold structure has not been determined for the superfamily or if the sequences are very divergent, prediction of secondary structure from the amino acid sequence presents a major means for identifying the correct fold. Numerous methods have been published for predicting secondary structure and the following section outlines some of those that are commonly used, and which were instrumental in the correct prediction of the von Willebrand factor type A domain structure (Edwards & Perkins, 1995).

The Chou-Fasman method is a statistical prediction method that was derived from the propensities for each amino acid to occur either in an α-helix or in a β-sheet in 15 protein structures (Chou & Fasman, 1978). According to these propensities, each amino acid is allocated to one of six classes depending on its likelihood of forming an α-helix, which vary from strong α-helix former to strong α-helix breaker, and to one of six classes depending on its likelihood of forming a β-sheet, which range from strong β-sheet former to strong β-sheet breaker. A series of rules is used to assign secondary

structure elements to clusters of probable α-helix and β-sheet residues in an amino acid sequence. Although this method is fairly uncomplicated in its concept, it has been criticised because of its simple statistical approach, its arbitrary prediction rules, and because it does not consider the chemical and physical properties of the amino acids (King *et al.*, 1996). Its accuracy is improved if it is used with a multiple sequence alignment.

The GOR method is a more complex statistical method (GOR-I; Garnier *et al.*, 1978). The method was developed using a database of 26 protein structures, which was later updated for a database containing 75 structures (GOR-III; Gibrat *et al.*, 1987). Each residue is unambiguously assigned to one of 4 possible conformations, α-helix, β-sheet, β-turn (2-residue turn) or random coil. The basis of this method is that the amino acid sequence and the secondary structure are two distinct messages that are related by a translation process that can be examined using information theory. Although in theory the conformation of any particular residue is dependent on every other amino acid in the protein, the most significant influence on the conformation of a residue is exerted by the eight residues either side of it (Robson & Pain, 1971; Robson & Suzuki, 1976). Structure prediction uses the information a residue carries about its own secondary structure, the information a residue carries on the secondary structure of a second residue within eight residues along the sequence that is independent of the second residue's type, and the information a residue carries about the secondary structure of a second residue that depends on the second residue's type. This method is theoretically elegant and it allows the separation of the different types of information involved in the folding of a protein. However, it also neglects the physical and chemical properties of the amino acids and it deliberately neglects protein folding (King *et al.*, 1996). Its accuracy is again improved if is used with a multiple sequence alignment.

PHD is a secondary structure prediction algorithm that is based on a neural net learning system and a multiple sequence alignment (Rost & Sander, 1993; Rost *et al.*, 1994). First, a profile of the frequencies of amino acids occurring at each sequence position is calculated from a multiple sequence alignment and this is then processed by a three-layered network. The first layer is a neural network that has been trained to

classify residues according to three states of secondary structure, α-helix, β-strand and loop. In the second layer, stretches of predicted residues are analysed and contiguous regions of residues that are predicted to have the same structure are assigned as secondary structure elements, and unlikely stretches of secondary structure elements are discarded. At this stage, the agreement of predicted segment lengths with those observed in protein structures is noticeably improved. The third layer averages the predictions from 12 networks that have been trained on different datasets and so acts to reduce the "noise" associated with the predictions. PHD had an accuracy greater than 70% when cross-validated on more than 100 unique structures (Rost & Sander, 1993).

The SAPIENS prediction method is based on the evaluation of mean propensities and environment-dependent substitution tables for amino acids in aligned sequences (Wako & Blundell, 1994a & b). Initially, sequences are considered individually and the preferred secondary structure state (α-helix, β-sheet, buried coil or exposed coil) is assigned to each residue using propensity and substitution tables. These assignments are modified for neighbouring residue cooperativity and according to the positions of residues that are typically found at the N- and C-terminal caps of secondary structure elements. Next, the secondary structure assignments are altered using predicted solvent accessibility patterns, which are compared to those observed for secondary structure elements in known protein structures. Finally, the conformational state at each residue position is averaged across the multiple sequence alignment and the most dominant state is used.

A recent comparison of commonly-used secondary structure prediction methods has been made using the sequence of the adaC O6 Methyl G.DNA Methyltransferase from *Escherichia coli*, which is an α + β protein (King, 1996). At the time the test was performed, the structure of this protein had been determined but not published which ensured that the predictions were performed blind. The secondary structure elements of the structure were determined using DSSP (Kabsch & Sander, 1983; Section 3.5.2). The PHD method was found to be the most successful prediction method with an accuracy of 79%. The comparison also included the other methods discussed above: the SAPIENS method had an accuracy of 64%, the Chou-Fasman method had an accuracy

of 53% and the initial GOR-I method had an accuracy of 51%. However it should be noted that the success of the PHD method may be partly due to its automatic generation of a multiple sequence alignment of homologous proteins, which is used to perform the secondary structure predictions. The other methods used only the target protein sequence, although the use of a multiple sequence alignment would probably lead to an improvement in their predictions. For the prediction of the von Willebrand factor type A domain, not only was a multiple sequence alignment utilized but a consensus of the predictions from these different methods was used. This resulted in high prediction accuracies of about 70% (Edwards & Perkins, 1996).

3.4.3. Tertiary structure predictions

Fold recognition by threading or inverse folding methods attempt to match the one-dimensional information contained in protein sequences directly to three-dimensional fold structures (Bowie & Eisenberg, 1993). The rationale behind these methods is that by aligning a residue from the sequence of interest with a target residue in a known fold structure, the tertiary location of the target residue can specify not only amino acid type but also the secondary structure, the exposure and the spatial interaction with other residues. These techniques enable the detection of more distant sequence-to-structure relationships than sequence comparisons alone, which can fail to recognise highly similar protein structures that have a sequence similarity between 20% and 30%. In general these methods use two principle measures of residue-to-environment compatibility, both of which are based on the analysis of known structures. The first principle is referred to as the solvation preference and it characterises a residue according to its degree of exposure to solvent. Typically, polar residues are expected to occur at exposed sites and apolar residues are expected to occur at buried sites. The second principle or the neighbour preference measure characterises the residue by what type of residues are found nearby. Thus, if a residue is surrounded by apolar residues then the residue is expected to be apolar.

THREADER is a fold recognition program that uses a dynamic programming algorithm based on a combination of neighbour and solvation preferences (Jones *et al.*, 1992b). A library of 102 unique protein fold structures at 2.8 Å resolution or better

was constructed. Each fold is considered only as a chain tracing through space. The test sequence is optimally fitted ("threaded") to the backbone coordinates of each fold and the pseudo-energy of each fitting is evaluated by the summing of pairwise interaction potentials between amino acids. This pseudo-energy evaluation is performed using a set of knowledge-based potentials that are derived from statistical analysis of known protein structures. A measure of the pseudo-energy is provided by considering a pair of atoms at a given residue sequence separation and a specified interaction distance. This relates to the probability of observing the proposed interaction in a native protein structure. These empirical potentials are divided into sequence separation ranges, where it is inferred that short range interactions specify secondary structural elements, medium range interactions specify super-secondary motifs, and long range interactions define tertiary packing. The interaction energies in THREADER are expressed as Z-scores: Z-score = (Energy - mean) / standard deviation, for the pairwise or solvation energies. The pairwise energy Z-score is used for ranking the library folds and the lowest pairwise interaction energy attributed to the most probable match. A Z-score less than -3.5 is regarded as very significant and is probably a correct prediction, but any such prediction needs to be substantiated by other methods.

3.4.4. Model building

Once the fold adopted by a target sequence has been identified its three-dimensional structure can be modelled. INSIGHT II (release 95.0; Biosym/MSI, San Diego, USA) is a suite of molecular graphics and computational chemistry software that was used for building and refining domain models. The homologous or analogous fold structure(s) is used as a template, from which an atomic coordinate model is built by a rigid fragment assembly method that is implemented using the HOMOLOGY program. In this program, the template structure is imported and the alignment of its sequence to the model sequence is recreated.

Before modelling can commence, the structurally conserved regions (SCRs) in the core of the fold must be defined. SCRs were commonly defined as the α-helix and β-sheet elements identified in the template structure by DSSP (Kabsch & Sander, 1983; Section 3.5.2). Alternatively, SCRs could be defined as a consensus between the

template fold secondary structure and secondary structure predictions for the target sequence. The SCR definitions are assigned to the sequence alignment in HOMOLOGY. The atomic coordinates of the template SCRs are then copied to the model, except when differences occur between the sidechain atoms of the two proteins, in which case these are modelled from a library of amino acid structures.

The loops which join the SCRs are modelled next. The conformations of SCRs are restricted by hydrogen bonding constraints, but such conformational restriction is not usually observed for loops and they therefore exhibit a higher degree of sequence and structural divergence between homologous proteins. Consequently loops are more difficult to model. Where possible loops from the template structure were used, but this approach is only applicable when the length of a loop in the template structure is identical to the corresponding loop in the target protein. This type of loop is termed a designated loop in HOMOLOGY and the template coordinates for designated loops are copied to the target model, except when differences occur between sidechain atoms, in which case these are modelled from library structures.

Other loops are termed searched loops, and these occur when the loop length is different between the template and target proteins. Searched loops are modelled using compatible loop structures (Hobohm & Sander, 1994). HOMOLOGY provides an algorithm that enables the database to be searched for loops that have the desired length and which best satisfy the geometric constraints required to join two SCRs. This search algorithm calculates an α-carbon distance matrix for the model residues on either side of a loop and this is compared to α-carbon distance matrices from structures in the database. The database loop structures that give the best fit to a model loop are defined as those with the lowest root mean-squared (RMS) distance values:

$$RMS\ distance = \left(\sum_{i=1}^{N} \frac{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}{N} \right)^{1/2} \qquad \text{Eq. 3.12.}$$

where the summation is performed for 1 to N pairs of α-carbon atoms between the two

structures; and x, y, z and x_0, y_0, z_0 are the atomic coordinates of each corresponding α-carbon atom between the two structures. The algorithm reports the ten best loops from a search. The number of residues before and after the loop in the model that are considered in the search can be varied to obtain different results. It is also possible to vary the orientation in which a loop is joined to the model: the searched loops are superimposed onto the model structure and either the α-carbon atoms of all the residues that were used to perform the search, or just the α-carbon atoms of the two residues that are adjacent to the loop can be selected for the superimposition. These two sources of variation allowed numerous loop structures to be tested in a model. The selection of a loop structure is typically governed by conditions such as; suitable mainchain torsion angles and distances between α-carbon atoms at the points that loops are joined to SCRs; the loop does not contain secondary structure; there is no gross steric overlap of the loop with other regions of the model. When a searched loop is selected, its coordinates are copied to the model, except when differences occur between sidechain atoms and these are modelled from library structures.

After atomic coordinates have been assigned to every model residue, manual rotamer searches can be used to relieve steric overlap. HOMOLOGY contains a library of commonly occurring sidechain rotamers and these are tested individually in an attempt to reduce the steric overlap of poorly modelled sidechains.

3.4.5. Model refinement

A model that has been built using a rigid fragment assembly method may contain structural artefacts. These include the substitution of large sidechains for small ones, strained peptide bonds between segments taken from different reference proteins, and non-optimum conformations for the loops. In order to overcome these artefacts, models can be refined using energy minimization.

DISCOVER is a molecular simulation program within INSIGHT II that can perform many routines, including energy minimisation, template forcing, torsion forcing, and dynamic trajectories as well as the calculation of interaction energies, derivatives, mean square displacements, and vibrational frequencies (DISCOVER

2.9.7/95.0/3.0.0 user guide). The consistent-valence forcefield was used to represent the potential energy of the molecular system. This forcefield contains terms which describe the energies necessary to stretch a bond, distort the angle between three atoms, rotate atoms about their bond axis, and move an atom out of the plane defined by the three atoms to which it is bonded, as well as extra terms that describe the energies representing the coupling effects of one of the above energies with another (cross terms), and associated with the attractive, repulsive and electrostatic forces between atoms that are not bonded to one another (charges). Energy minimization is used to produce a model that is chemically and conformationally reasonable. DISCOVER allows the minimization algorithm to be specified and choices include steepest descent and conjugate gradients (Section 3.3.1.4). Other options include the number of minimization steps, and whether cross terms and charges are included in the evaluation.

HOMOLOGY is used to call up several DISCOVER routines that have been developed specifically for minimizing the energy of protein models. The first stage of protein minimization would commonly be *splice repair*. This minimizes the energy of the peptide bonds at the junctions between regions from different reference structures. Bond lengths and mainchain omega torsion angles are displayed by the program so that the progression of the minimization can be monitored. The *relax* option enables different minimization strategies to be set up: the model is divided into individual fragments from different reference structures, SCRs and loops, mainchains and sidechains, and mutated sidechains and non-mutated sidechains, and from these different regions can be selected for minimization. It was usual to test numerous combinations of minimization strategies and the resulting models could be assessed by comparing bond lengths, bond angles and torsion angles to consensus values derived from X-ray crystal protein structures.

3.5. Structural analysis of models

3.5.1. Structure validation

After building an atomic coordinate model, whether by means of X-ray crystallography, NMR or homology modelling techniques, it is important to assess its quality. PROCHECK (Laskowski *et al.*, 1993) is a suite of programs that assesses a

PDB-format atomic coordinate file using stereochemical parameters derived from high-resolution protein crystal structures (Morris *et al.*, 1992), and bond lengths and bond angles derived from a comprehensive analysis of small-molecule structures (Engh & Huber, 1991). The stereochemical quality of the model is output as a residue-by-residue listing that enables the clear identification of regions that are in error. A useful feature of PROCHECK is that it produces a Ramachandran plot of the ϕ and ψ mainchain torsion angles (Ramachandran & Sassiekharan, 1968). PROCHECK was routinely used to validate models in this thesis (Chapters 4, 5, 6 and 7).

3.5.2. Secondary structure

Knowledge of a protein's secondary structure can be used for several purposes. In homology and analogy modelling strategies, secondary structure is used to define the core of the reference protein structure. Determination of a modelled protein's secondary structure will enable its structure to be compared to the reference structure, and this is especially useful in guarding against poor energy minimization strategies. More generally, the identification of significant conserved and divergent features between homologous fold structures may require stringent definitions of helices, strands and loops within domains and linkers between domains. It is therefore necessary to have consistent definitions of secondary structure elements.

DSSP identifies secondary structure in a protein model according to standard, unambiguous definitions of secondary structure classes (Kabsch & Sander, 1983). Secondary structure recognition is based mainly on mainchain hydrogen-bonding patterns. The program assigns a hydrogen bond when the electrostatic interaction energy between a hydrogen-bond donor and a hydrogen-bond acceptor is less than -0.5 kcal $mole^{-1}$. The interaction energy is a function of the distance d and the angle θ between the CO and NH groups. An ideal hydrogen bond has an energy of -3.0 kcal $mole^{-1}$ at $d = 2.9$ Å and $\theta = 0°$ between the two groups. The cutoff of -0.5 kcal $mole^{-1}$ allows misalignment of up to 63° at the ideal distance, and an N to O distance of up to 5.2 Å for perfect alignment. DSSP implements a secondary structure recognition algorithm that defines eight classes of secondary structure based on two elementary hydrogen-bonding patterns: "*n*-turns" with a hydrogen bond between the CO of residue

i and the NH of residue *i* + *n*, and "bridges" with hydrogen bonds between residues not near each other in sequence. An isolated "*n*-turn" is defined as a turn (T). Two or more consecutive "*n*-turns" are defined as helices; a 3_{10} helix for *n*=3 (G), an α-helix for n=4 (H), and a π-helix for *n*=5 (I). Two residues *i* + *j* that form an isolated "bridge" are termed B, whereas consecutive bridges of identical type form extended β-strands (E). DSSP also defines regions of high mainchain curvature, or bends. A bend (S) at *i* requires a curvature of at least 70°, where curvature is defined by the angle between the vector joining the carbon atoms $C^{\alpha i}$ to $C^{\alpha i-2}$ and the vector joining $C^{\alpha i+2}$ to $C^{\alpha i}$. Unclassified regions are termed coil (C).

3.5.3. Accessible surface area

As a rule, the structure of a globular protein domain is partitioned into a hydrophobic core and a polar, hydrophilic surface. The term "accessible surface area" is used to refer to the exterior of the protein that is in contact with solvent and its calculation has been used to consider the hydrophobic effect in many aspects of protein modelling. It is also possible to quantify the area that is buried at the interface between protein molecules or domains using the same calculation. The accessible surface is calculated as the surface that is traced by the centre of a probe molecule as it rolls on the van der Waals surface of the protein molecule (Lee & Richards, 1971). A sphere of radius 1.4 Å is commonly used as the probe, because its dimensions approximate to that of a water molecule. PSA, which is part of the of the COMPARER suite of programs, was used to calculate the accessible surface area of protein models (Sali & Blundell, 1990). It uses the algorithm of Richmond & Richards (1978) to calculate the accessible surface area and sidechain accessibility calculations and normalisation is carried out as described by Hubbard & Blundell (1987).

3.5.4. Surface electrostatic potentials

Electrostatic interactions play a central role in a variety of biological processes (Honig & Nicholls, 1995). DELPHI is part of the INSIGHT II suite of programs and is used to calculate the electrostatic properties of charged molecules. The electrostatic potential in and around macromolecules can be calculated using a finite difference solution to the Poisson-Boltzmann equation. DELPHI permits the ionic strength and the

dielectric constants of both the solvent and the protein molecule to be varied. The output from DELPHI can be mapped onto the molecular surface of the protein via its interface with INSIGHT II.

The classical treatment of electrostatic interactions in solution is based on the Poisson-Boltzmann equation:

$$\nabla\cdot[\varepsilon(\mathbf{r})\nabla\cdot\phi(\mathbf{r})] - \varepsilon(\mathbf{r})\kappa(\mathbf{r})^2 \sinh[\phi(\mathbf{r})] + 4\pi\rho^f(\mathbf{r})/kT = 0 \qquad \text{Eq. 3.13.}$$

where $\phi(\mathbf{r})$ is the electrostatic potential at any point in space, and has units of kT/e (k is the Boltzmann constant; T is the absolute temperature and e is the elementary charge); ε is the dielectric constant; and ρ^f is fixed charge density (in proton charge units). The term $\kappa^2 = 1/\lambda^2 = 8\pi q^2 I/ekT$, where λ is the Debye length and I is the ionic strength of the bulk solution. The variables ϕ, ε, κ and ρ are all functions of the vector $\mathbf{r}$.

The molecular surface is defined as the contact surface formed between the van der Waals envelope of the molecule and a probe molecule of 1.4 Å radius. All internal regions are assigned a low value of ε (around 2-4), whereas exterior regions are assigned the standard dielectric constant of water (ε of around 80). Using iterative procedures for the solution of the above equation, $\phi(\mathbf{r})$ can be calculated for a molecule in solution of arbitrary ionic strength. In the context of proteins, unique patterns of $\phi(\mathbf{r})$ are seen that in many cases have an important functional role (Honig & Nicholls, 1995).

Chapter 4

Multi-domain Structure of Human Carcinoembryonic Antigen by X-ray and Neutron Scattering

4.1. Introduction

There has been extensive clinical interest in the human carcinoembryonic antigen (CEA) since it was first described by Gold and Freedman (1965) as an antigen that is present in the foetal gut, absent from normal adult tissues, but re-emerging in certain carcinomas. It is now known that CEA is actually expressed on a number of normal adult tissues and this expression is up-regulated on a wide variety of tumours including colon, ovarian, breast, stomach, pancreas and lung tumours (Thompson, 1995; Wittekind, 1995). Today, the difference in CEA expression between normal and cancer cells forms the basis of many strategies for the targeting, monitoring and treatment of tumours, and this is especially important for colon cancer. Also it has been revealed in recent years that, rather than being just a passive tumour marker, CEA possesses numerous and wide-ranging properties, where is belongs to a family of closely-related proteins, with cell adhesion and signalling functions, and binds to microorganisms. These suggest complex roles for this molecule in both normal and carcinoma tissues.

4.1.1. The CEA molecule

Early biochemical characterizations of CEA demonstrated it to be a heavily glycosylated protein with a commonly assumed M_r of 180 000, of which more than 50% is carbohydrate. The amino acid sequence of CEA was determined simultaneously by cDNA and polypeptide sequencing techniques (Oikawa *et al.*, 1987; Paxton *et al.*, 1987; Figure 4.1). The sequence showed that CEA has a 34 residue N-terminal peptide, which is cleaved after signalling the transport of CEA out of the cytosol. The remainder of the sequence consists of 668 residues, which can be divided into an N-terminal region of 108 residues, followed by three homologous 178 amino acid repeats, and a hydrophobic C-terminal peptide of 26 residues. Each of the three homologous repeats consists of a 95 residue domain followed by an 83 residue domain. When discussing the seven domain structure of CEA, the three homologous repeats are typically referred to as I, II and III, and the two domains of each repeat are distinguished by designating the 95 residue domain as subtype A and the 83 residue domain as subtype B (Figure 4.1). The C-terminal peptide is post-translationally replaced by a glycosyl phosphatidyl inositol (GPI) anchor for attachment to membranes (Jean *et al.*, 1988; Hefta *et al.*, 1988). Like

(a)

(b)

```
                5   10   15   20   25   30
                ....|....|....|....|....|....
Leader      : MESPSAPPHRWCIPWQRLLLTASLLTFWNPPTTA

                5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95  100  105
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
N           : KLTIESTPFNVAEGKEVLLLVHNLPQHLFGYSWYKGERVDGNRQIIGYVIGTQQATPGPAYSGREIIYPNASLLIQNIIQNDTGFYTLHVIKSDLVNEEATGQFRVYP

                5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
IA          : ELPKPSISSNNSKPVEDKDAVAFTCEPETQDATYLWWVNNQSLPVSPRLQLSNGNRTLTLFNVTRNDTASYKCETQNPVSARRSDSVILNVLYG
IIA         : EPPKPFITSNNSNPVEDEDAVALTCEPEIQNTTYLWWVNNQSLPVSPRLQLSNDNRTLTLLSVTRNDVGPYECGIQNELSVDHSDPVILNVLYG
IIIA        : ELPKPSISSNNSKPVEDKDAVAFTCEPEAQNTTYLWWVNGQSLPVSPRLQLSNGNRTLTLFNVTRNDARAYVCGIQNSVSANRSDPVTLDVLYG

                5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
IB          : PDAPTISPLNTSYRSGENLNLSCHAASNPPAQYSWFVNGTFQQSTQELFIPNITVNNSGSYTCQAHNSDTGLNRTTVTTITVYA
IIB         : PDDPTISPSYTYYRPGVNLSLSCHAASNPPAQYSWLIDGNIQQHTQELFISNITEKNSGLYTCQANNSASGHSRTTVKTITVSA
IIIB        : PDTPIISPPDSSYLSGANLNLSCHSASNPSPQYSWRINGIPQQHTQVLFIAKITPNNNGTYACFVSNLATGRNNSIVKSITVSA

                5   10   15   20   25
                ....|....|....|....|.
C-terminal  : SGTSPGLSAGATVGIMIGVLVGVALI
```

Figure 4.1. Domain structure and amino acid sequence of the human CEA molecule. (*a*) A schematic representation of the linear arrangement of domains in the CEA molecule. CEA has nine distinct regions: a 34-residue leader sequence (grey); a 108-residue N-terminal domain (yellow); followed by three homologous regions I, II and III, each of which consists of a 94-residue domain and an 84-residue that are designated domains A (green) and B (magenta) respectively; and a 26-residue hydrophobic membrane-attachment C-terminal peptide (black) which is post-translationally replaced by a GPI anchor. (*b*) The CEA sequence. This sequence is the CEA entry from the SWISSPROT sequence database (accession code: P06731). The sequence is divided into the segments defined in (*a*), and the A domains and B domains are aligned independently to demonstrate the homology between these regions. Putative glycosylation sites are shown in boldface and underlined.

many other cell surface molecules, CEA is predicted from its sequence to belong to the immunoglobulin (Ig) superfamily (Williams, 1987; Williams & Barclay, 1988; Chapter 1). The N-terminal region is homologous to V-set Ig-fold domains, but lacks the disulphide bridge between β-strands B and F. Both the A- and B-domains of CEA were classified as having the C2-set topology (Williams & Barclay, 1988; see Footnote). The amino acid sequence contains 28 putative N-linked glycosylation sites. The N-linked glycosylation sites are not evenly distributed over the protein domains of CEA, and the seven domains have 2, 5, 6, 4, 3, 5 and 3 N-linked sites respectively, in order from the N-terminal domain to domain IIIB. The CEA carbohydrate comprises GlcNAc, Man, Gal, Fuc and NeuNAc residues, and the oligosaccharide chains are mainly complex-type (Chandrasekaran *et al.*, 1983; Hammarstõm *et al.*, 1975; Kessler *et al.*, 1978; Pavlenko *et al.*, 1990; Slayter & Codington, 1973; Yamashita, *et al.*, 1987, 1989).

CEA particles have been visualised by electron microscopy and shown to have a twisted, rod-like morphology (Slayter & Coligan, 1975). A model of CEA has been produced using its homology to the structures of the Bence-Jones dimer REI, the N-terminal pair of domains in human CD4 and the N-terminal domain of rat CD2, and the linker between the V_L and C_L domains of the human New IgG1 Fab fragment (Bates *et al.*, 1992). This model predicted that CEA has an extended structure with its oligosaccharide chains lying in a compact conformation, covering the surface of the protein, and may be dimeric by analogy with antibody structures.

4.1.2. The human CEA gene family

There exists a number of antigens that possess high sequence similarity to CEA, and consequently these antigens cross-react with CEA. Together they form a subset of the Ig superfamily which is commonly referred to as the CEA gene family. In humans, the CEA gene family comprises approximately 29 closely-related genes (Teglund *et al.*, 1994), of which 17 are active genes and 12 are pseudogenes (Thompson, 1995). These

---

Footnote: Since that time, it was predicted that the A-type CEA domains actually belong to the I-set of Ig-fold structures (Harpaz & Chothia, 1994), and this is returned to in Chapter 7 below.

genes are clustered together in a region that spans approximately 1.1 to 1.2 Mb on the long arm of chromosome 19 (Zimmermann *et al.*, 1988; Thompson *et al.*, 1990). Sequence comparisons identify subgroups within the human CEA gene family. Conventionally, it is divided into two subgroups, the CEA subgroup and the pregnancy-specific glycoprotein (PSG) subgroup (Thompson *et al.*, 1991), although a third subgroup has been identified recently (Teglund *et al.*, 1994). Members of the same subgroup possess greater than 70% sequence identity, while the identities between members of different subgroups are generally between 50 and 70%. Table 4.1 summarises the members of the human CEA gene family. The CEA subgroup contains CEA, non-specific cross-reacting antigen (NCA; von Kleist *et al.*, 1972), biliary glycoprotein (BGP; Svenberg, 1976) and the CEA gene family members CGM1, CGM2 and the CGM6 to CGM12 genes (Thompson *et al.*, 1989; Thompson *et al.*, 1991; Khan *et al.*, 1992a). The PSG subgroup consists of the genes for PSG1 to PSG8 and for PSG11 to PSG13 (Thompson *et al.*, 1989; Thompson *et al.*, 1991; Khan *et al.*, 1992b). The third subgroup consists of the CGM13 to CGM18 pseudogenes (Teglund *et al.*, 1994).

It is convenient to describe the structures of CEA gene family members in terms of the arrangements of domains in CEA. For the human CEA gene family, the predicted polypeptide products of the CEA subgroup and PSG subgroup genes and pseudogenes all contain an N-terminal V-set Ig-fold domain which is followed by variable numbers of A- and B-type domains. In contrast, all members of the third subgroup lack the N-domain but do contain exons for A- and B-domains. The number of CEA-related antigens is increased by alternative splicing of the CEA family gene products. For example, there are at least 13 splice-variants of human BGP (Barnett *et al.*, 1993). Post-translational modifications such as variable glycosylation and proteolytic products may also increase the diversity of the CEA family of antigens. CEA-related antigens are not restricted to humans and, in particular, numerous examples have been identified in rodents (Thompson *et al.*, 1991). Although the rodent CEA-like antigens can also be divided into sequence-based subgroups, sequence identities cannot be used to determine their direct counterparts in humans. Indeed, certain rodent CEA-like antigens contain multiple copies of the N-terminal type domain (Rebstock *et al.*, 1990; Keck *et al.*, 1995),

Table 4.1. The human CEA gene family[a]

CEA subgroup	PSG subgroup	Third subgroup
BGPa,b,c,d,e,f,g,h,i,x,x',y,z (CD66a)[b]	PSG1a,b,c,d,e,f	*CGM13*
CGM6 (CD66b)	PSG2a	*CGM14*
NCA (CD66c)	PSG3m	*CGM15*
CGM1a,b,c (CD66d)	PSG4a	*CGM16*
CEA (CD66e)	PSG5n,m	*CGM17*
CGM2	PSG6r,s	*CGM18*
CGM7	PSG7	
CGM8	*PSG8*	
CGM9	PSG11s,w	
CGM10	PSG12	
CGM11	PSG13	
CGM12		

[a] The genes and pseudogenes of the human CEA-gene family are listed in their respective subgroups. Pseudogenes are shown in italics. Where splice variants are known for the products of these genes, they are listed as lower-case letters (Adapted from Teglund *et al.*,1994 and Stanners *et al.*, 1995).

[b] The corresponding CD definitions for the protein products of these genes are shown in parentheses (Skubitz *et al.*, 1995a).

a phenomenon that is not observed for the human CEA family.

Protein products have not been identified for all of the CEA gene family members. Release 35.0 of the SWISSPROT sequence database (November 1997) contained entries for twelve human CEA family proteins, and the CEA homologues BGP and CEA10 from mouse and CCAM105 from rat (Figure 4.2). The CEA subgroup proteins are all cell-surface molecules. However, it should be noted that during instances of cancer, the increased expression of CEA results in a secreted form that is detectable in the patient's serum. It is possible to further divide the CEA subgroup of proteins according to the means by which they are attached to the membrane. CEA, NCA (Barnett *et al.*, 1988) and CGM6 (Berling *et al.*, 1990) are attached to the cell membrane by a GPI anchor. BGP (Hinoda *et al.*, 1988) and CGM1 (Nagel *et al.*, 1993) are attached to the cell membrane by a transmembrane peptide region (Figure 4.2). The GPI-anchored molecules have only been identified in humans, whereas transmembrane peptide anchored molecules are found in rodents as well as humans. It is therefore proposed that the GPI-anchored CEA-molecules evolved later (Rojas *et al.*, 1996) and this proposal is supported by the alignment of CEA family nucleotide sequences (Thompson *et al.*, 1991).

In some respects, the GPI- and transmembrane-anchored CEA family proteins have opposite functions. For example, while CEA and NCA are up-regulated in a number of cancers, BGP is down-regulated (Rojas *et al.*, 1996). Further examples are outlined in the following sections. The PSG subgroup proteins PSG1 (Watanabe & Chou, 1988), PSG1a (Streydio *et al.*, 1988), PSG1c (Streydio *et al.*, 1988), PSG2n (Streydio *et al.*, 1988), PSG4 (Thompson *et al.*, 1989), PSG6 (Zimmermann *et al.*, 1989) and PSG11s (Arakawa *et al.*, 1990) are secreted from cells, most notably into the serum by the placenta during pregnancy (Figure 4.2).

All of the human CEA family proteins have a single N-domain and in all of these proteins this domain lacks the disulphide bridge between β-strands B and F. The numbers of A- and B-type domains vary in the human CEA family proteins from zero in CGM1 to six in CEA. The alternation of A and B domains that is observed in CEA

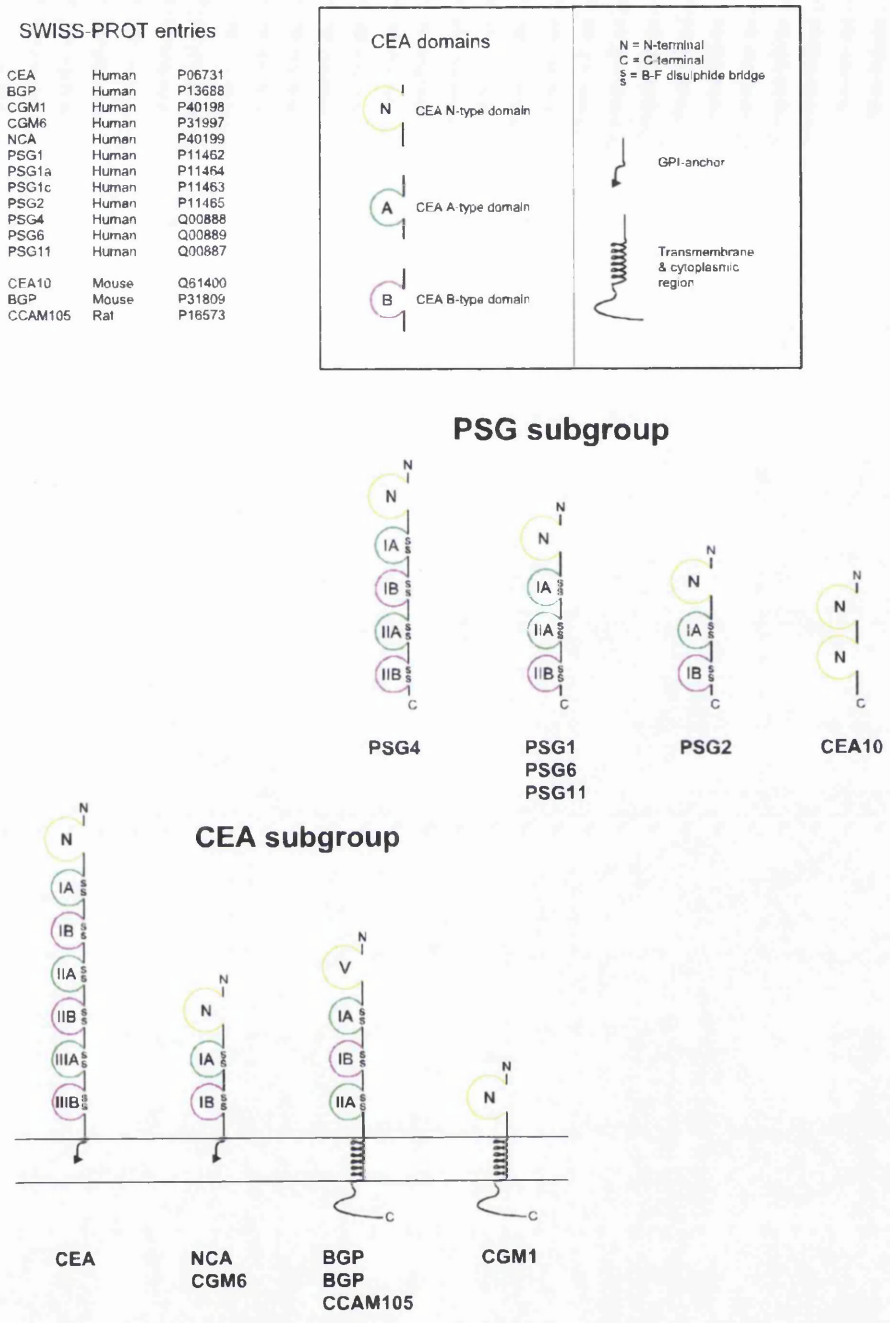


Figure 4.2. The known proteins of the CEA-gene family. These proteins were identified from the SWISSPROT database (Release 35.0; November, 1997; Bairoch & Apweiler, 1997), and the accession codes and source of each protein is listed. The arrangements of the Ig-fold domains in each protein are shown. Domains are labelled according to whether they have homology to the N-, A- or B-domains of CEA (see text). The N- and C-termini of the proteins are indicated and the common B-F disulphide bridge is labelled if it is present. The CEA subgroup of proteins are attached to the membrane by either a GPI-anchor or a transmembrane peptide. The PSG subgroup of proteins and the mouse CEA10 protein are secreted.

is conserved for the CEA subgroup proteins but not for PSG1, PSG1a, PSG1c and PSG11s, which have two consecutive A domains (Figure 4.2). Both mouse BGP (Dveksler *et al.*, 1993) and rat CCAM105 (Lin & Guidotti, 1989) are similar to human BGP, in that they have the same domain arrangements and are attached to the cell-membrane through a transmembrane peptide. The mouse CEA10 molecule (Keck *et al.*, 1995), which is akin to the human PSG proteins, is secreted but unlike all human CEA proteins it contains two domains homologous to the N-terminal domain of CEA (Figure 4.2).

4.1.3. Biological functions of CEA

4.1.3.1. CEA is a cell adhesion molecule

In their analysis of cell-surface members of the immunoglobulin superfamily, Williams & Barclay (1988) observed that those which contain predicted C2-set Ig-fold domains commonly act as cell adhesion molecules. This observation is true for CEA, which has both homophilic (Benchimol *et al.*, 1989) and heterophilic cell adhesion properties *in vitro* (Oikawa *et al.*, 1989). Additional members of the CEA family have also been demonstrated to possess cell adhesion functions. These include human NCA (Oikawa *et al.*, 1989), BGP (Rojas *et al.*, 1990) and CGM6 (Oikawa *et al.*, 1991), and rat CCAM105 (Sippel *et al.*, 1996). The N-domain has been implicated in the homophilic interactions of CEA family proteins (Oikawa *et al.*, 1991). Recombinant forms of the isolated CEA N-terminal domain, expressed either in HeLa cells (Hefta *et al.*, 1992) or in *E. coli* (Krop-Watorek *et al.*, 1998), associate to form oligomers (up to pentamer). Because proteins expressed in *E. coli* are unglycosylated, this would suggest that the carbohydrate chains on the CEA N-domain are not required for its self-association. The importance of the N-domain polypeptide sequence for CEA family homophilic interactions has been confirmed by mutation studies on rat CCAM105 and CEA. CCAM105 has ecto-ATPase activity, and accordingly it contains a consensus ATPase sequence motif (GPAYSGRET) in its N-domain. This sequence motif is almost completely conserved in the N-domain of CEA (in Figure 4.1 this motif starts at residue 58 of the N-domain). Mutation of the arginine residue in this motif to an alanine not only abolishes the ecto-ATPase activity of CCAM105 (Sippel *et al.*, 1994), but it also abrogates the adhesion properties of both CCAM105 and CEA (Sippel *et al.*, 1996).

Although the N-domain is essential for the homophilic interaction of CEA molecules, it is not generally considered to be the sole contributor to this function (Oikawa *et al.*, 1991). Based on a study of the interactions between truncated forms of the CEA molecule, Zhou *et al.* (1993) have proposed that normal CEA homophilic interaction involves the reciprocal bonding between the N-domain and the IIIA-domain on separate molecules, and that less favourable bonding can also occur between the N-domain and either the IA- or IIA-domains. An interesting feature of the cell adhesion properties of the human CEA family is that the two divisions of the CEA subgroup behave differently (Rojas *et al.*, 1996; Section 4.1.2). The cell-adhesion mediated by BGP, which contains a transmembrane peptide, is reversibly dependent on Ca^{2+} or Mg^{2+}, is temperature-dependent and ATP-inhibitable, whereas the adhesions mediated by CEA and NCA, which are GPI-anchored, are opposite to BGP in all of these aspects. CEA undergoes heterophilic cell adhesion interactions with other CEA family of proteins (Oikawa *et al.*, 1989).

4.1.3.2. The role of CEA homophilic adhesion in carcinogenesis

In vitro, homophilic interactions cause aggregation of CEA-expressing cells (Benchimol *et al.*, 1989). The precise biological effects of this mode of interaction can only be speculated on at present. However consideration of the general functions of cell adhesion molecules does provide some insights. The term cell adhesion molecule typically refers to those cell surface molecules which have the ability to form interactions with either molecules on adjacent cells or with the extracellular matrix, and encompasses several classes of molecule in addition to immunoglobulin superfamily proteins; these being the integrins, the selectins and the cadherins. As a whole, adhesion molecules perform a number of functions in health and disease, which include pattern formation and morphogenesis during embryonic development, maintaining the three-dimensional organization of tissues and organs, wound repair, regulation of immune responses, migration of inflammatory cells from blood vessels into inflamed tissues, and tumour growth and metastasis. Cell adhesion typically has an associated cell-signalling function which acts to couple adhesion to other cell functions. For recent reviews of cell adhesion molecules and their functions the reader is referred to Gumbiner (1996), Ruoslahti and Öbrink (1996) and Buckley and Simmons (1997). The predominant

expression patterns of CEA, namely in the embryonic intestine and in colonic tumours, are compatible with involvements of CEA in embryonic development, and in colon carcinogenesis (Benchimol *et al.*, 1989). Recently, increased homophilic adhesion has been demonstrated to directly increase the metastatic potential of CEA-expressing cells to the liver (Yoshioka *et al.*, 1998). This might be due to a clustering of tumour cells, which is promoted by homophilic interactions between CEA molecules, increasing their chances of survival in the circulation. It is also possible that an up-regulation of CEA will produce cell adhesion interactions that disrupt the normal cell-to-cell and cell-substratum interactions which are required for correct tissue architecture. Indeed, the cell-surface expression of CEA produces a change in cell morphology and mediates random, multilayered associations between tumour cells, rather than the monolayer that is formed by normal colon epithelial cells (Grimm & Johnson, 1995; Yan *et al.*, 1997).

4.1.3.3. Additional roles of CEA in carcinogenesis

The role of CEA in carcinogenesis appears to extend further than its cell adhesion properties. One significant factor is the presence of a CEA receptor in Kupffer cells from the liver (Gangopahyay & Thomas, 1996; Gangopahyay *et al.*, 1996a, 1996b). Given that CEA is involved in the development of liver metastases from colon cancer (Wagner *et al.*, 1992), evidence is accumulating to support the hypothesis that the interaction of CEA with this receptor influences the metastatic process (Gangopahyay *et al.*, 1996b). This receptor binds the peptide sequence PELPK, which forms the linker between the N– and AI-domains of CEA (in Figure 4.1 this motif starts at residue 108 of the N-domain and stretches to residue 4 of the AI-domain). Ligation of this receptor leads to the secretion of cytokines, such as Il-1α, IL-1β IL-6 and TNF-α, by Kupffer cells and it is suggested that these cytokines stimulate the production of adhesion molecules by the liver endothelium which results in the retention of tumour cells in the liver (Gangopahyay *et al.*, 1996b).

CEA also has an effect on malignant cellular transformation. The transformation of normal cells to tumour cells can be divided into four events, these being increased proliferation, prevention of apoptosis, immortalization, and loss of differentiation. Both CEA and NCA expression levels are inversely correlated with the degree of cell

differentiation in colon tumours and it is therefore possible that CEA and NCA contribute directly to carcinogenesis by inducing the blocking colonocyte differentiation (Ilantzis *et al.*, 1997). Interestingly, BGP which is down-regulated in tumour cells does not inhibit differentiation. Screaton *et al.* (1997) have demonstrated that CEA expression also inhibits the differentiation of rat myoblasts and the action of CEA was found to be cooperative with the action of the oncogenes v-Myc and Bcl-2. V-myc functions to increase the rates of cell division rate, apoptosis and differentiation, while Bcl-2 functions to block the apoptosis signal of v-Myc. CEA produces a dominant signal that blocks differentiation. The three proteins are thus considered to act in concert to promote malignant transformation. These authors propose that CEA may be considered as a new class oncogene, which is capable of blocking cell differentiation. The cell-surface expression of CEA also leads to an up-regulation in the expression of specific, unrelated cell adhesion molecules in a melanoma cell line, which may further contribute to the general deregulation of adhesive interactions in tumour cells (Grimm & Johnson, 1995).

4.1.3.4. Interaction of CEA with microorganisms

The discussion so far has pointed to roles for CEA in tumour development. However, CEA is also expressed on normal cells. Normal colonic epithelial cells have a polarized structure, in which CEA is localized to the face lining the lumen of the large intestine. Because cell adhesion interactions between these cells occur mainly in a sideways direction, it is hard to reconcile the localization of CEA with a cell adhesion function. It is therefore predicted that CEA has an alternative biological function in normal colon epithelia. Leusch *et al.* (1990, 1991) demonstrated that both CEA, NCA and BGP bind to *E. coli* strains of human origin. These authors have therefore proposed that CEA family proteins may regulate bacterial colonization of the body. In the case of CEA, regulated colonization of the gut by *E. coli* could provide a mechanism for protecting against infection by pathogenic bacteria. Conversely, the CEA family of proteins also act as receptors for invasive gonococcal bacteria *in vitro*. This interaction is mediated by gonorrhea opacity (Opa) proteins, and the OpaA protein is capable of binding to CEA, CGM1, CGM6, NCA and BGP (Bos *et al.*, 1997; Chen *et al.*, 1997b). The N-domain of CEA family proteins appears to be necessary for the interaction with

Opa proteins, in particular the first 59 residues of this domain (Virji *et al.*, 1996; Gray-Owen *et al.*, 1997). The case of human phagocytes, which express BGP, CGM6 and NCA, and are susceptible to invasion by *Neisseria gonorrhoeae*, is of particular clinical relevance. The phagocytosis of bacteria by these cells is a crucial event in combatting infection. The classical activation of phagocytosis is via the Fc-γ-receptor which recognises specific antibodies bound to the bacterial surface. However, if phagocytosis is mediated by CEA family proteins, the signalling cascade that is initiated is different to that activated by the Fc-γ-receptor, and it is therefore speculated that *N. gonorrhoeae* exploit this mechanism to infect phagocytes, while avoiding their normal bactericidal mechanisms (Hauck *et al.*, 1998). A further example of CEA family proteins acting as receptors for microorganisms is the mouse BGP, which is the receptor for mouse hepatitis virus (Dveksler *et al.*, 1993). This virus can also infect primates causing lesions of the central nervous system. CEA family proteins in primates are therefore likely candidates for a mouse hepatitis virus receptor. Studies have recently shown that mouse cells lacking functional mouse BGP become susceptible to mouse hepatitis infection only after transfection with plasmids expressing human CEA or BGP (Chen *et al.*, 1997a). The N-domain of CEA family proteins is involved in many of their functions and unsurprisingly it is also implicated in binding to this virus (Chen *et al.*, 1997a).

4.1.3.5. Cell signalling functions of CEA

The actions of CEA in blocking cell differentiation, altering cell morphology and promoting protein expression indicate that it possesses an associated signalling function. For the CEA family protein BGP, interactions could occur between its cytoplasmic region and cytoplasmic protein kinases. Indeed, the BGP cytoplasmic region contains the sequence motif YXXL which is known to be critical for signal transduction in antigen receptors (Reth, 1989; Weiss, 1993), its cytoplasmic domain undergoes tyrosine phosphorylation (Skubitz *et al.*, 1992), and BGP associates with Src-like tyrosine kinases (Brummer *et al.*, 1995; Skubitz *et al.*, 1995b). In neutrophils, protein kinase activities are also detected in association with CGM6 and NCA (Skubitz *et al.*, 1995b). Also, the binding of *N. gonorrhoeae* to BGP, CGM6 and NCA on human phagocytes activates signalling cascades via Rac1 and Src-like tyrosine kinases (Hauck *et al.*, 1998).

CGM6 and NCA, like CEA, are GPI-anchored proteins (Figure 4.2), and although the phenomenon of cell-signalling by GPI-anchored proteins has been recognised for some time (Stefanova *et al.*, 1991), the precise nature of signalling by CEA and its related proteins is unknown. Despite research into the cell-signalling properties of CEA family proteins being relatively recent, the findings so far indicate that, whatever the precise biological functions of these molecules, they are intimately coordinated to the other functions of the cell.

4.1.4. CEA-based strategies for the treatment of cancer

In the effort to combat colon cancer and others, the up-regulation of CEA expression by tumour cells has resulted in its use as a target for several detection, monitoring and treatment strategies. The specific binding of antibodies raised against CEA is often exploited to detect tumours in patients. The *in vivo* localization of an anti-CEA antibody labelled with a radioactive isotope can be determined using a γ-emission camera. The detection of CEA localized to a tumour can then be used to guide decisions on surgical procedures. More radically, an anti-CEA antibody can be coupled to a cytotoxic agent, such as to a powerful radioisotope or to an enzyme that will activate a cytotoxic agent, and this will enable cells at the site of the tumour to be specifically killed. Antibody-based strategies are currently at the forefront of anti-cancer research and an overview of the design of antibody fragments for such targeting strategies is given in Chapter 6.

CEA has also been used in strategies which attempt to induce a human immune response to tumour cells (Zbar *et al.*, 1998). For example, a mouse anti-idiotypic antibody that mimics a specific CEA epitope and a recombinant vaccinia virus that expresses CEA have both been developed for vaccinations, with the aim of stimulating an immune response against CEA-expressing cancer cells (Pervin *et al.*, 1997; McAneny *et al.*, 1996). Alternatively, a bi-specific antibody that recognises CEA and the T-cell marker CD3 has been developed for directly targeting activated T-cells to CEA-expressing tumour cells (Kuwahara *et al.*, 1996). In another approach, an *in vitro* cytotoxic T-cell response against CEA-expressing cells is generated by exposing them to either dendritic cells that have been pulsed with a CEA peptide (Alters *et al.*, 1998),

or dendritic cells transfected with CEA mRNA (Nair *et al.*, 1998).

A different approach to utilizing the up-regulation of CEA in tumours focuses on the expression of CEA at the DNA level. It is proposed that the CEA gene promoter could be used in gene therapy: a gene under the control of the CEA promoter will be specifically activated upon exposure to the transcription factors produced by CEA-expressing cells. After identification of the *cis*-acting DNA sequences that are important for transcription of the CEA gene, Richards *et al.* (1995) linked this promoter to the cytosine deaminase gene. Upon introducing this construct into CEA-expressing cells, they demonstrated the subsequent expression of cytosine deaminase, the enzymatic activity of which produces the cytotoxic drug 5-fluorouracil from inactive 5-fluorocytosine. Adenovirus-based vectors have been developed to carry chimeric gene constructs under the control of the CEA promoter (Lan *et al.*, 1996; Tanaka *et al.*, 1996). A degree of success has been achieved by infecting cells with these vectors, whereby CEA-expressing cells specifically promote expression of the chimeric gene product (Lan *et al.*, 1997; Tanaka *et al.*, 1997).

4.1.5. The objective of CEA X-ray and neutron solution scattering studies

Chapter 4 here details the determination of a first model for the glycosylated, seven-domain structure of CEA by X-ray and neutron solution scattering (Boehm *et al.*, 1996). Targeted cancer therapy depends on efficient ligand binding to a target on tumour cells. Both rapid on-rate and slow off-rate are required as components of such ligands. Antibodies through their great potential for diversity provide ligands for targeting to most common types of cancer. Manipulation of antibody structures to provide slower off-rates should facilitate tumour retention of antibody, and will offer a powerful means of cancer therapy. In order to improve the reliability of CEA as a tumour marker, the different epitopes in CEA require identification, together with those in the other CEA family proteins, in order to suggest improvements to existing anti-CEA antibodies by the use of molecular biology methods.

No atomic structure is known for CEA at present. For reason of its glycosylation, it is most unlikely that the extracellular domains could be crystallised

intact, and alternative strategies are required for a structure determination. Electron microscopy has suggested that CEA is rod-shaped (Slayter & Coligan, 1975), although this may be limited by the non-physiological conditions of measurement. An extended CEA model was predicted on the basis of homology with crystal structures for the Bence-Jones dimer REI, the cell surface proteins CD4 and CD2 and the linker between the V_L and C_L domains of the human New IgG1 Fab fragment (Bates *et al.*, 1992). Such a model however requires experimental validation. X-ray and neutron scattering will provide structural information to a resolution of approximately 3 nm, and is well-suited for such a study of the CEA domains. No special preparation of CEA is required, apart from its solubilisation from membranes, and data are obtained in near-physiological conditions. The joint use of X-ray scattering in H_2O buffers and neutron scattering in ^2H_2O buffers enables the contributions of protein and carbohydrate to the scattering curve to be properly analysed. In X-ray scattering, the CEA carbohydrate makes a greater contribution to the intensity of the observed curve than the CEA protein, and in neutron scattering it contributes less. The scattering data will show whether CEA contains an extended or compact arrangement of domains, and whether it is monomeric or dimeric.

Scattering analyses will be enhanced if a new method of constrained curve modelling is applied which permits a systematic automated evaluation of possible models for CEA, and places limits on the precision of such models (Perkins *et al.*, 1991; Mayans *et al.*, 1995; Beavil *et al.*, 1995). It had been shown that scattering curves are fully calculable from atomic coordinates (Smith *et al.*, 1990; Perkins *et al.*, 1993). The CEA domains can be represented using known atomic structures from homologous proteins from the Ig superfamily. At the time this modelling study was performed, crystal structures of related cell-surface proteins were only known for CD2 and CD4 (Wang *et al.*, 1990; Ryu *et al.*, 1990; Jones *et al.*, 1992; Brady *et al.*, 1993; Bodian *et al.*, 1994). This meant that both the A- and B-type domains from CEA could be modelled assuming a C2-set Ig-fold structure (Williams & Barclay, 1988; Figure 4.3). Starting from these crystal structures, it is possible to analyse all arrangements of CEA domains to determine which are the most compatible with the experimental

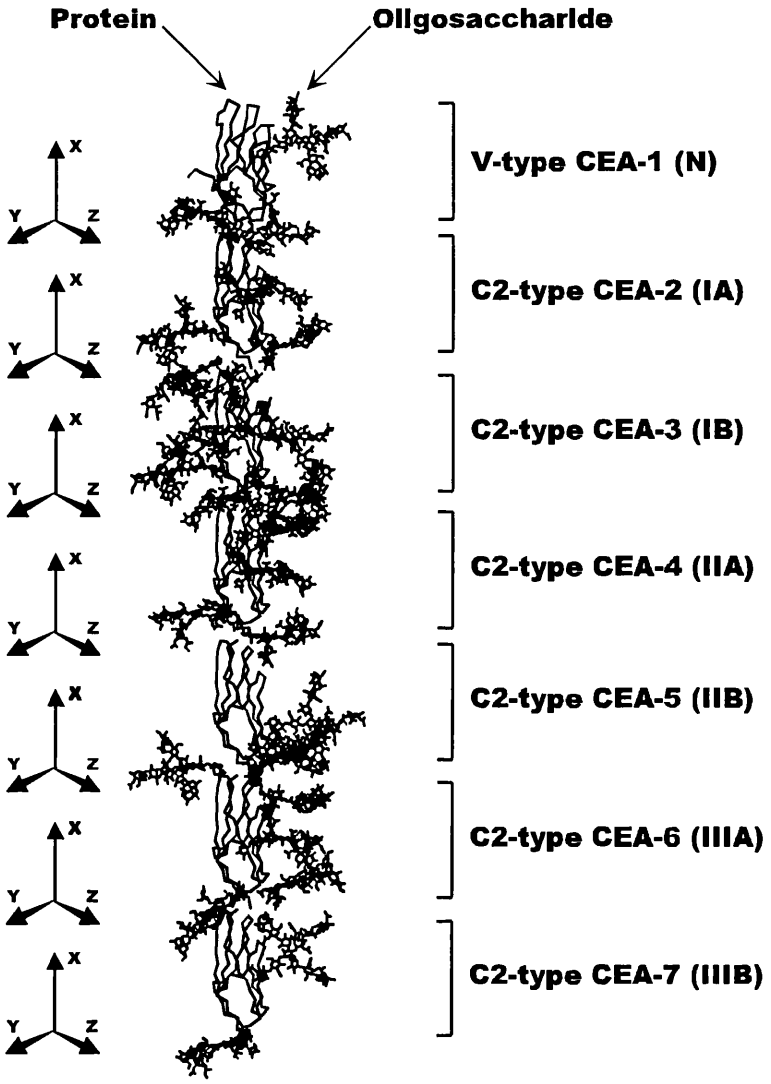


Figure 4.3. A linear model of CEA derived from the crystal structure of human CD2. The CEA domains are numbered as 1 to 7, and as N to IIIB to follow Oikawa *et al.* (1987). Seven separate CD2 C2-set domains (shown as α-carbon traces) were orientated such that the N-terminus and C-terminus of each C2-set domain (α-carbons of Glu104 and Pro180 in 1hnf) were aligned on the same X-axis, with a 0.416 nm spacing between adjacent C2-set domains. A variable (V) -set CD2 domain was superimposed on the N-terminal C2-set domain. X-, Y- and Z-axes were assigned to each domain at the α-carbon of Pro180 as origin so that each domain could be rotated independently of each other. Twenty-eight oligosaccharide chains were added in full at the putative N-linked glycosylation sites (see text).

scattering curves. The outcome of this constrained conformational search is used to propose a low resolution model for the complete glycoprotein structure of CEA solubilised from membranes, as well as the limits of error on such a structure. The resulting model permits assessment of the steric accessibility of the Ig fold domains in CEA in terms of its physiological role in cellular adhesion events (Benchimol *et al.*, 1989; Zhou *et al.*, 1993), and also in terms of binding to antibodies and scFv antibody fragments (Hammarström *et al.*, 1989; Schwarz *et al.*, 1988; Ikeda *et al.*, 1992; Nap *et al.*, 1992; Murakami *et al.*, 1995).

4.2. Materials and methods

4.2.1. Preparation of CEA for solution scattering

CEA was prepared in two batches by Dr P. A. Keep and Mr J. D. Thornton, one from a pool of five liver metastases of colon cancer and one from a single liver metastasis. In each case, the extraction procedure was the same (Keep *et al.*, 1978). Slices of frozen liver tissue were homogenised with 0.01 M phosphate-buffered saline at 4°C to give a thick slurry, which was stored at -20°C until required. To portions of this was added an equal volume of cold 2M perchloric acid with stirring for 5 min. The resulting homogenate was centrifuged at 4°C and the supernatant dialysed against 4 changes of distilled water for 20 h at 4°C. The extract was concentrated by Amicon PM10 ultrafiltration for the gel filtration step. For the second batch, the extraction buffer also contained the protease inhibitors pepstatin A, chymostatin, leupeptin and antipain, each at 10 mg/l. Sodium azide (0.02%) was also added.

The crude extract from the first batch was purified by gel filtration on Sepharose 6B and Sephacryl S-300 in 0.05 M sodium phosphate buffer, pH 7.5 with 0.02% sodium azide. The CEA-containing fractions were identified by radioimmunoassay and by double diffusion in agar against rabbit anti-CEA. The peak corresponding to an M_r of 180,000 was taken in each case. 4 mg of CEA as determined by radioimmunoassay (goat polyclonal anti-CEA) was purified further by immunoadsorption on a column of A5B7 monoclonal anti-CEA (43 mg) coupled to Sepharose 4B. Non-bound material was recovered by elution with phosphate-buffered saline (12 mM Na, K phosphate; 140 mM NaCl; pH 7.4) and the bound material by elution with 3M ammonium thiocyanate

(40 ml). The bound fraction was concentrated by ultrafiltration at 4°C in an Amicon stirred cell with a PM10 membrane to 3.7 ml, then dialysed overnight against phosphate-buffered saline and filtered using a 0.2 μm filter. Aggregates were removed by fast protein liquid chromatography gel filtration on a Superose-12 (Pharmacia) column with two runs, each with 200 μl of CEA applied to the column and eluted with phosphate-buffered saline. Fractions (20 drops) were collected and tested against PK1G goat anti-CEA antiserum by double diffusion in agar. Fractions 9 and 10 from each run were pooled, filtered (0.2 μm filter) and concentrated by Centricon-30 (Amicon) centrifugation to 380 μl. A 30 μl aliquot was removed for characterisation and the remainder (0.90 mg CEA/ml by Hybritech CEA assay) was used for scattering or sedimentation analyses. The second batch of perchloric acid-extracted CEA was processed as for the first batch. Each batch gave a single diffuse band on non-reduced SDS-PAGE at a molecular weight of 200,000. Whilst perchloric acid cleavage reduces CEA yields (Kimball & Brattain, 1978), antibodies raised to such CEA preparations have been successfully used to locate colonic tumours in patients (Lane *et al.*, 1994).

Preparations were cleared of aggregates shortly before scattering data collection by gel filtration through Superose-12 using a fast protein liquid chromatography system (Pharmacia). Proteins were dialysed at 6°C with 4 changes of buffer over at least 36 hours into phosphate buffer saline in H_2O (X-rays) or 99.8% ^2H_2O (neutrons). The amino acid composition of CEA was taken from Oikawa *et al.* (1987) after exclusion of the signal peptide and membrane-spanning regions. Yamashita *et al.* (1987, 1989) showed that 92% of CEA oligosaccharides are the complex type with a standard $Man_3GlcNAc_2$ core. In this core structure, 40% contain an additional bisecting GlcNAc residue, and 86% contain a further fucose on the proximal GlcNAc residue. On average, there are 2.7 outer chain branches attached to the core, which are mainly Galß1→4GlcNAc (43.8%) or Galß1→4(Fucα1→3)GlcNAc (35.1%) repeats. The full analysis indicated that an average oligosaccharide chain contained 3.1 Man, 5.5 GlcNAc, 3.1 Gal and 2.7 Fuc residues. The sialic acid content per oligosaccharide chain was estimated to be 0.6 (Chandrasekaran *et al.*, 1982; Slayter & Coligan, 1975; Hammarström *et al.*, 1975; Kessler *et al.*, 1978; Pavlenko *et al.*, 1990). The CEA carbohydrate was therefore represented by 28 triantennary complex-type oligosaccharide

chains of composition $Man_3GlcNAc_6Gal_3Fuc_3NeuNAc_1$ which were located at putative NXT or NXS sites in the CEA sequence (except when X = Pro). This composition is 53% by weight of CEA and is within error of the carbohydrate analyses of Kessler *et al.* (1978) and Chandrasekaran *et al.* (1983). The total M_r of CEA is calculated as 152,500, which is 15% less than that of the commonly reported value of 180,000 for CEA from SDS-PAGE (Slayter & Coligan, 1975). This difference is often observed for heavily glycosylated glycoproteins (Gordon, 1975), e.g. a 32% reduction was seen for $C\bar{1}$ inhibitor of complement (Perkins *et al.*, 1990). Protein concentrations for the M_r calculations from $I(0)$ Guinier data (see below) were calculated from optical density measurements at 280 nm using an absorption coefficient (1%, 1 cm) of 6.41 calculated from the CEA composition by the corrected Wetlaufer procedure (Perkins, 1986).

4.2.2. Synchrotron X-ray data collection at Station 8.2 at SRS

X-ray scattering data were obtained in six independent sessions using the low angle solution scattering camera at Station 8.2 (Towns-Andrews *et al.*, 1989) at the SRS Daresbury, Warrington, UK. Experiments were performed with beam currents in a range of 120-230 mA and a ring energy of 2.00 GeV. Samples were measured for 10 minutes in 10 time frames for protein concentrations that ranged between 1.2 to 7.0 mg/ml. The use of a 500-channel quadrant detector (Worgan *et al.*, 1990) with sample-detector distances of 3.26 m to 3.54 m resulted in a usable Q range between 0.07-2.2 nm^{-1}. The Q range was calibrated using fresh, wet, slightly stretched rat tail collagen, based on a diffraction spacing of 67.0 nm. Samples were held in Perspex cells of sample volume 20 µl, contained within mica windows of thickness between 10 to 15 µm, and cooled at 15°C. Buffers and samples were measured in alternation for equal times to minimise background subtraction errors. Data were only accepted if the subsequent Guinier plots were linear and reproducible in repeated measurements. Each of the 10 time frames was individually checked using Guinier analysis to check for the absence of time-dependent radiation damage effects. Data reduction was performed using the standard Daresbury software package OTOKO (P. Bendall, J. Bordas, M. H. C. Koch & G. R. Mant, EMBL Hamburg and Daresbury Laboratory, unpublished software). Curves were normalised using an ion chamber monitor positioned after the sample for individual runs, and a detector response measured for at least 8 hours using a uniform

^{55}Fe radioactive source. Reduced curves were calculated by subtraction of the buffer runs from those of the samples.

4.2.3. Pulsed neutron data collection at Instrument LOQ at ISIS

Neutron scattering data were obtained in four different beam sessions on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, U.K. (Heenan & King, 1993). The pulsed neutron beam was derived from proton beam currents of 160-190 μA. Monochromatisation was achieved using time-of-flight techniques. A ^{3}He ORDELA wire detector was employed to record intensities at a fixed sample-to-detector distance of 4.3 m. The samples and $^{2}H_2O$ buffers were measured in 2 mm-thick rectangular Hellma cells positioned in a thermostatted rack at 15°C. Data acquisitions were for fixed totals of 4.0×10^6 monitor counts in runs lasting 50-60 min each for protein concentrations between 3.6-7.3 mg/ml. Spectral intensities were normalised relative to the scattering from a partially deuterated polystyrene standard (Wignall & Bates, 1987). Transmissions were measured for all samples and backgrounds. Reduction of the raw data collected in 100 time frames of 64 × 64 cells utilised the standard ISIS software package COLETTE (Heenan *et al.*, 1989). Scattered intensities were binned into individual diffraction patterns based on the wavelength range from 0.22 nm to 1.00 nm, and were corrected for a linear wavelength-dependence of the transmission measurements. The patterns were merged to give the full curve $I(Q)$ in a maximal Q range of 0.05 - 2.2 nm^{-1} ($Q = 4\pi \sin\theta / \lambda$; 2θ = scattering angle; λ = wavelength). The Q range was based on 0.04% logarithmic increments, which was optimal both for Guinier analyses at low Q, and for better signal-noise ratios at large Q.

4.2.4. Analysis of reduced X-ray and neutron data

In a given solute-solvent contrast, the radius of gyration R_G is a measure of structural elongation if the internal inhomogeneity of scattering densities has no effect. Guinier analyses at low Q give the R_G and the forward scattering at zero angle $I(0)$ (Glatter & Kratky, 1982):

$$\ln I(Q) = \ln I(0) - R_G^2 \, Q^2/3.$$

This expression is valid in a $Q.R_G$ range up to 0.7 for extended rod-like particles, and is approximate in a $Q.R_G$ up to 1.5 in which it underestimates the true R_G. The relative $I(0)/c$ values (c = sample concentration) for samples measured in the same buffer during a data session gives the relative molecular weights M_r of the proteins when referenced against a suitable standard (Kratky, 1963; Wignall & Bates, 1987). If the structure is elongated, the mean radius of gyration of the cross-sectional structure R_{XS} and the mean cross-sectional intensity at zero angle $[I(Q).Q]_{Q\to 0}$ (Hjelm, 1985) can be obtained from:

$$\ln [I(Q).Q] = \ln [I(Q).Q]_{Q\to 0} - R_{XS}^2\, Q^2/2.$$

The R_G and R_{XS} analyses lead to the triaxial dimensions of the macromolecule. If the structure can be represented by an elongated elliptical cylinder, $L = [12(R_G^2 - R_{XS}^2)]^{1/2}$, where L is its length (Glatter & Kratky, 1982). Alternatively, L is given by $\pi I(0)/[I(Q).Q]_{Q\to 0}$ (Perkins *et al.*, 1986). The two semi-axes, A and B, of the elliptical cylinder are calculated by combining the dry or hydrated volume V ($V = \pi ABL$) with the R_{XS} value ($R_{XS}^2 = (A^2 + B^2)/4$). The hydrated volume is obtained on the basis of a hydration of 0.3 g of water/g of glycoprotein and 0.0245 nm^3 per water molecule (Perkins, 1986). Data analyses employed an interactive graphics program SCTPL4 (A.S. Nealis & S.J. Perkins, unpublished software) on a Silicon Graphics 4D35S Workstation.

Indirect transformation of the scattering data $I(Q)$ in reciprocal space into real space to give $P(r)$ was carried out using the ITP-91 program of Glatter (Glatter & Kratky, 1982).

$$P(r) = \frac{1}{2\pi^2} \int_0^\infty I(Q)\, Qr\, \sin(Qr)\, dQ$$

$P(r)$ corresponds to the distribution of distances r between any two volume elements within one particle weighted by the product of their respective electron or nuclear densities relative to the solvent density. This offers an alternative calculation of the R_G and $I(0)$ that is based on the full scattering curve, and gives the maximum dimension of

the macromolecule L. For CEA, the X-ray $I(Q)$ contained 247 data points extending out to 2.04 nm^{-1} and was fitted with 10 splines with D_{max} set as 35 nm. The neutron $I(Q)$ contained 78 data points extending out to 2.01 nm^{-1} and was fitted with 6 splines with D_{max} set as 35 nm. $P(r)$ was defined by 101 points. Criteria for the correct choice of parameters for $P(r)$ were: (i) $P(r)$ should exhibit positive values; (ii) the R_G from ITP and Guinier analyses should agree; (iii) $P(r)$ should be zero when r is zero; (iv) $P(r)$ should be stable and reproducible for different experimental $I(Q)$ curves when the number of splines and D_{max} is varied over a reasonable range. L was determined from $P(r)$ when this became zero at large r; however errors in L can be significant as a result of the low intensity of $P(r)$ in this region.

4.2.5. Automated Debye scattering curve modelling of CEA

INSIGHT II V2.3.0 molecular graphics software (Biosym Technologies Inc.) on Silicon Graphics Indigo and Indy Workstations (R3000 and R4000 series with 48-64 Mb memory) were utilised for all manipulations. Atomic coordinates for the CEA-related cell surface proteins CD2 and CD4 corresponded to pairs of V-set and C2-set Ig folds (Brookhaven database codes: human CD4 domains 1 and 2, 1cdh, 1cdi, 3cd4, 1cd4; rat CD4 domains 3 and 4, 1cid; human and rat CD2 domains 1 and 2, 1cdb, 1hnf, 1hng), except for 1cdb which is a single V-set domain and corresponds to an NMR solution structure. The CEA oligosaccharide structure was generated from the 9-residue carbohydrate structure in the Fc fragment of human IgG1 KOL (Brookhaven code 1fc1), to which seven additional carbohydrate residues were attached singly or in pairs at appropriate positions to generate the full structure with composition as above. For the AUTOSCT automated curve fitting procedure (Beavil *et al.*, 1995), a Biosym Command Language macro was written to rotate each domain relative to one another to generate CEA conformations. The rotational centre of each domain was defined as the α-carbon atom of the C-terminal residue. For each C2-set domain, the X-axis was defined by the line joining the N-terminal and C-terminal α-carbon atoms, and the plane of its Y-axis was defined by an α-carbon atom. The corresponding X-axis and Y-axis for the V-set domain were defined by superimposition of the C2-set domain onto the V-set domain (see Results).

Each CEA coordinate model was converted into a scattering curve fit. A CEA sphere model was created by placing the full atomic coordinate set within a three dimensional array of cubes, each of side length 0.572 nm. This length is much less than the nominal resolution of $2\pi / Q_{max}$ of the scattering curves (3.1 nm for $Q_{max} = 2.0$ nm^{-1} in X-ray experiments; 3.7 nm for $Q_{max} = 1.7$ nm^{-1} in neutron experiments). If the number of atoms within a cube exceeded a user-defined cutoff, a sphere of the same volume as the cube (sphere diameter 0.710 nm) was placed at the centre of the cube. This cutoff was determined by the requirement that the total volume of spheres was within 1% of the dry volume of 178.9 nm^3 for CEA calculated from its composition (Chothia, 1975; Perkins, 1986). This cutoff was set as a minimum of three and four atoms to define each protein and carbohydrate sphere respectively. The total of 485 protein and 474 carbohydrate spheres corresponded respectively to volumes of 90.6 nm^3 and 88.3 nm^3 calculated from the CEA composition. X-ray scattering experiments are influenced by a hydration shell surrounding the glycoprotein. A hydration of 0.3 g H_2O / g of glycoprotein and an electrostricted volume 0.0245 nm^3 per bound water molecule (Perkins, 1986) was used to calculate a hydrated CEA volume of 241.1 nm^3 for X-ray curve fits based on sphere diameters of 0.783 nm. In neutron experiments, the hydration shell was not detectable and the dry sphere models were used for curve fits.

The scattering curve $I(Q)$ was calculated using Debye's Law adapted to spheres, essentially by computing all the distances r from each sphere to the remaining spheres and summing the results. The different scattering densities of protein and carbohydrate were incorporated in two-density scattering curves calculated from (Glatter & Kratky, 1982):

$$[I(Q)/I(0)] = g(Q)\,[n_1\rho_1^2 + n_2\rho_2^2 + 2\rho_1^2 \sum A_j^{11} (\sin Qr_j / Qr_j)$$
$$+ 2\rho_2^2 \sum A_j^{22} (\sin Qr_j / Qr_j)$$
$$+ 2\rho_1\rho_2 \sum A_j^{12} (\sin Qr_j / Qr_j)]$$
$$\times (n_1\rho_1 + n_2\rho_2)^{-2}$$

The CEA model is constructed from n_1 and n_2 spheres of different densities ρ_1 and ρ_2;

$g(Q) = 3(\sin QR - QR \cos QR)^2/Q^6R^6$ (the squared form factor of the spheres of radius R); A_j^{11}, A_j^{22}, and A_j^{12} are the number of distances r_j for that increment of j between the spheres 1 and 1, 2 and 2, and 1 and 2 in that order; the summations Σ are performed for $j = 1$ to m, where m is the number of different distances r_j. For the X-ray data, no corrections were applied for wavelength spread or beam divergence as these are thought to be negligible. For the neutron data, a 16% wavelength spread for a nominal λ of 1.0 nm and a beam divergence of 0.016 radians were used as an approximation to correct the calculated neutron scattering curve for the reasons discussed in Mayans *et al.* (1995). The quality of the curve fits was assessed by calculations of the R_G and R_{XS} values of the model from the scattering curve in the same Q ranges used for Guinier fits, and the crystallographic R-factor in the Q range extending to 1.6 nm^{-1} (denoted as $R_{1.6}$) or to 2.0 nm^{-1} ($R_{2.0}$) (Smith *et al.*, 1990; Beavil *et al.*, 1995). Curve fits were assessed using spreadsheets in which the absence of steric overlap was verified from the volume of the model, which should be close to 959 spheres. The models were filtered and sorted on the basis of comparisons between the observed and calculated R_G and R_{XS} values and the R-factor values.

4.2.6. Hydrodynamic analyses and modelling of CEA

Sedimentation coefficients $s^o_{20,w}$ for CEA were measured at 0.42, 0.59 and 0.83 mg/ml at 20°C on a Beckmann XL-A analytical ultracentrifuge operated at 30,000 rpm and equipped with scanning absorption optics by Dr O. Byron at the National Centre for Macromolecular Hydrodynamics, Leicester. Traces were measured at 280 nm and analysed using a digitising pad interfaced with an Apple II computer to yield experimental $s^o_{20,w}$ values after correction for the density and viscosity of the buffer. Frictional coefficients f were calculated from the $s^o_{20,w}$ values for comparison with f values calculated directly from the hydrated models with approximately 959 spheres used for fits of the X-ray scattering curves (but now using non-overlapping spheres to satisfy the algorithm in use). These f values were derived using the modified Oseen tensor procedure in the program GENDIA (Garcia de la Torre & Bloomfield, 1977a, 1977b; Perkins *et al.*, 1993), and were imported into Microsoft Excel 5.0a spreadsheets for joint analyses with the scattering fits. A more recent hydrodynamic modelling program HYDRO permitted the use of overlapping spheres (Garcia de la Torre, 1989).

As this gave results that were very similar to those from GENDIA at the cost of much increased CPU time, HYDRO was only used as a control of the outcome of the GENDIA simulations.

4.3. Results and discussion

4.3.1. Synchrotron X-ray scattering measurements on CEA

CEA was solubilised to cleave it from its membrane anchor by treatment with perchloric acid (Materials and Methods), and pretreated by gel filtration to remove non-specific aggregates. Using synchrotron X-ray scattering data, linear Guinier plots could be obtained for ten different CEA samples in the concentration range between 1.6 to 7.0 mg/ml (Figure 4.4a). Occasionally steeply curved Guinier regions at low Q were observed in place of the linear plot seen in Figure 4.4a, and these data were rejected as these correspond to CEA aggregates. Time-frame analysis of the Guinier region showed that CEA was resistent to radiation damage effects that often occur with synchrotron X-ray beams. Guinier analyses gave a consistent R_G value of 8.0 ± 0.6 nm in a $Q.R_G$ range of 0.8 to 1.5. Since CEA is expected to be rod-like in structure, the R_G values will underestimate the true value for CEA as the available Q range for measurement is not sufficiently low. All R_G values cited below are the apparent values from Guinier fits except when specified otherwise. Calculation of the elongation or anisotropy ratio of CEA from the ratio R_G/R_O (where R_O is the R_G of the sphere with the same hydrated volume of 241.1 nm^3 as CEA) gave a minimum value of 2.7 ± 0.2. Since R_G/R_O values for typical globular proteins are close to 1.28 (Perkins, 1988), it is concluded that CEA possesses a highly elongated structure in solution.

CEA gave satisfactory linear cross-sectional R_{XS} analyses (Figure 4.4b) in an acceptable $Q.R_{XS}$ range of 0.5 to 0.9 and beyond to higher Q. The ten CEA curves gave a consistent R_{XS} value of 2.1 ± 0.2 nm. The combination of the R_G and R_{XS} analyses (Materials and Methods) gave a length L of 27 ± 2 nm for CEA. The combination of the $I(0)$ and $[I(Q).Q]_{Q\to 0}$ analyses gave a similar length L of 31 ± 4 nm. If the hydrated volume of CEA is 241.1 nm^3, combination of the L and R_{XS} values showed that to a first approximation CEA can be represented by a cylinder of dimensions $L \times 2A \times 2B$ of 29 nm × 8 nm × 1 nm. This showed that CEA has an elongated cross-section. While the

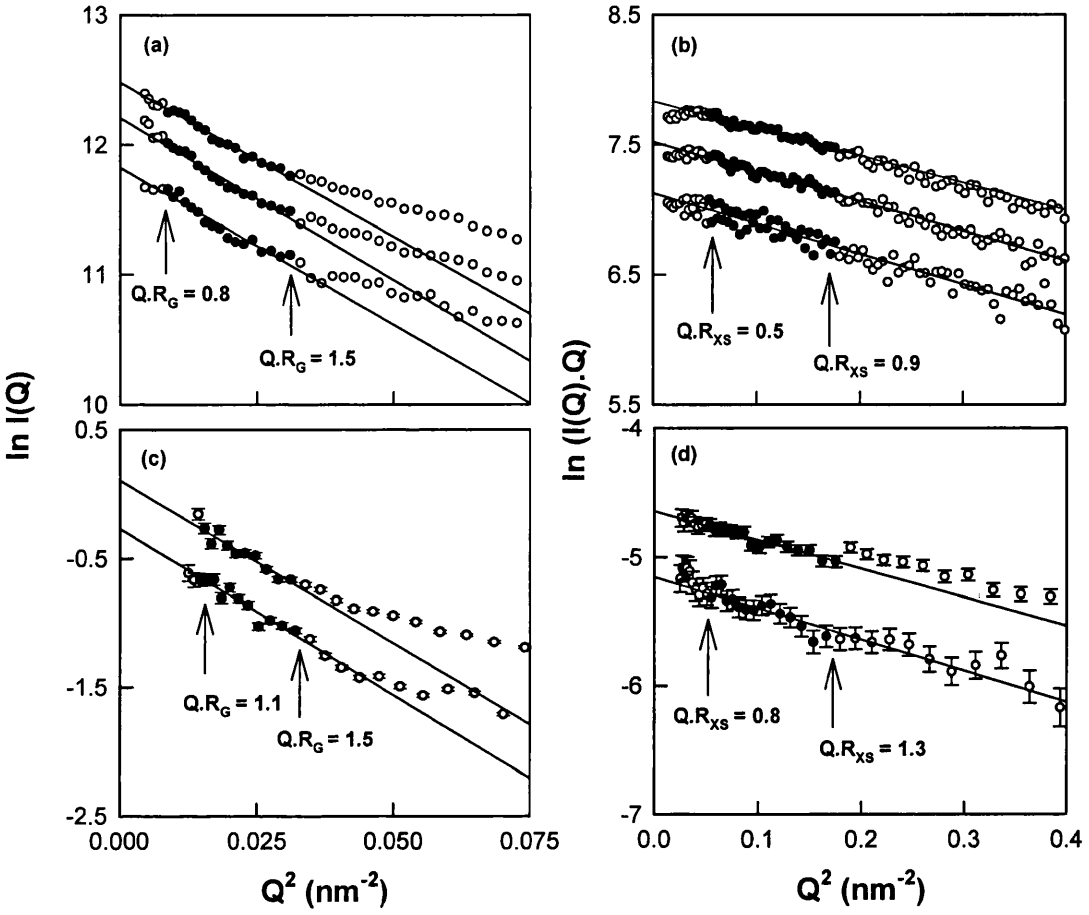


Figure 4.4. X-ray and neutron Guinier R_G and R_{XS} plots for CEA. (*a*) and (*b*) depict the X-ray plots for CEA concentrations of 3.3, 2.5 and 1.6 mg/ml. The filled circles between the $Q.R_G$ and $Q.R_{XS}$ ranges as arrowed show the data points used to determine R_G and R_{XS} values. The R_G values were extracted using a Q range of 0.09-0.18 nm^{-1}, and the R_{XS} values likewise from a Q range of 0.23-0.42 nm^{-1}. Data outside these ranges are denoted by open circles. (*c*) and (*d*) show the neutron plots for CEA concentrations of 7.3 and 3.6 mg/ml, with other details as for (*a*) and (*b*) except that the Q range used for the R_G values was 0.12-0.18 nm^{-1}. Error bars correspond to the statistical errors of LOQ neutron data collection.

cross-sectional dimensions of 8 nm × 1 nm are not physically realistic, they can be attributed to the consequence of highly extended carbohydrate chains on the protein surface.

The $I(Q)$ curves in reciprocal space for CEA were transformed into distance distribution functions $P(r)$ in real space (Figure 4.5a). The R_G and $I(0)$ values calculated from $P(r)$ analysis were within 6% and 2% respectively of the Guinier values (Figure 4.4a), which indicates that these analyses are self-consistent. The point at which the $P(r)$ curve intersects the zero axis at large inter-vector distances R gave the maximum dimension L of CEA as 33 nm. This is comparable to the values calculated from the two Guinier analyses above, although the precision of this determination is not high for reason of the low intensity of $P(r)$ at large r. The maximum M in $P(r)$ gives the most frequently occurring distance within CEA and this was determined as $r = 4.9$ nm.

4.3.2. Pulsed neutron scattering measurements on CEA

Two CEA samples were also studied by neutron scattering as a control for X-ray-induced radiation damage and contrast-dependent properties. In common with other glycoproteins, CEA was prone to aggregation for reason of the weaker hydrogen bonding properties of the ^2H_2O solvent used for neutron work. All samples yielding curved Guinier plots were not considered further. The neutron Guinier R_G analysis of Figure 4.4c showed that CEA in ^2H_2O buffers resulted in an R_G value of 8.8 ± 0.5 nm. This is within error of the X-ray data, although the precision of measurement on LOQ is less for reason of a reduced $Q.R_G$ range of fit. The neutron Guinier $I(0)/c$ values are standardised relative to a standard deuterated polymer, and the M_r of CEA can be determined by comparison with other proteins in ^2H_2O buffers measured on LOQ. The mean $I(0)/c$ value for CEA was determined to be 0.183 ± 0.043. Simulations show that the systematic error of this determination is maximally 5%. The $I(0)/c$ value for bovine IgG1 and IgG2 were determined to be 0.180 and 0.182 (± 0.006) (Mayans *et al.*, 1995). As IgG1 and IgG2 have M_r values of 144,000, the M_r for CEA was determined as 150,000 ± 35,000. This agrees well with the M_r of 152,500 calculated from the CEA composition (Materials and Methods). The M_r calculation validated the CEA scattering data and showed that CEA as prepared is monomeric, thus ruling out dimer models for

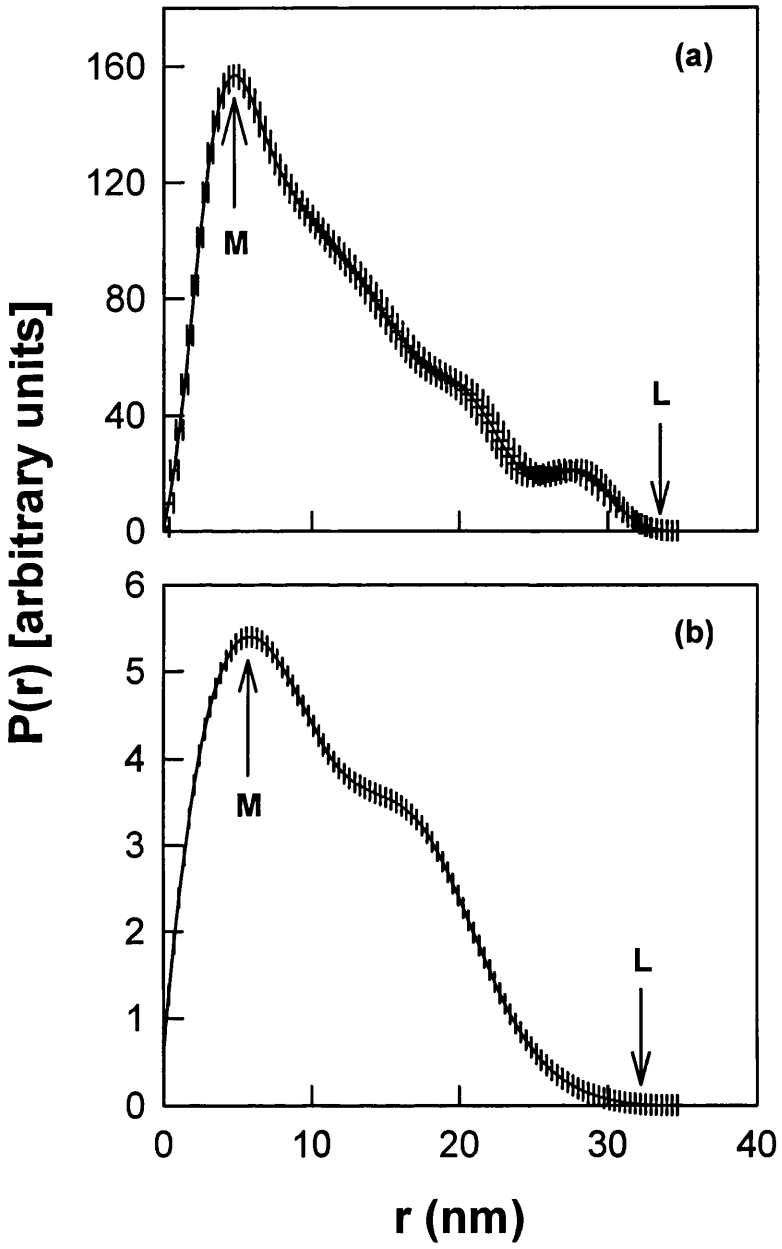


Figure 4.5. X-ray and neutron distance distribution functions $P(r)$ for CEA. (*a*) shows the X-ray $P(r)$ curve at a CEA concentration of 3.3 mg/ml as a continuous line with error bars as calculated using ITP-91, and a maximum M at 4.9 nm. (*b*) shows the neutron $P(r)$ curve at a CEA concentration of 3.6 mg/ml, with M at 6.0 nm. The maximum dimension of CEA was determined to be 28-32 nm from (*a*) and (*b*).

CEA.

While less precise than the X-ray data, the neutron analyses validated the results from X-ray scattering. The neutron Guinier cross-sectional R_{XS} analysis of Figure 4.4d resulted in an R_{XS} value of 2.3 ± 0.3 nm for CEA in ^2H_2O buffers. The neutron R_G and R_{XS} analyses gave a length L of 37 ± 2 nm for CEA, while the $I(0)$ and $[I(Q).Q]_{Q\to 0}$ analyses gave L as 29 ± 1 nm. The dimensions $L \times 2A \times 2B$ of CEA were determined to be 33 nm × 9 nm × 1 nm from a dry volume of 178.9 nm^3. Interestingly the R_{XS} values did not exhibit a contrast dependence as reported elsewhere for glycoproteins (Perkins *et al.*, 1990). Such a dependence was in fact found at larger Q beyond the R_{XS} region in the X-ray and neutron scattering curves of CEA (below). The neutron distance distribution function $P(r)$ for CEA in Figure 4.5b gave R_G and $I(0)$ values within 5% and 1% respectively of those found in the Guinier analyses, and showed that the neutron $P(r)$ and Guinier analyses were self-consistent. The maximum M in $P(r)$ was 6.0 nm, and the maximum dimension L was 29 nm. All these analyses gave values similar to the X-ray values.

4.3.3. Initial molecular graphics model for CEA

As the CEA domains belong to the Ig superfamily (Williams & Barclay, 1987; Bates *et al.*, 1992), the experimental X-ray and neutron curves of CEA were modelled using atomic coordinates for one V-set and six C2-set Ig fold domains. The most relevant crystal structures were those for the related cell-surface proteins CD2 and CD4. This was examined further by means of manual and MULTAL automated sequence alignments (Taylor, 1988). Of the 108 residues in the CEA V-set sequence, 42% showed >60% residue type conservation with the CD2 and CD4 V-set sequences. If the C2-set sequences of CEA were compared one-by-one with the CD2 and CD4 sequences, 31-34% of residues showed >60% conservation. Higher similarities were found with CD2 than with CD4, and this analysis confirms and extends the earlier observation that CEA belongs to a subset of the Ig superfamily (Killeen *et al.*, 1988). The DSSP program (Kabsch & Sander, 1983) was used to locate the ß-strands in the crystal structures of CD2 and CD4 in order to position these in the CEA domains. These were consistently located in all 8-9 structures (Figure 4.6), and their positions could be

Figure 4.6. (Overleaf) Sequence alignment of CEA with CD2 and CD4 and their known structures. (*a*) The V-set domain of CEA (CEA-1) is compared with V-set domains in CD2 and CD4. (*b*) The C2-set domains of CEA (CEA-2 to CEA-7) are compared with C2-set domains in CD2 and CD4.

In (*a*) and (*b*), residues are asterisked if residue conservation is at least 60%, with conserved groups of residues as defined in brackets: (G, A, S), (I, L, M, V), (F, H, W, Y), (D, E), (H, K, R), (S, T), (N, Q). Putative N-linked glycosylation sites are bolded and underlined. Sequences are labelled with the appropriate Brookhaven Protein Database code. The secondary structure elements identified by the DSSP program are labelled as follows: E, ß-strand (bolded); B, single residue ß-ladders; T, turns; S, bends; G, 3_{10}-helix. In (*b*), the inter-domain link residues in human and rat CD2 are ERVS and EMVS respectively. At the bottom of each of (*a*) and (*b*), the location of at least one putative Asn glycosylation site in sequence alignments of subgroups of the CEA gene family is denoted by N.

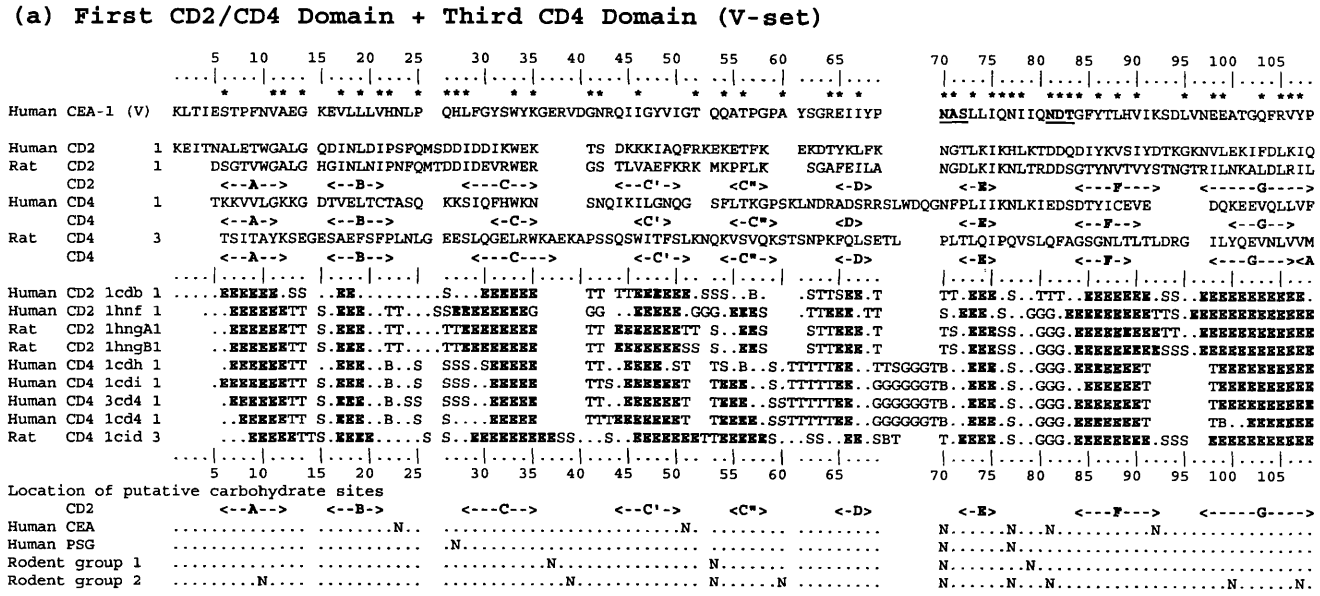


Figure 4.6. (*a*) Sequence alignment of CEA with CD2 and CD4 and their known structures (legend on page 154).

(b) Second CD2/CD4 Domain + Fourth CD4 Domain (C2-set)

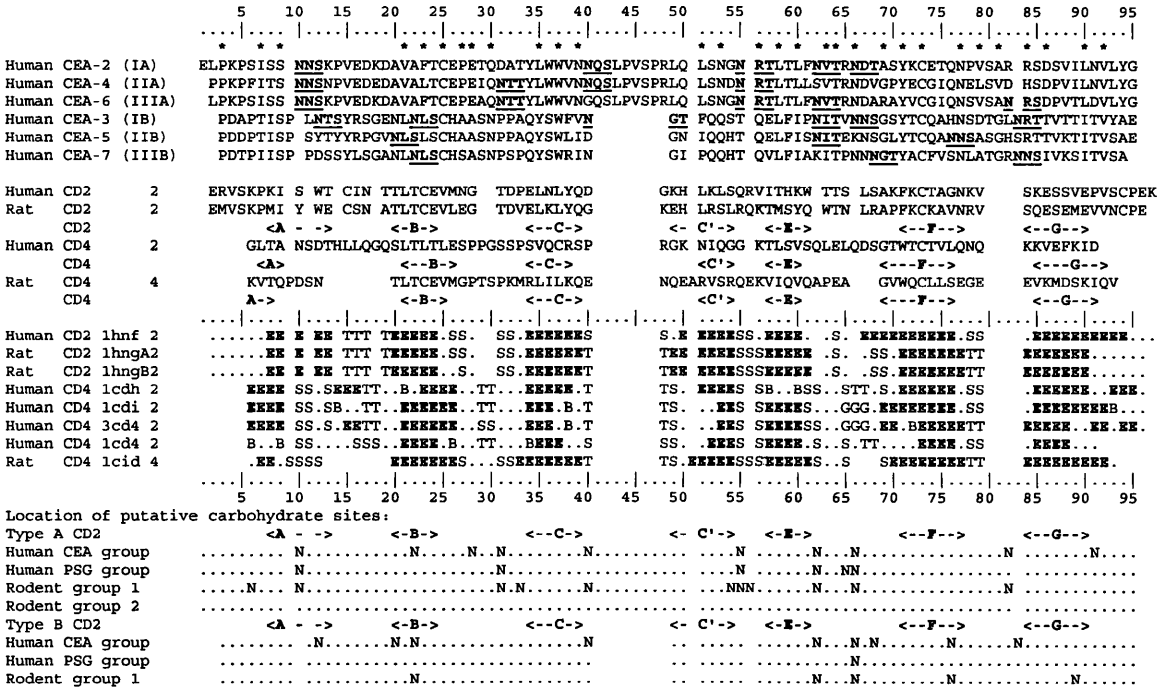


Figure 4.6. (*b*) Sequence alignment of CEA with CD2 and CD4 and their known structures (legend on page 154).

assigned in CEA from the sequence alignment. The DSSP analysis shows that only CD2 has the linker sequences ERVS and EMVS between the V- and C2-set domains which are analogous in length to the six links between the seven CEA domains. For these two reasons, the human CD2 crystal structure was selected for modelling the protein core of CEA.

Two different protein models for CEA were created. The "CD2-derived" model was based on the interdomain orientation of CD2 throughout CEA. The β-strands C, F and G (Figure 4.6) are structurally equivalent in the V- and C2-set domains (Jones *et al.*, 1992). α-Carbon atoms located within ß-strand C (Glu131, Leu132, Leu134), ß-strand F (Lys159, Lys161, Thr163) and ß-strand G (Val174, Glu175, Pro176) of the CD2 C2-set domain were readily superimposed upon those in the V-set domain (Asp32, Ile33, Trp35; Ile80, Lys82, Ser84; Ile97, Phe98, Asp99). After linking 12 V-set and C2-set domains in six CD2 structures, the five surplus V-set domains were deleted to leave a seven-domain CEA model. The C-terminal α-carbon of Pro180 and the N-terminal α-carbon of Glu104 between two adjacent C2-set domains was found to be separated by 0.42 nm. The "linear" CEA model was based on treating the seven domains in the CD2-derived model as independent objects, in which the first and last α-carbon atoms of Glu104 and Pro180 defined the X-axis (Figure 4.3; Methods). These α-carbon atoms were separated by 0.42 nm in neighbouring domains. The Y-plane was defined by the α-carbon atom of Phe160. The origin was set at the C-terminal residue Pro180 (Figure 4.3; Methods). All seven domains were aligned in the same relative orientation, and CEA models for curve fitting were generated by rotations of each domain about its origin.

Using the known carbohydrate composition of CEA and the similar sizes of each oligosaccharide as constraints, two types of oligosaccharide conformations were constructed. Examination of over 50 glycoprotein coordinate files in the Brookhaven database showed that the three largest oligosaccharide structures were those of Fc Kol, human leucocyte elastase and glucoamylase (Deisenhofer, 1981; Bode *et al.*, 1989; Aleshin *et al.*, 1992, 1994). The oligosaccharides found in the database showed that the first three residues GlcNAc.GlcNAc.Man have a common near-linear conformation by

virtue of β1→4 links between them. Despite the further addition of carbohydrate residues, the resulting oligosaccharide structures have similar spatial dimensions. Each oligosaccharide was represented by a triantennary complex-type structure $Man_3GlcNAc_6Gal_3Fuc_3NeuNAc_1$, whose structure was based on that found in Fc Kol (Figure 4.7; Methods). Each one was added at known putative sites in the CEA model according to the alignment of Figure 4.6. However, while the carbohydrate chains in human leucocyte elastase and CD2 have conformations extended from the protein surface (Bode *et al.*, 1989; Wyss *et al.*, 1995), glucoamylase is observed to possess compact oligosaccharide structures positioned against the protein surface (Aleshin *et al.*, 1992, 1994). The two conformational types were therefore represented in CEA models by positioning all 28 chains either extended away from or in proximity to the protein surface. This modelling procedure permits a realistic evaluation of the two possible extremes of carbohydrate structures in CEA.

4.3.4. Single-density X-ray scattering curve modelling for CEA

The objective of curve modelling was to show what family of domain structures best represented the solution structure of CEA. The automated curve-fitting procedure AUTOSCT (Beavil *et al.*, 1995) was applied to a range of rotamers calculated from the linear CEA model. This is based on the procedure previously calibrated with known single-domain crystal structures (Smith *et al.*, 1990; Perkins *et al.*, 1993). A full three-axis rotational search in 30° steps over 360° for the six interdomain links in CEA would involve 18 parameters and $(12 \times 12 \times 12)^6$ steps = 3×10^{19} models. As 2.5 min is required for each curve fit, the conformational search was only feasible for 10^3-10^4 models. The similarity of both the link regions and domain types in CEA (Figure 4.6) suggested that a simplified three-parameter approach could be usefully applied to CEA in which the same X-, Y- and Z-axis rotations were applied to all six interdomain interfaces. While this procedure is biased in that it excludes consideration of CEA models in which the six sets of rotations differ between the domains, such models are less likely. This procedure is adequate to cover a suitable survey of the major families of CEA structures while remaining within the precision of the technique.

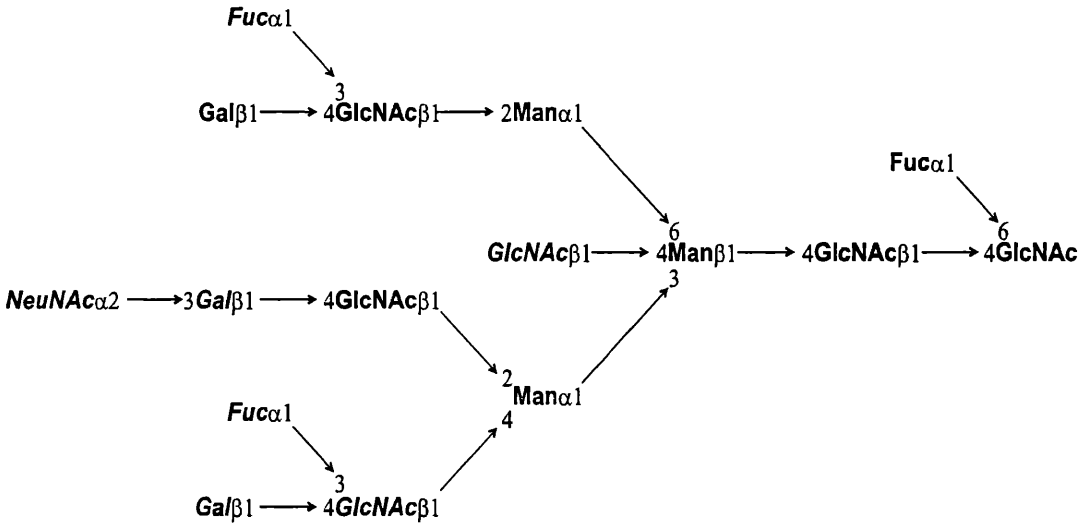


Figure 4.7. The averaged structure for a single oligosaccharide site on CEA. This was determined to be $Man_3GlcNAc_6Gal_3Fuc_3NeuNAc_1$ from the carbohydrate analysis of CEA (Yamashita *et al.*, 1987, 1989). A model of this was constructed using the nine-residue carbohydrate coordinates in the crystal structure of the Fc fragment of immunoglobulin G (Deisenhofer, 1981). Seven further residues (in italics) were added to the Fc coordinates to produce the complete oligosaccharide model for CEA.

Contour plots showed that a large proportion of possible CEA structural families could be rejected. Figure 4.8 illustrates how the calculated R_G and $R_{2.0}$ values depend on the Y-axis and Z-axis rotations between -90° and 90° in 15° increments when the X-axis rotation is fixed at 0°. The R_G values ranged from 8.2 nm for the starting linear model (Y = 0°, Z = 0°) to 3.1 nm for the most compact structure (Y = 90°, Z = 0°). The $R_{2.0}$ values ranged from 5.3% for the best curve fits (Y = -90°, Z = -90°) to 29.8% for the worst fit (Y = 90°, Z = 0°). The two contour plots showed that models with R_G values close to the experimental range of 8.0 ± 0.6 nm produced low $R_{2.0}$ values. Models with suitable $R_{2.0}$ values at the four corners of the contour map could be ruled out as these were incompatible with the R_G data.

The full search of CEA domain rotations between 0° and 345° in 15° steps about the X-axes, and between -90° and 90° in 15° steps about the Y- and Z-axes generated 24 × 13 × 13 = 4,056 single density models with extended carbohydrate chains. This calculation required 7 days of R4000 CPU time. Histograms showed that the $R_{2.0}$ values ranged from 4.3% to 30.4% (mostly less than 9%), the R_G values ranged from 3.09 nm to 8.34 nm, and the R_{XS} values ranged from 0.13 nm to 4.14 nm (mostly between 1.7 - 2.3 nm). Four families of structures could be distinguished, as defined by the structure of the protein core (Figure 4.9). None of the models gave consistently good curve fits, although zig-zag models gave the most promising fits, as shown by comparisons of the calculated and observed curves in the Q range of 0.09-1.0 nm^{-1} in Figure 4.9.

(i) The linear model (X = Y = Z = 0°) gave a poor X-ray curve fit, as shown also from the low R_{XS} value of 1.62 nm and high $R_{2.0}$ value of 8.1%, even though the R_G value is reasonable at 8.2 nm. Such a CEA structure was too elongated with too narrow a cross-section.

(ii) Curved CEA models resulted from rotations only about the Y- or Z-axes (domain tilting). While the curved model of Figure 4.9 (X = Y = 0°; Z = 30°) had reasonable R_G and R_{XS} values of 7.8 nm and 1.82 nm respectively, the $R_{2.0}$ value was high at 9.0%.

(iii) The combination of domain tilting with domain twisting about the X-axes produced zig-zag and helical CEA models. The zig-zag models have X-axes rotations between 90° and 270° for all Y-axes and Z-axes rotations (Table 4.2), and produced

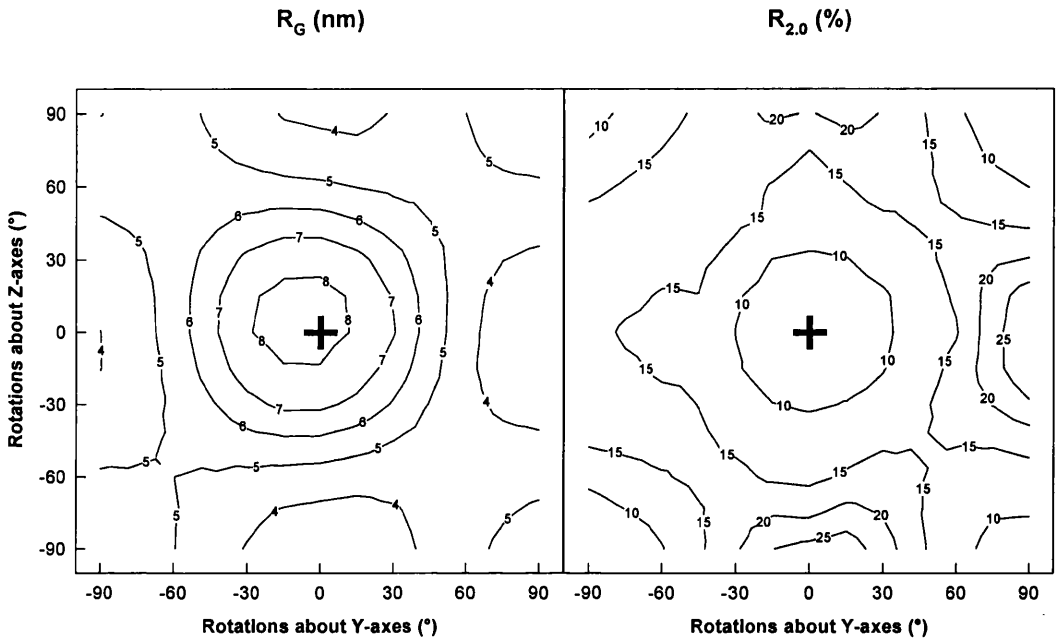


Figure 4.8. Contour maps of the dependence of the R_G and R-factor on domain rotations in the starting linear model of CEA. The X-axis rotation is set as 0°. The map was generated by domain rotations in 15° steps from -90° to 90° about the Y- and Z-axes. The starting linear model is obtained with Y = Z = 0° and is marked by a plus symbol.

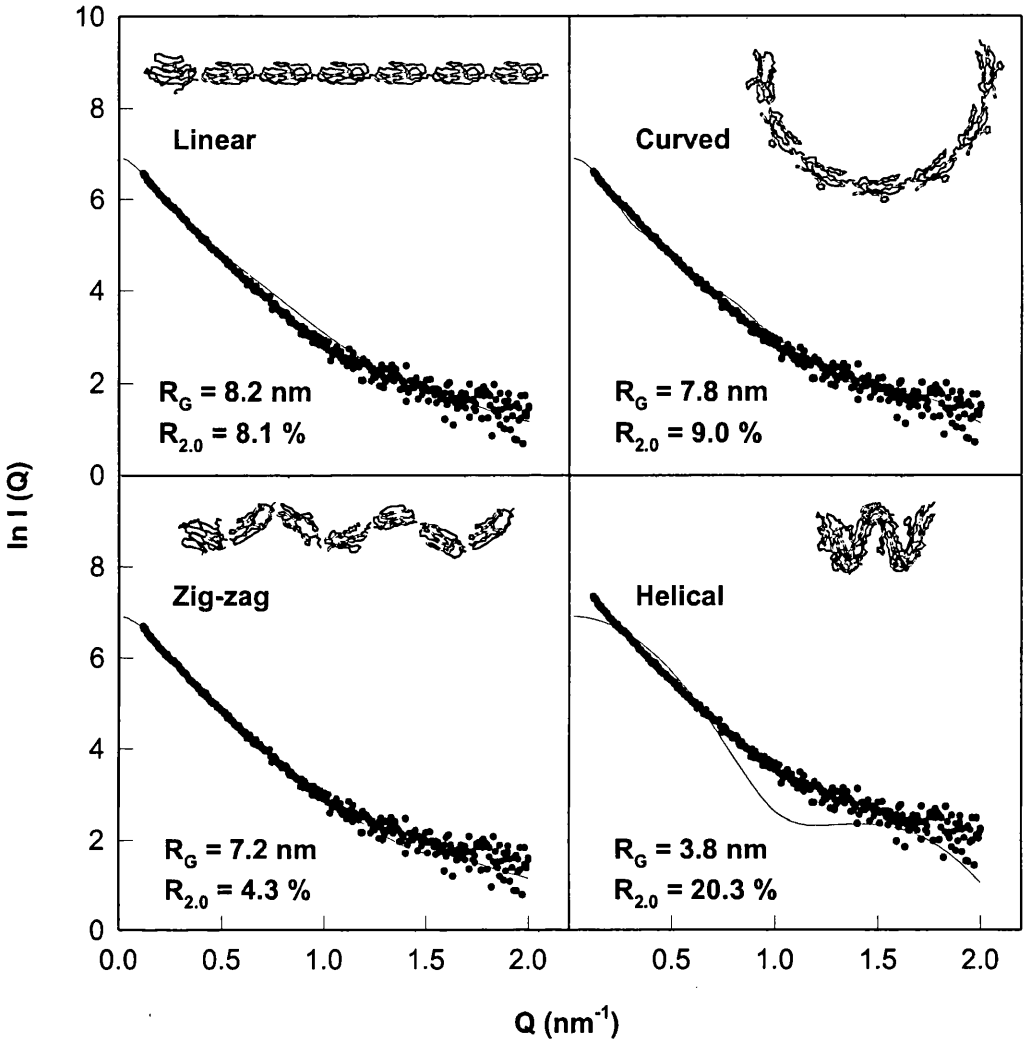


Figure 4.9. Comparison of the calculated scattering curves for four families of CEA structures with the X-ray scattering curves. Linear, curved, zig-zag and helical CEA models are depicted without the oligosaccharide chains in order to clarify the protein cores used to define each structural family. The *R*-factors were calculated using the X-ray experimental data out to $Q = 2.0$ nm^{-1}. Each model scattering curve is shown as a continuous line, and experimental curves are shown as filled circles.

Table 4.2. Summary of CD2, CD4 and CEA rotational angles that define their models

Model	Rotations (°): X-axis	Y-axis	Z-axis	R_G [a] (nm)	R_{XS} (nm)	$R_{2.0}$ (%)	$s^o_{20,w}$ (S)
Experimental values:				8.0 ± 0.6	2.1 ± 0.2	=	6.04±0.22
CEA scattering models:							
Sterically-allowed ranges	0 to 345	-90 to + 90	-90 to 90				
Single density and linear	0	0	0	(8.2)	(1.62)	(8.1)	5.7
Single density and curved	0	0	30	(7.8)	(1.82)	(9.0)	5.7
	0	0	(-90 to 90)				
Single density zig-zag (Figure 4.9)	120	30	60	(7.2)	(2.12)	(4.3)	6.1
	(90 to 270)	(-90 to 90)	(-90 to 90)				
Single density and helical	30	90	0	(3.8)	(3.54)	(20.3)	8.2
	(0-90 and 270-345)	(-90 to 90)	(-90 to 90)				
Double density zig-zag (Figure 4.10)	165	30	15	8.0	1.99	4.7	5.9
Double density CD2-derived (Figure 4.11)	=	=	=	7.9	2.10	4.9	6.1
Crystal structures (PDB code) [b]:							
Human CD2 (1hnf)	210	-10	-40	8.0	2.14	5.0	5.9
Rat CD2 (1hng A)	220	10	-50	7.7	2.16	5.0	5.9
Rat CD2 (1hng B)	220	5	-50	7.7	2.22	5.4	6.0
Human CD4 [c] (1cdh)	200	-45	-40	7.6	2.20	5.9	5.8
Rat CD4 (1cid)	230	-45	15	7.9	1.99	4.6	6.0

[a] The R_G, R_{XS} and $R_{2.0}$ values correspond to two-density models for X-ray curve fits except when single-density models (bracketted) are specified.

[b] Rotational angles were derived from the following crystal structures to apply to the linear CEA model (see Materials and Methods) to generate these structures.

[c] Other human CD4 structures (1cdi, 3cd4, 1cd4) give similar rotational angles.

elongated structures. That in Figure 4.9 (X = 120°; Y = 30°; Z = 60°) gave the most reasonable curve fit with a satisfactory R_{XS} of 2.12 nm, and a low $R_{2.0}$ value of 4.3%, although the R_G value of 7.2 nm was low.

(iv) Helical models were primarily generated from X-axes rotations in the remaining ranges of 0° to 90° and 270° to 345°, and had more compact structures. The helical model (X= 30°; Y = 90°; Z = 0°) was too compact as evidenced by the low R_G value of 3.8 nm, the high R_{XS} of 3.54 nm, and the high $R_{2.0}$ value of 20.3%.

4.3.5. Control of X-ray and neutron scattering curve modelling for CEA

Control calculations were performed to assess the 4,056 CEA models further in sections (v), (vi) and (vii) below.

(v) Based on previously tested procedures (Smith *et al.*, 1990; Perkins *et al.*, 1993), curve fitting was now applied jointly to the X-ray and neutron scattering curves of CEA. This required consideration of the different scattering densities of protein and carbohydrate of CEA. Their electron densities were 419 $e.nm^{-3}$ and 492 $e.nm^{-3}$ respectively, compared to that of water at 334 $e.nm^{-3}$ (Perkins, 1986), and the ratio of electron densities of protein: carbohydrate was 1.00 : 1.86. The neutron scattering densities were equivalent to 42.3% and 46.0% ^2H_2O respectively (Perkins, 1986), which gives a ratio of 1.00 : 0.93. The protein and carbohydrate spheres in the CEA models were assigned weight of 2 : 3 and 1 : 1 in 4,056 X-ray and 4,056 neutron curve fits respectively. The joint analysis also involved hydrated models in the X-ray modelling, but no corrections for beam effects, while dry models were used in neutron modelling but the curves require corrections for beam wavelength spread and divergence.

To follow Beavil *et al.* (1995), the two-density model curve fits were first sorted in order of the *R*-factors, then the models were filtered to ensure that at least 460 protein and 460 carbohydrate spheres were present in each model (i.e. no steric overlap), the R_G values were between 7.4 nm and 8.6 nm, and the R_{XS} values were between 1.9 nm and 2.3 nm. The filtering procedure resulted in a limited family of models. Comparison of the 100 and 400 best-fit models gave similar outcomes, and this showed that the analysis was stable. The 100 best-fit models corresponded to the zig-zag family with a mean X-axis rotation of 160° ± 25°, Y-axis rotation of 10° ± 30° and Z-axis rotation of -5° ± 35°.

Interestingly, the mean R_G of 7.8 ± 0.2 nm and R_{XS} of 2.02 ± 0.06 nm were now both within error of experiment (Table 4.2). Figure 4.10 shows that a two-density zig-zag CEA model close to the mean with X= 165°, Y = 30° and Z = 15° (Figure 4.11b) gave a good fit to the experimental X-ray data (R_G = 8.00 nm; R_{XS} = 1.99 nm; $R_{2.0}$ = 4.7%. The neutron model had R_G = 6.9 nm, R_{XS} = 1.73 nm, $R_{1.6}$ = 8.7%. While the quality of the neutron curve is poorer, and the R_G is less than that observed, it should be noted that the neutron and X-ray curves are noticeably different in Figure 4.10. Despite this difference, the two-density zig-zag model is able to offer a good curve fit for the neutron data. Comparison of the single- and two-density models in Table 4.2 shows that the X-, Y- and Z-axis rotations were within 45° of each other. Allowance for two densities has now selected CEA models for which both the X-ray R_G and R_{XS} values agreed well with observation, unlike the single density model which gave too low an R_G value. In other words, the single density CEA models had increased their diameter by coiling to compensate for the reduced weight of carbohydrate, which then causes the CEA models to become shorter in overall length (smaller R_G). The use of two-density models avoided this effect to offer good X-ray R_G and R_{XS} values.

(vi) A two-density CEA model was derived directly from the crystal structure of human CD2 (Figure 4.11c; Methods) for comparison with the zig-zag model. This had interdomain rotations equivalent to X = 210°, Y = -10° and Z = -40° (Table 4.2), which are within 1-2 standard deviations of the best-fit two-density zig-zag model, and therefore was structurally similar to this (Figure 4.11b). Table 4.2 shows that the double-density CD2-derived model, and those derived from human and rat CD2 all gave comparable X-ray R_G, R_{XS} and $R_{2.0}$ values to those of the double-density zig-zag model (curve fits not shown). The rotations for human and rat CD4 are comparable with those for CD2, despite the small differences noted by Jones *et al.* (1992), and Table 4.2 shows that these CD4-derived models also generated comparable good agreements. The success of CEA models based on CD2 crystal structures is further support for a family of zig-zag structures for CEA (Figures 4.9b and 4.9c).

(vii) The premise that CEA has extended oligosaccharide chains on its surface was tested by performing curve simulations with 4,056 single- and double-density

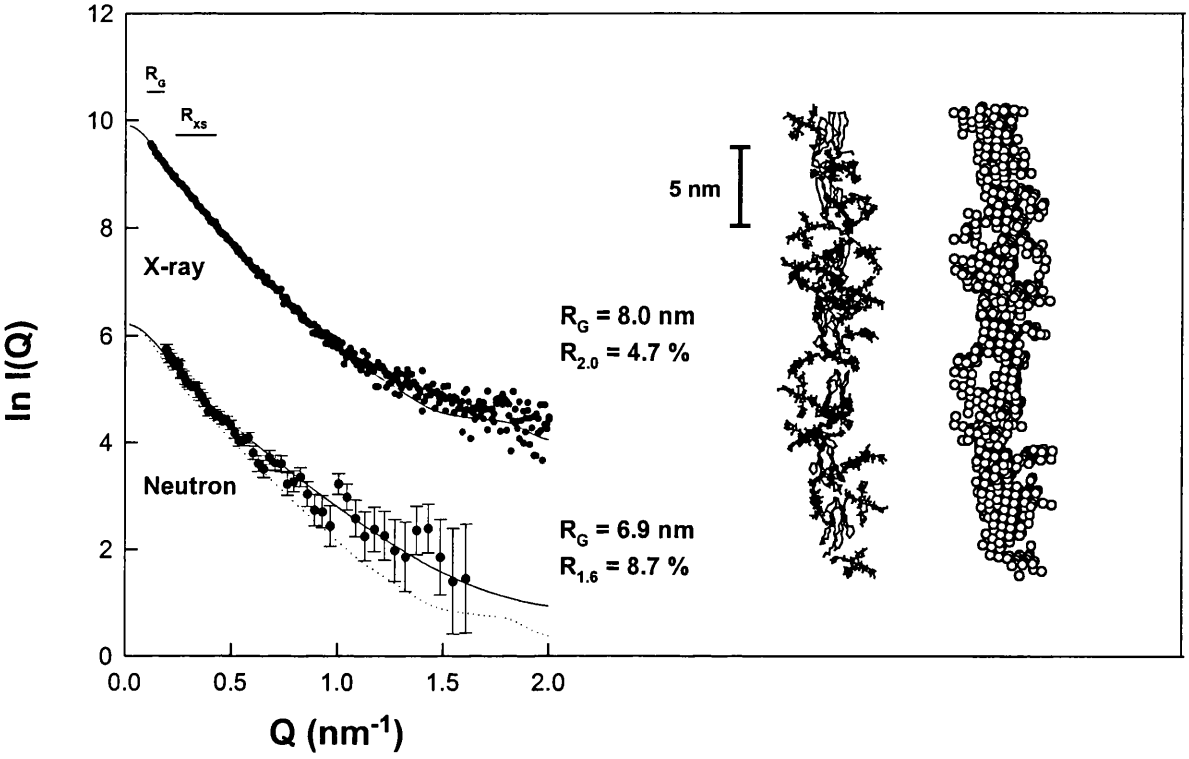


Figure 4.10. Comparison of the simulated X-ray and neutron scattering curves for the best-fit CEA model with experimental X-ray and neutron data. The seven CEA domains in each model are shown as α-carbon traces, whereas the carbohydrate chains are represented in full. The corresponding sphere model (sphere diameter, 0.572 nm) is also shown. A two-density hydrated sphere model with a protein : carbohydrate density ratio of 2 : 3 was used to calculate the X-ray scattering curve, and a single density unhydrated model was used for the neutron curve. The calculated neutron curve was corrected for wavelength resolution and beam divergence. For the experimental X-ray curve (●), the R_G value is 8.0 nm and the $R_{2.0}$ value for the X-ray data is 4.4%. For the experimental neutron curve (●), the R_G value is 7.3 nm and the $R_{1.6}$ value is 7.9%. The calculated X-ray curve is shown as a dashed line for comparison purposes with the neutron curve. The Q-ranges used for the R_G and R_{XS} analyses (Figure 4.4) are denoted by horizontal bars.

Figure 4.11. (Overleaf) Molecular graphics stereoviews of the final zig-zag and CD2-derived models for CEA. The domains are shown as α-carbon traces, whereas the carbohydrate chains are represented in full. The immunoglobulin G model from Mayans *et al.* (1995) is shown on the same scale for comparison with its Fv region boxed.

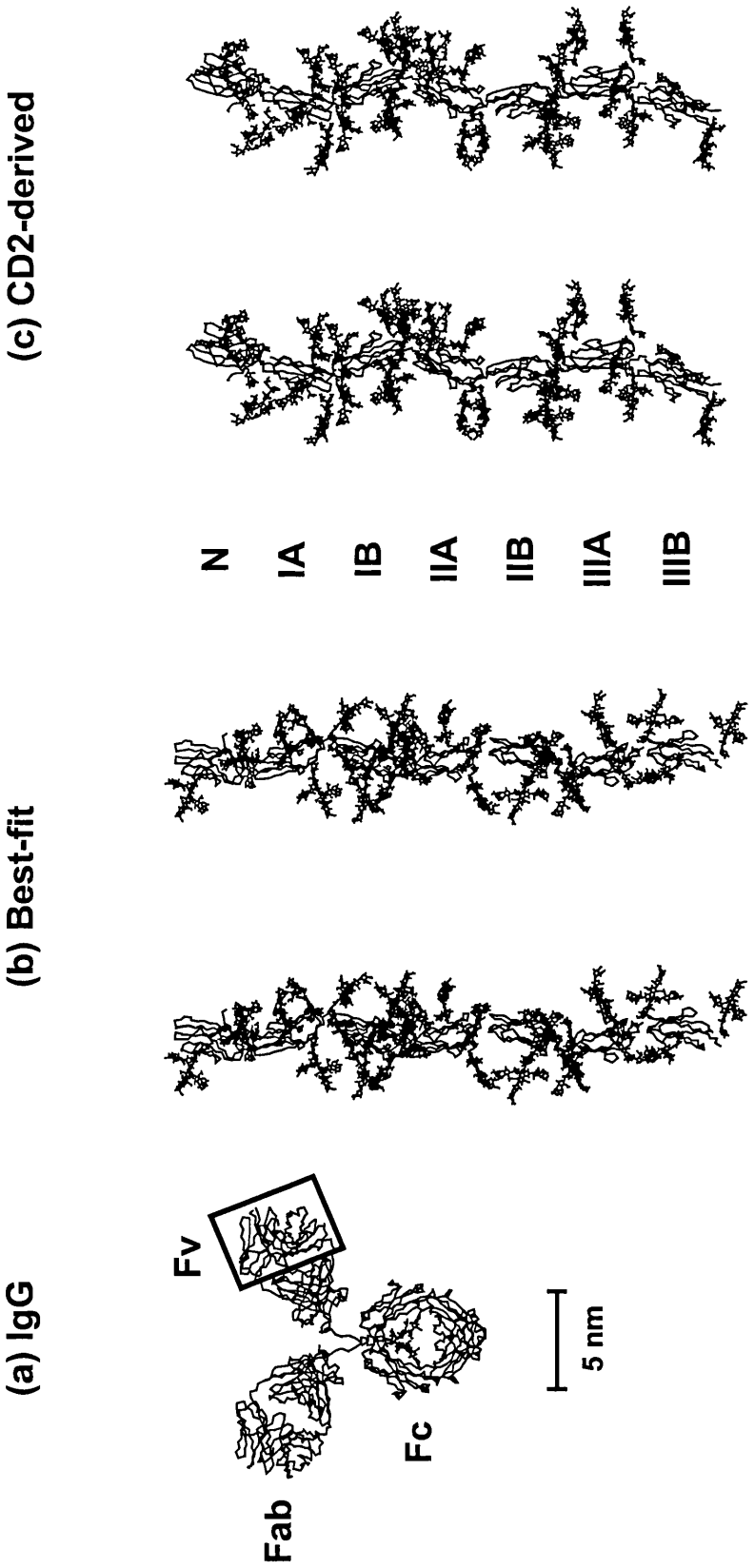


Figure 4.11. Molecular graphics stereoviews of the final zig-zag and CD2-derived models for CEA (legend on page 167).

models in which the oligosaccharide chains were positioned close to the protein surface. The sorting of the curve fits based on $R_{2.0}$ values and filtering based on the observed R_G and R_{XS} parameters showed that it was not possible to match the R_G and R_{XS} values simultaneously. The best X-ray models from these searches exhibited low R_G values of 5.9-6.1 nm, satisfactory R_{XS} values of 2.21-2.25 nm, and poorer $R_{2.0}$ values of 6.9-7.8%. The recalculation of 4,056 neutron curve fits confirmed this result. This shows that the CEA carbohydrate structures are generally extended in their structures, rather than being generally compact.

4.3.6. Sedimentation velocity of CEA and its hydrodynamic modelling

Sedimentation coefficients $s^{\circ}_{20,w}$ are a monitor of macromolecular elongation akin to R_G data, and provide a control of the scattering data. In sedimentation velocity runs, the mean $s^{\circ}_{20,w}$ value from three CEA samples between 0.42-0.83 mg/ml was determined to be 6.04 ± 0.22 S. From this, a high frictional ratio f/f_o of 1.80 was calculated, where f_o is the frictional coefficient of a sphere with the same hydrated volume as that of CEA. The value of $s^{\circ}_{20,w}$ is comparable with the range of previously published values of 5.47 S - 8.04 S (Krupey *et al.*, 1968) and 6.8 S (Slayter & Coligan, 1975).

Procedures for hydrodynamic simulations based on sphere models have been tested in Smith *et al.* (1990) and Perkins *et al.* (1993). The GENDIA program was used to calculate the $s^{\circ}_{20,w}$ values from the 4,056 hydrated X-ray CEA models, each with about 959 spheres. This required 30 days of R3000 CPU time. The 100 best two-density X-ray models gave $s^{\circ}_{20,w}$ values between 5.82 S to 6.26 S. The mean $s^{\circ}_{20,w}$ value was 6.00 ± 0.10 S. The two CEA models of Figure 4.11 gave $s^{\circ}_{20,w}$ values of 5.9 S and 6.1 S (Table 4.2). As these agreements were within the acceptable precision of ± 0.3 S for these calculations (Perkins *et al.*, 1993), the hydrodynamic data and their modelling provide further support for the CEA zig-zag model derived from X-ray curve fits. Table 4.2 showed that the zig-zag model offered better agreement than the linear, curved and helical models for CEA.

4.4. Conclusions

4.4.1. Low resolution models for CEA

In its natural state, CEA occurs primarily in the colonic epithelium, with increased expression in colon cancer. It functions as a cell adhesion molecule (Benchimol *et al.*, 1989; Oikawa *et al.*, 1991; Zhou *et al.*, 1993). Solution scattering has resulted in an improved understanding of its solution structure. Up to now, it was not clear that CEA is monomeric. The neutron molecular weight agreed well with a monomer molecular weight of 152,500. The success of the scattering curve fits likewise indicated that CEA is monomeric. Gel filtration and SDS-PAGE studies on CEA have generally found that this is monomeric (Slayter & Coligan, 1975), although Lisowska *et al.* (1983) have suggested that CEA may dimerize depending on the conditions of sample preparation. The monomeric state of CEA as studied here implies that the homotypic cell-adhesion interactions between individual CEA molecules on different cells are weak, and that a large number of CEA molecules are required to generate a significant adhesive interaction between cells.

The present X-ray scattering data are more precise than the neutron data. These indicate that CEA is 27-33 nm in length and up to 8 nm in width, and contains extended carbohydrate structures. Knowledge of the CEA composition and curve modelling based on the structural homology with CD2 resulted in two rod-like models of length 27 nm and width 8 nm (Figure 4.11). In agreement with this, electron microscopy has visualised CEA as distinctive rod- or crueller-shaped macromolecules with larger dimensions of 9 × 40 nm (Slayter & Coligan, 1975), although the rotary shadowing method in use can overestimate macromolecular dimensions (Slayter & Codington, 1973). Rod-shaped structures have also been reported for related members of the Ig superfamily such as neural cell adhesion molecule (N-CAM) and intercellular adhesion molecule-1 (I-CAM) by electron microscopy (Becker *et al.*, 1989).

The application of automated curve modelling to the domains of CEA demonstrates the applicability of this method to multidomain proteins that cannot in all probability be crystallised. The present curve modelling was constrained on the basis of the protein sequence, the carbohydrate composition, the CD2 crystal structure and the

distance between its two domains, and the carbohydrate conformations seen in known crystal structures. The curve fits were implemented on the basis of tested scattering densities and procedures for curve modelling. In combination with the experimental X-ray curve, these constraints place limits on molecular structures for CEA. For CEA, the definition of a full range of interdomain rotations enables a basic set of 4,056 models to be tested against experiment. The two time-consuming stages were the initial definition of a suitable search strategy, which will differ from protein to protein, and the length of the simulations themselves. The advantages of this procedure are that a full range of rotational structures is tested automatically, the assumptions in CEA curve modelling could be tested systematically, and the statistical precision of the curve fitting can be defined. In this way, a basic set of four major conformational families from 4,056 CEA models could be identified, of which only one resulted in good curve fits. The repetitive evaluation of 4,056 models to test the effect of one- or two-density modelling, the use of the CD2 crystal structure, the use of extended or compact carbohydrate structures, and hydrodynamic sedimentation coefficients consistently indicated that the zig-zag family of CEA structures offered the most satisfactory account of the CEA scattering curve.

In relation to the final low resolution structure of CEA, it should be noted that solution scattering only shows CEA structures that are compatible with scattering curves, and do not determine a unique structure. The X-ray curves close to Q of 1 nm^{-1} were sensitive to the relative orientations of the seven domains in CEA. Analyses based on R_G or R_{XS} values alone were restricted to the Q range of 0.09 - 0.18 nm^{-1} and 0.23 - 0.42 nm^{-1}, and were therefore less discriminatory than the R-factor (Figure 4.9). All three parameters, together with sedimentation coefficients, were nonetheless useful to assess stereochemically-correct models. The existence of six inter-domain links in CEA implies that as many as 10^{19} models should be tested to explore all inter-domain orientations, of which a significant small proportion will give good curve fits. While several types of interdomain conformation may exist, as observed in the two molecules in the crystal structure of rat CD2 (Jones *et al.*, 1992), it is not possible to allow for this in the simulations. Here, the assumption that all six interdomain links are similar in CEA permitted a realistic survey of allowed conformations. In summary, by two approaches, the present modellings were able to demonstrate that the two-domain crystal

structures of CD2 (and CD4 as well) offer a good explanation of the overall solution structure of CEA.

4.4.2. Implications of the CEA structure for biological activity

The determination that CEA has an extended zig-zag structure in solution (Figure 4.11) implies that such a structure on cell membranes would protrude perpendicular from the membrane into solution. All seven domains would then be accessible for protein-protein interactions. A CEA N-terminal V-set domain and one of the C2-set domains are both required to be present for homotypic cell-adhesion interactions (Oikawa *et al.*, 1991; Zhou *et al.*, 1993; Hashino *et al.*, 1993). The IIIA-IIIB domain pair of CEA is involved, although interactions with the IA-IB and IIA-IIB domain pairs have not been excluded.

In terms of the CEA model, the CD2 and CD4 crystal structures show that each C2-set Ig fold is approximately half-rotated about its X-axis relative to its neighbour such that the EBA and GFCC' β-sheets (Figure 4.6b) in the β-sheet sandwich of each C2-set Ig fold will present alternate faces along one side of the long axis of CEA. Note that this steric relationship cannot be deduced at the resolution of solution scattering, and an atomic structure determination will be required to verify this. Such a half-rotation is also seen in the crystal structure of tissue factor and the growth hormone receptor, both of which have two adjacent fibronectin type III domains that resemble Ig fold domains (Chapter 1; de Vos *et al.*, 1992; Harlos *et al.*, 1994; Campbell & Spitzfaden, 1994). This arrangement is attributed to the manner in which the last β-strand G of the first domain runs almost unchanged in direction into the first β-strand A of the second domain, and this packing scheme causes the two Ig folds to become twisted relative to one another (see Campbell & Spitzfaden, 1994 for a discussion).

In members of the Ig superfamily, the AGFCC'C" β-sheet in the V-set domain has been associated with the CD2 and CD4 ligand binding faces (Wang *et al.*, 1990; Ryu *et al.*, 1990; Jones *et al.*, 1992; Bodian *et al.*, 1994), and the GFCC' β-sheet in the C2-set second domain of the α-chain of the IgE FcεRI receptor is the determinant of the IgE-receptor interaction (Beavil *et al.*, 1993), and the AGFCC'C" face in VCAM-1 is

likewise involved in ligand binding (Jones *et al.*, 1995). The GFCC' β-sheet in the structurally-related C2-set second domain of tissue factor has been implicated in interactions with factor VIIa (Harlos *et al.*, 1994). Computer modelling has suggested that the GFCC' β-sheet in the C2-set first domain of the intercellular adhesion molecule-1 (ICAM-1, CD54) has been implicated in binding to the integrin LFA-1 (Staunton *et al.*, 1990; Berendt *et al.*, 1992). It is thus tempting to propose that the GFC β-sheet forms possible protein ligand sites for the interactions between different CEA molecules from different cells. If this is the case, such adhesion sites on the IA-IB, IIA-IIB and IIIA-IIIB domains would be predicted from Figure 4.11 to be equally presented on two or three sides of the long axis of CEA through rotational twisting of the domains to facilitate these interactions.

The location of the 28 putative carbohydrate sites in CEA supports the possible role of the GFCC' β-sheet of the C2-set domains in CEA adhesion. From the sequence alignment of Figure 4.6, 20 sites are predicted to lie on exposed loops and only 8 are positioned on β-strands. Carbohydrate sites are not found on the β-strands of the V-set domain. Carbohydrate sites are found at the centre of either strand A or B in all six of the EBA β-sheets in the C2-set domains of human CEA. These EBA β-sheets are unlikely to present protein ligand sites to a CEA V-set domain. In contrast, five of the six GFCC' β-sheets have no carbohydrate sites and are potentially available for antiparallel homotypic CEA interactions. The single exception is an oligosaccharide site at the end of strand F in domain IIB. Molecular graphics inspection of the best-fit and CD2-derived models in Figure 4.11 shows that the six GFCC' faces in human CEA were free of steric hindrance caused by the extended carbohydrate structures in CEA.

Sequences for the human CEA gene family, which includes CEA, nonspecific cross-reacting antigen, biliary glycoprotein and CEA-group members (Thompson *et al.*, 1991), were obtained from the Entrez CD-ROM database Release 17.0 (June 1995). These confirm the accessibility of the GFCC' β-sheet on the basis of putative carbohydrate sites.

(i) None of 3-4 putative oligosaccharide sites in 58 V-set sequences occur in the β-sheet AGFCC'C", apart from two exceptions at the end of strand F (bottom of Figure

4.6a). This holds also for 70 sequences from the related human pregnancy-specific glycoprotein (PSG) gene family (Thompson *et al.*, 1991) and 31 rodent group 1 sequences, but not for 41 rodent group 2 sequences, although the rodent groups have a different domain organisation to the human groups.

(ii) The 67 type A and 61 type B C2-set sequences of the human CEA gene family contain between 3-6 putative oligosaccharide sites in a single domain. The 128 sequences generally showed carbohydrate-free GFCC' β-sheets, with the occasional exception of the end of strand F in the type B sequences (bottom of Figure 4.6b). While the human PSG gene family showed reduced glycosylation levels in the C2-set domains, this also conforms to carbohydrate-free GFCC' β-sheets in 130 type A and 79 type B sequences. The analysis of rodent group 1 and group 2 C2-set domains (57 type A and 4 type B sequences in group 1; 25 type A sequences in group 2) also demonstrates accessible GFCC' β-sheets and blocked EBA β-sheets.

Two hypothetical models for an antiparallel homotypic interaction of CEA molecules from different cells are shown in Figure 4.12. Two CEA structures can be positioned such that the AGFCC'C" face of a V-set domain is close to a GFCC' face of a C2-set domain. The domain pair in question is dependent on the assumed rotational twist along the long axis of CEA, and this is not known. Figure 4.12 shows how an interaction can occur with either the IB or IIB or IIIB domains, or with the IIA or IIIA domains. The models show that the GFCC' face of each C2-set domain is angled at about 45° to the vertical. This presents an upward-angled surface that can match with its complementary V-set domain. Such a match is not possible with the EBA faces in the C2-set domains of CEA, all of which are angled downwards to face the cell membrane.

The best-fit and CD2-derived CEA models can be compared with a model for immunoglobulin G in order to assess binding sites for anti-CEA antibodies (Figure 4.11). There is concern that the high carbohydrate content may hinder the binding of antibodies to protein epitopes (Thompson *et al.*, 1991). Mapping demonstrates that epitopes to monoclonal antibodies are present on all seven domains of CEA (Schwarz *et al.*, 1988; Ikeda *et al.* 1992; Murakami *et al.* 1995). In a survey of 52 anti-CEA

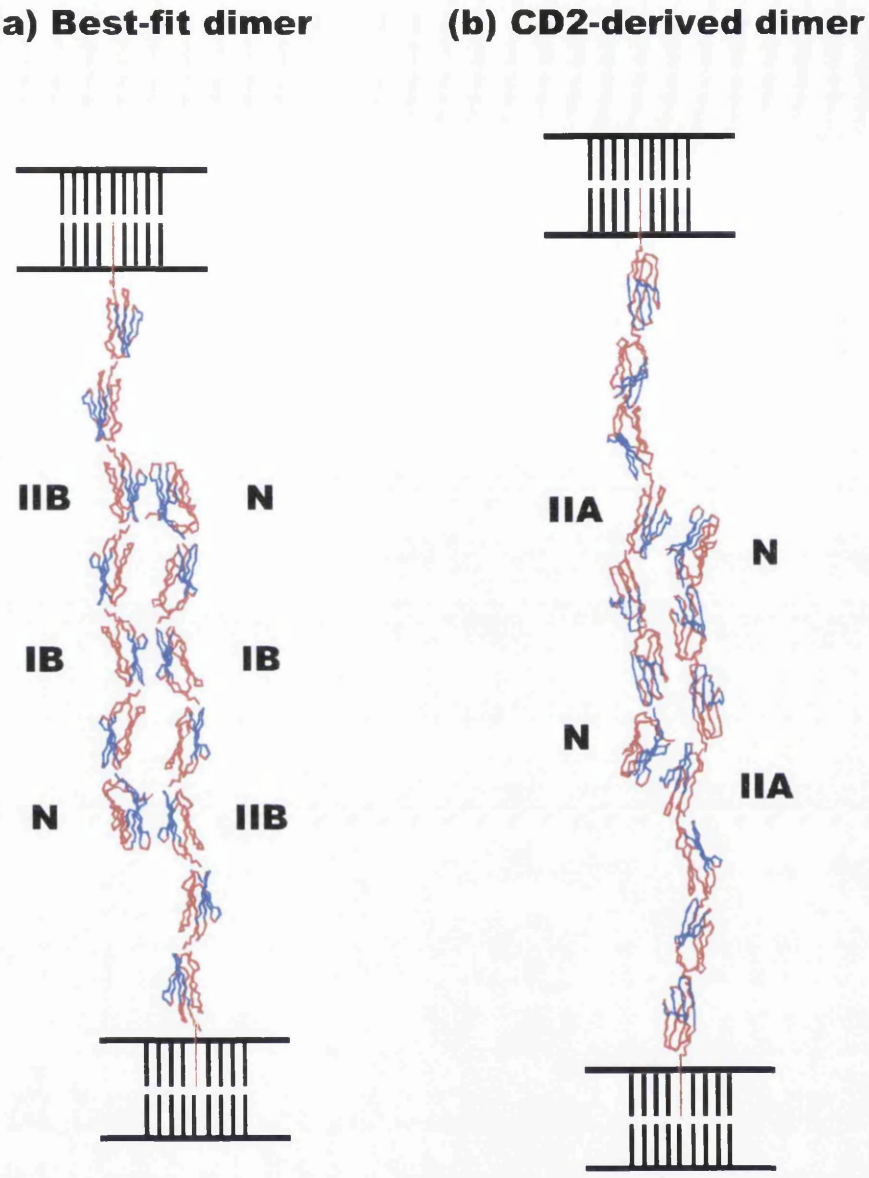


Figure 4.12. Schematic possible models for the homotypic interaction between two CEA molecules from different cells. The CEA α-carbon trace of its polypeptide chain is depicted in red, except for the GFCC' β-sheet which is shown in blue. No carbohydrate is shown for reason of clarity. The C-terminal membrane anchor of CEA is shown in red. The best-fit model from Figure 4.11b shows GFCC' β-sheet interactions between the N- IIB pair of domains and also between the IB-IB pair of domains. The CD2-derived model from Figure 4.11c shows an interaction between the N-IIA pair of domains. It is conceivable that the exposed non-glycosylated GFCC' face on six of the seven Ig fold domains in CEA may present many of these epitopes. The extended carbohydrate conformations on CEA occur mostly on loops located at the six junctions between domains or at the EBA face in the six C2-set domains, and these should not restrict access to these GFCC' faces. The CEA models are consistent with what is known of anti-CEA antibody interactions.

monoclonal antibodies, 43 of these were shown to bind to one of five independent noninteracting protein epitopes GOLD 1-5, while the rest were directed against carbohydrate epitopes or were inactive (Hammarström *et al.*, 1989). The use of deletion mutants of CEA showed that GOLD 5 antibodies reacted with domain N, GOLD 4 antibodies reacted with domains IA-IB and IIA-IIB, GOLD 2 antibodies reacted with domains IIA-IIB, GOLD 1 antibodies reacted with domains N, IIA-IIB and IIIA-IIIB, and GOLD 3 antibodies reacted with domains N and IIIA-IIIB (Murakami *et al.*, 1995). For this to occur, an exposed protein surface on CEA of area at least 4 nm × 5 nm must be available to permit an antibody Fv fragment to interact with these five independent epitopes. Figure 4.11 shows that the size of the antibody Fv fragment is comparable to that of a single CEA domain face. It is thus conceivable that the exposed non-glycosylated GFCC' face on six of the seven Ig fold domains in CEA may present many of these epitopes. The extended carbohydrate conformations on CEA occur mostly on loops located at the six junctions between domains or at the EBA face in the six C2-set domains, and these should not restrict access to these GFCC' faces. The CEA models are consistent with what is known of anti-CEA antibody interactions.

Chapter 5

Arrangement of the Fab and Fc Fragments in Human IgA1 by X-ray and Neutron Scattering and a Comparison with IgG

5.1. Introduction

Antibody therapies for CEA require structural knowledge of intact antibody molecules. In humans there are nine immunoglobulin isotypes (IgA1, IgA2, IgD, IgE, IgG1, IgG2, IgG3, IgG4 and IgM). The most abundant ones are IgA and IgG. Each isotype is the product of a separate heavy chain constant region gene segment (termed C_H gene segment), which is denoted by the corresponding Greek character (therefore the gene segment that determines the IgG1 isotype is denoted $C_\gamma 1$). The antibody isotype produced by a B-cell is governed by specific rearrangements and fusions of coding segments (for reviews see Harriman *et al.*, 1993; Burrows *et al.*, 1995). In the heavy chain multigene family, which consists of V_H, D_H, J_H and C_H gene segments, the gene segments are arranged linearly in a fashion dictated by the structure of the heavy chain protein and the developmental appearance of the immunoglobulin isotypes during the course of an immune response. When a mature B-cell is released from the bone marrow it expresses membrane-bound forms of IgM and IgD on its surface. The binding of antigen to cell-surface IgM induces the activation and proliferation of the B-cell expressing it. For most types of antigen, B-cell activation also requires the interaction of the B-cell with a T_H-cell: T_H-cells produce cytokines that are instrumental in switching the class of antibody produced by a B-cell from IgM to alternative isotypes, and induce the proliferation of B-cells into plasma cells that produce soluble immunoglobulins. The different immunoglobulin isotypes are associated with different effector functions. The structure of IgG antibodies by crystallography and by neutron scattering is well understood (Mayans *et al.*, 1995; Harris *et al.*, 1998b). The remainder of this chapter is concerned with the structure and function of the human IgA class of immunoglobulin, the other major antibody in serum, in particular to investigate how the Fab and Fc fragments are arranged and how this compares to that in IgG.

5.1.1. Overview of the domain structure of IgA

An IgA monomer consists of two identical light (κ or λ) and two identical heavy polypeptide chains (Figure 5.1). The four chains fold up into twelve domains, and all of the domains have the characteristic Ig β-sandwich structure. The domain topology is DEBA|GFCC'C" for the nine β-strand V-set Ig domains and DEBA|GFC(C') for the seven (or eight) β-strand C1-set Ig domains (Bork *et al.*, 1994; Jones & Chothia, 1997).

Figure 5.1. (Overleaf) Schematic diagram of the Ig fold domains in human IgA1 and IgG1. Each heavy chain contains the V_H, C_H1, C_H2 and C_H3 domains, and each light chain contains the V_L and C_L domains, each of which is represented by a rectangle. The four-domain Fab fragments are linked to the four-domain Fc fragment by a 23-residue hinge in IgA1. The C-terminus of each IgA1 heavy chain contains an 18-residue tailpiece (dashed). The hinge of IgG1 is one residue shorter and there is no tailpiece. Each Ig fold contains a conserved internal Cys-Cys disulphide bridge (S-S). The C_H1 domain of IgA1 has an additional disulphide bridge between Cys196-Cys220. The heavy and light chains of IgA1 are linked between Cys133 in the C_H1 domain, and the C-terminal Cys residue of the light chain. The heavy and light chains of IgG1 are linked by a Cys residue in the hinge and the C-terminal Cys residue of the light chain. In IgA1, Cys241 and Cys242 in the hinge and Cys299 and Cys301 in the C_H2 domain form intra- and inter-heavy chain disulphide bridges. In IgG1, two Cys residues in the hinge form two inter-chain bridges. In IgA1, Cys311 on the C_H2 domain and Cys471 in the tailpiece are shown free. IgA1 has two N-linked oligosaccharide sites on β-strand B of the C_H2 domain and on the tailpiece (●). That on the C_H2 domain of IgA1 is at a different position to that on the C_H2 domain of IgG1 which occupies a central cavity in the IgG1 Fc structure and which is conserved in the other human Ig classes. IgA1 also has five O-linked oligosaccharide sites in the hinge (○) which are not present in IgG1. PTerm455 is a recombinant IgA1 molecule which lacks the tailpiece.

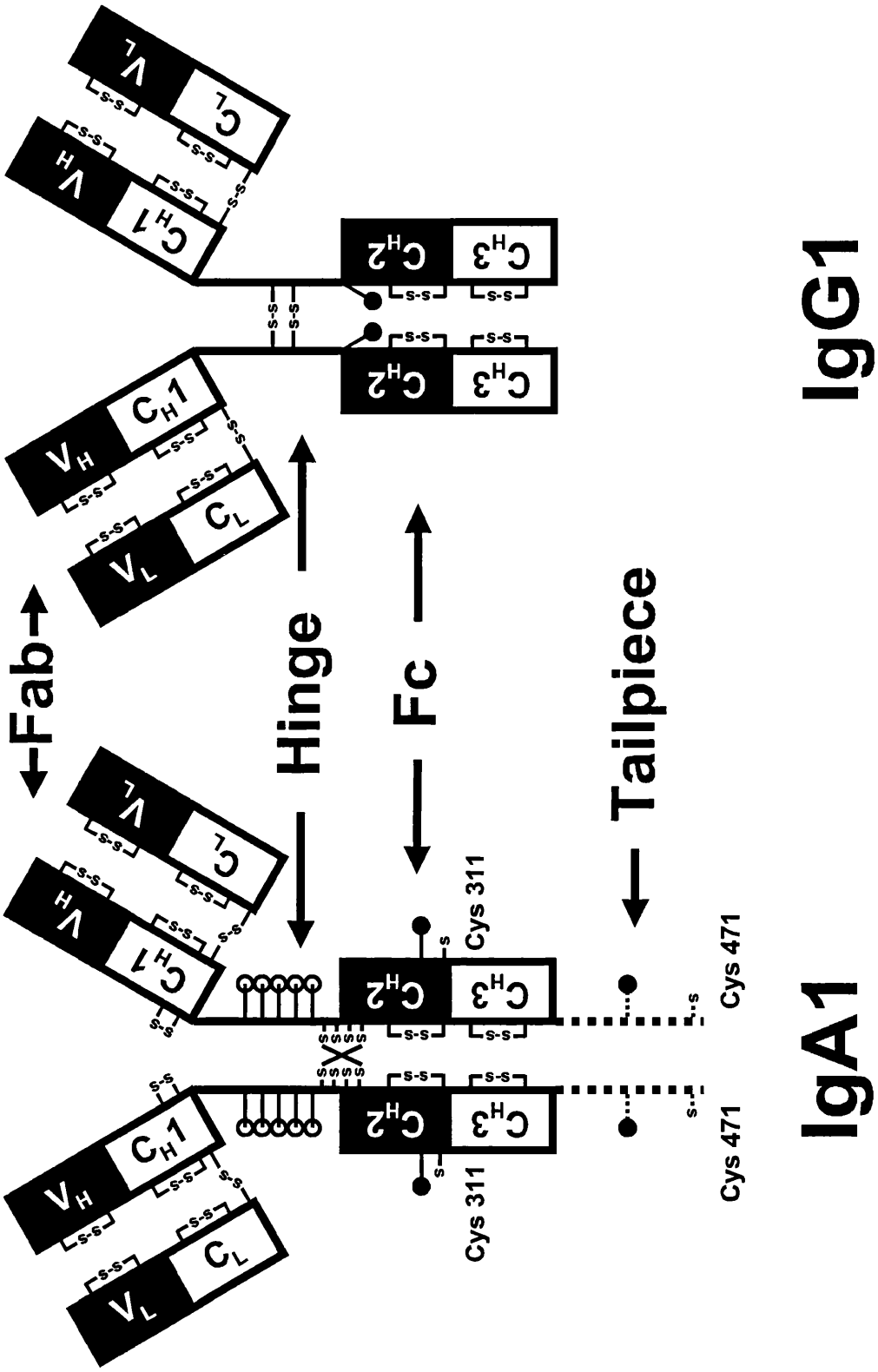


Figure 5.1. Schematic diagram of the Ig fold domains in human IgA1 and IgG1 (legend on page 179).

Each light chain has an N-terminal V-set Ig domain (V_L) and a C1-set Ig domain (C_L). Each heavy chain has an N-terminal V-set Ig domain (V_H) followed by three C1-set Ig domains (C_H1, C_H2 and C_H3). There is an 18 residue tailpiece at the C-terminus of the C_H3 domains. In immunoglobulin heavy chains, the whole of the heavy chain except for its V-set domain is encoded by a C_H gene segment. Therefore, the sequence and hence the functional differences that characterize a particular isotype are localized to this region. The IgA1 isotype is encoded by the $C_\alpha1$ gene segment and the IgA2 isotype is encoded by the $C_\alpha2$ gene segment. In addition, IgA2 exists as two known allotypes, namely IgA2m(1) and IgA2m(2) with a third form IgA2(n) possibly representing a third allotype (Tsuzukida *et al.*, 1979; Torano *et al.*, 1978; Chintalaruvu *et al.*, 1994). Human IgA1 and IgA2 exhibit differences at only 22 residue positions. However these give rise to several significant structural differences between the isotypic and allotypic variants (Kerr, 1990). The IgA1 isotype has an elongated hinge region between its C_H1 and C_H2 domains and there are five O-linked glycosylation sites within each hinge region (Baezinger & Kornfeld, 1974; Mattu *et al.*, 1998). Each IgA1 heavy chain also has two N-linked glycosylation sites and these are located on its C_H2 domain and on its tailpiece. All of the IgA2 proteins have a deletion of 13 residues in the hinge region and hence lack the O-linked glycosylation sites. The differences in the hinge regions of IgA1 and IgA2 account for their different susceptibilities to bacterial proteinases (Senior *et al.*, 1991), and for differences in binding to a human T-cell receptor which binds IgA1 O-linked sugars (Rudd *et al.*, 1994). However, both of the IgA1 N-linked glycosylation sites are conserved in the IgA2 proteins. All three IgA2 variants have two additional shared N-linked sites in each of their heavy chains; one is located in the C_H1 domain and the other is located in the C_H2 domain. In addition, the IgA2m(2) and IgA2(n) forms have a common fifth N-linked site, which is situated in the C_H1 domain. The IgA proteins contain a number of cysteine residues and the disulphide bridges formed by these cysteines are responsible for maintaining the structure of the IgA monomers and for stabilizing complexes with other protein molecules.

The IgA present in human serum is mainly monomeric (Kerr, 1990). Serum IgA is produced in the bone marrow and, although it consists of both isotypes, the IgA1 isotype accounts for almost 90% of it. Although the concentration of IgA in serum is

relatively low, only constituting between 10 and 15% of the total serum immunoglobulin, the synthesis of IgA in humans is actually greater than for all of the other immunoglobulin classes combined. The reason for this is that IgA is present in mucosal secretions. The mucosal surfaces (gastrointestinal tract, airways, oral cavity and genital mucosa) form the largest exposed area of the body and are consequently the main sites for colonization and invasion by microorganisms.

Secretory IgA (sIgA) differs from serum IgA in several respects. Secretory IgA is not synthesized in the bone marrow but by plasma cells within mucosal lymphoid tissues (Kraehenbuhl & Neutra, 1992). As a result of this localized production, the proportions of the two IgA isotypes in sIgA vary according to the type of secretion (Kerr, 1990). The composition of sIgA also differs significantly from that of serum IgA. Secretory IgA is invariably polymeric, consisting mainly of dimers, and has a J-chain molecule and a secretory component (SC) molecule associated with it.

The J-chain is a small polypeptide (molecular weight of 15,000) (Max & Korsmeyer, 1985), which is also found in association with pentameric IgM. In IgA-producing plasma cells, intracellular dimerization of IgA occurs with a J-chain attached (Zikan *et al.*, 1986). The tailpiece of IgA is essential for dimerization (Atkin *et al.*, 1996) and the cysteine residue that occupies the penultimate position on the tailpiece has been shown to play an important role by the formation of a disulphide bridge with the J-chain (Figure 5.1) (Bastian *et al.*, 1992; Atkin *et al.*, 1996).

After soluble dimeric IgA is released from plasma cells, it can be recognized by the polymeric immunoglobulin receptor (pIgR), which is a cell-surface glycoprotein found at the basolateral membrane of glandular or mucosal epithelial cells. The pIgR has a molecular weight of 100,000 and consists of a five-domain extracellular region, a 23-residue transmembrane segment and a 103-residue cytoplasmic tail (Krajči *et al.*, 1989). Its five extracellular domains all have Ig-fold structures, of which the N-terminal four domains are predicted to adopt the V-set fold structure and the fifth domain is predicted to adopt the C2-set fold structure (Williams & Barclay, 1988). Its primary interaction with dimeric IgA is a high affinity non-covalent interaction (Kuhn &

Kraehenbuhl, 1979) that is mediated by its N-terminal domain (Frutiger *et al.*, 1986). This interaction involves the loops in the N-terminal domain of pIgR that correspond to the complementarity determining regions (CDRs) of antibodies (Coyne *et al.*, 1994) and the first CDR-like loop is of particular importance (Bakos *et al.*, 1991a, 1991b). Transcytosis of the pIgR transports the dimeric IgA from the basolateral surface to the apical surface of epithelial cells (Apodaca *et al.*, 1991; Kraehenbuhl & Neutra, 1992). During the process of transcytosis, a disulphide bridge is formed between pIgR and dimeric IgA. This disulphide bridge involves Cys311 on one of the C_H2 domains in one of the IgA monomers and a free cysteine in the fifth Ig domain of pIgR (Figure 5.1) (Fallgreen-Gebauer *et al.*, 1993). The formation of this disulphide bridge occurs late in the transcytosis pathway and is not absolutely necessary for transcytosis (Chintalacharuvu *et al.*, 1994). Either during transcytosis or at the apical surface, the extracellular region of pIgR is cleaved from its transmembrane and cytoplasmic regions. The five-domain cleaved portion of pIgR is the so-called secretory component and the resulting sIgA is then secreted into the lumen at the apical surface of the epithelial cell.

5.1.2. Functions of IgA

The abundant production of sIgA is generally viewed as providing an immunological barrier that serves to prevent foreign matter, including microorganisms, from adhering to and entering the body (Underdown & Schiff, 1986; Kraehenbuhl & Neutra, 1992; Killian & Russell, 1994). It is suggested that complexes formed between sIgA and antigens become readily entrapped in mucus and are then cleared by peristalsis in the gut or mucociliary transport on respiratory tract surfaces and so preventing contact between antigens/microorganisms and epithelial surfaces. Alternatively, the attachment of microorganisms to epithelial cells may be restricted by sIgA blocking the sites that mediate this attachment. In addition to forming a secreted immunological barrier, sIgA may also exert protective functions within the body (Mazanec *et al.*, 1993). Intracellular neutralization of viruses by sIgA could play an important role in immune defence. Monoclonal IgA antibodies have been shown to neutralize Sendai virus and influenza virus within epithelial cells, seemingly demonstrating an intersection of the sIgA transcytosis pathway and the intracellular pathway for synthesis and assembly of virus (Mazanec *et al.*, 1992). Secretory IgA could also play a role in directly removing

antigens/microorganisms from the body. *In vitro* experiments indicate that IgA-antigen complexes may be secreted by the pIgR transcytosis pathway (Kaetzel *et al.*, 1991, 1994). The importance of IgA in mediating immune protection is demonstrated by the effectiveness of monoclonal IgA and sIgA antibodies against specific bacterial or viral antigens in preventing invasion (Winner *et al.*, 1991; Michetti *et al.*, 1992; Mazanec *et al.*, 1987; Renegar & Small, 1991a, 1991b). The antigen-binding function of sIgA is presumably enhanced by the multivalency provided by its two or more constituent IgA monomers. Indeed, it has recently been shown that polymeric IgA and sIgA are more effective than monomeric IgA at neutralizing viruses (Renegar *et al.*, 1998).

IgA antibodies can also exert antimicrobial effects through active mechanisms, that is those which are coupled to the effector arm of the immune systems. There exists a specific α-chain Fc receptor (FcαR or CD89) (Monteiro *et al.*, 1990; Maliezewski *et al.*, 1990). This receptor has been identified on neutrophils, eosinophils, monocytes and macrophages and binds serum and secretory forms of IgA (Weisbart *et al.*, 1988; Shen *et al.*, 1989; Monteiro *et al.*, 1993; Morton *et al.*, 1996; Kerr & Woof, 1998). The FcαR sequence reveals that it contains two extracellular Ig-fold domains, a transmembrane peptide and a short cytoplasmic tail (Maliezewski *et al.*, 1990). The two Ig-fold domains have homology with IgG receptors (FcγRI, II and III) and the high affinity IgE receptor (FcεRI). An IgA1 mutant lacking the C_H2 domain glycosylation site at Asn263 is unable to bind FcαR (Carayannopoulos *et al.*, 1994 and 1996). Binding between IgA1 and FcαR is also lost upon mutation of residues in loops at the interface between the C_H2 and C_H3 domains, and it is therefore proposed that the interaction site occurs within this region (Carayannopoulos *et al.*, 1996). Despite their sequence homology, this putative interaction site is not comparable to those for FcγRI, II and III and FcεRI which all contact their respect immunoglobulin ligands at a hinge-proximal position (Canfield & Morrison, 1991, Chappel *et al.*, 1991; Sarmay *et al.*, 1992; Nissim & Eshar, 1992). Instead in IgG, the interface between the C_H2 and C_H3 domains is the site of a number of other interactions. X-ray crystal structures reveal that *Staphylococcus aureus* protein A (Deisenhofer, 1981), streptococcal protein G (Sauer-Eriksson *et al.*, 1995), rheumatoid factor (Corper *et al.*, 1997) and rat neonatal Fc receptor (Burmeister *et al.*, 1994b) all bind near to the C_H2-C_H3 interface. Perhaps significantly, reorientations of

the C_H2 domains relative to the C_H3 domains are observed in IgG Fc crystal structures. The large helical loop between strands A and B of the C_H2 domains, which is located at the C_H2-C_H3 interface, is proposed to be critical for these reorientations, forming a "pivot" about which the C_H2 domain rotates (Harris *et al.*, 1998b).

There is also evidence that IgA functions to regulate complement-activation, however this remains a debatable issue. While it is generally accepted that IgA does not activate the classical complement pathway, several studies have shown activation of complement by IgA1, IgA2 and sIgA via the alternative pathway (Hiemstra *et al.*, 1987, 1988; Lucisano-Valim & Lachmann, 1991; Bogers *et al.*, 1991). In studies demonstrating complement-activation by the IgA1 isotype, the IgA1 was either aggregated chemically or by high temperatures, or it was bound onto plastic and so might not reflect the *in vivo* properties of IgA1. Recently it was shown that IgA1 binds to C3 but is unable to activate complement (Chuang & Morrison, 1997). Other studies have also questioned whether IgA can activate complement and instead it has been proposed that it may have the opposite function, namely the inhibition of complement activation by other immunoglobulin classes (Griffiss & Goroff, 1983; Russell *et al.*, 1989; Jarvis & Griffiss, 1991; Nikolova *et al.*, 1994).

5.1.3. Issues of IgA structure

To date, structural studies on IgA have utilised protein crystallography and electron microscopy. The crystal structure of a Fab fragment from murine IgA J539 was reported, and this confirmed the β-sheet structure that is expected from numerous human and murine IgG crystal structures (Suh *et al.*, 1986). No crystal structure is presently known for an IgA Fc fragment, although these are known for human, rabbit and mouse IgG Fc fragments (Deisenhofer, 1981; Sutton & Phillips, 1983; Harris *et al.*, 1998a). Structures of intact IgA have been reported for human and rabbit IgA by electron microscopy, and this showed that the two Fab fragments are attached to the Fc fragment to give a Y-shaped structure (Svehag & Bloth, 1970). Recently, crystal structures have been determined for intact murine IgG1 and IgG2a (Harris *et al.*, 1992, 1997, 1998a, 1998b). These showed two pseudo-dyad axes, one relating the two Fab fragments and the other the two halves of the Fc fragment. The arrangements of the two Fab fragments

relative to the Fc fragment were notably asymmetric, where murine IgG1 was described as a distorted Y-shape and murine IgG2a was a distorted T-shape. It was therefore deduced that the hinge regions in murine IgG1 and IgG2a had an inherent flexibility which allowed the Fab fragments to adopt numerous conformations relative to the Fc fragment in IgG. The other Ig classes have very different hinge regions. That of IgA is similar in size to that of IgG, but is heavily glycosylated, while those in immunoglobulins E and M are replaced by an additional domain.

The determination of the arrangement of the Fab and Fc fragments in human IgA1 is essential for evaluating its function. In particular, the structural significance of the glycosylated hinge is IgA1 compared to the IgG hinge is presently unclear. The application of X-ray crystallography or multidimensional NMR to determine a high resolution structure will be restricted by the structural flexibility, high glycosylation levels and large size of IgA1. Small angle X-ray and neutron solution scattering offers an alternative means of studying protein structures. Even though the structural resolution is about 3 nm, and unique structures cannot be determined by scattering, the use of known structural models as tight constraints on the interpretation of scattering data considerably restricts the structures that are allowed. By this method, the experimental precision of a structure determination can be of the order of 0.2-1 nm (Perkins *et al.*, 1998a, 1998b). Here, in application to IgA1, crystal structures for murine IgA and human IgG1 Fab fragments and the human IgG1 Fc fragment were used to construct and test homology models for the individual C_H1, C_H2 and C_H3 domains of IgA1 and the full IgA1 Fab and Fc fragments. Solution scattering data for IgA1 and a mutant lacking the tailpiece, PTerm455, were used with the homology models and the development of a new conformational search procedure for scattering curve fits that involves the molecular dynamics modelling of the IgA1 hinge. This resulted in the creation of tens of thousands of IgA1 models that could be tested against the scattering data. This procedure revealed that the optimal solution structure of IgA1 is significantly different from several structures determined for IgG, including one recently determined by neutron scattering (Harris *et al.*, 1998b; Mayans *et al.*, 1995). The relevance of the solution structure of human IgA1 to the IgA2 isotype, the dimeric IgA and secretory IgA structures, and IgA function is discussed.

5.2. Materials and methods

5.2.1. Preparation of IgA1 and PTerm455 for solution scattering

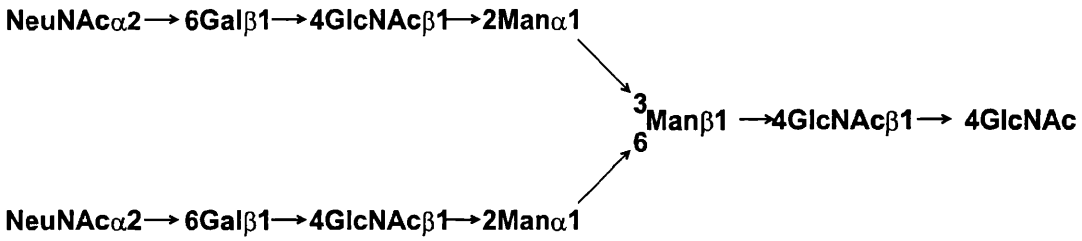
Human IgA1 was purified from serum using jacalin-affinity chromotography by Prof. M. A. Kerr (Dundee) (Loomes *et al.*, 1991; Kerr *et al.*, 1997). Briefly, a crude globulin preparation was precipitated from serum using 50% saturated ammonium sulphate solution. The redissolved pellet was gel-filtrated on Sepharose 6B and IgA-containing fractions were pooled. IgA was separated from IgG using Q-Sepharose anion exchange chromatography. IgA1 was purified using a jacalin-agarose column, and its purity was confirmed using SDS-PAGE. PTerm455 is a recombinant form of IgA1 that contains a translation stop codon immediately prior to the tailpiece coding region (Atkin *et al.*, 1996). PTerm455 binds the antigen 3-nitro-4-hydroxy-5-iodophenylacetate (NIP). Its purification by J. M. Woof (Dundee) from the culture supernatant of Chinese hamster ovary K1 transfectants using NIP-Sepharose affinity chromatography was performed according to Morton *et al.* (1993).

Immediately prior to X-ray and neutron scattering data collection, IgA1 and PTerm455 samples were gel-filtrated on a Superdex 200 column using fast performance liquid chromatography to remove nonspecific aggregates. The buffer used for data collection was Dulbecco's phosphate buffer (12.5 mM sodium phosphate, 140 mM NaCl at pH 7.4). X-ray experiments utilised H_2O buffer while neutron experiments involved dialysis at 6°C into 100% ^2H_2O buffer for at least 36 hours with four buffer changes.

5.2.2. Composition of IgA1 and PTerm455

Amino acid and carbohydrate sequences were required for data analyses and the modelling of IgA1. The α-chain sequence of IgA1 for the C_H1, C_H2 and C_H3 domains was taken from SWISSPROT (database code: P01876). The IgA1 α-chain sequence was numbered according to the commonly adopted scheme used for IgA1 Bur (Putnam *et al.*, 1979). The V_H, V_L and C_L sequences were taken from the crystal structure of human TR1.9 IgG1 Fab (Brookhaven database code: 1vge; Chacko *et al.*, 1996), and included the human κ light chain. For PTerm455, the C-terminal 18 residues of the IgA1 α-chain sequence were deleted. The IgA1 carbohydrate composition was based on data for the O- and N-linked oligosaccharides in serum IgA1 (Figure 5.2; Field *et al.*,

(a)



(b)

Figure 5.2. The N-linked and O-linked carbohydrate structures used in the modelling of PTerm455 and IgA1.

(*a*) The biantennary complex-type N-linked oligosaccharide structure used in the modelling of PTerm455 and IgA1 was built to satisfy carbohydrate sequence analyses (Field *et al.*, 1994; Mattu *et al.*, 1998) using the carbohdyrate coordinates in the crystal structure of the IgG1 Fc (Brookhaven code 1fc1; Deisenhofer, 1981) and an isolated glycopeptide fragment of human lactotransferrin (Brookhaven code 1lge; Bourne *et al.*, 1994).

(*b*) The O-linked oligosaccharide structure used in the modelling was built to satisfy PTerm455 IgA1 and IgA1 carbohydrate analyses (Field *et al.*, 1989; Field *et al.*, 1994; Mattu *et al.*, 1998) using carbohydrate coordinates from the crystal structure of cholera-toxin B-pentamer bound to receptor G(M1) pentasaccharide (Brookhaven code 1chb; Merritt *et al.*, 1994).

1989, 1994; Mattu *et al.*, 1998). The N-linked carbohydrates at Asn263 on the C_H2 domain and Asn459 on the tailpiece were each modelled by a standard biantennary complex-type oligosaccharide with a $Man_3GlcNAc_2$ core and two NeuNAc.Gal.GlcNAc antennae (Figure 5.2a). The five O-linked carbohydrates on each IgA1 α-chain were modelled using a NeuNAc.Gal.GalNAc oligosaccharide (Figure 5.2b). The modelling searches used O-linked carbohydrate sites at Ser224, Ser230, Ser232, Ser238 and Ser240 to correspond to human myeloma IgA1 (Baezinger & Kornfeld, 1974). The final models used O-linked sites at Thr225, Thr228, Ser230, Ser232 and Thr236 to correspond to human serum IgA1 (Mattu *et al.*, 1998). The compositions of IgA1 and PTerm455 were used to calculate their molecular weights, unhydrated and hydrated glycoprotein volumes (based on a hydration of 0.3g H_2O/g of glycoprotein and an electrostricted volume of 0.0245 nm^3 per bound water molecule) and absorption coefficients at 280 nm using the corrected Wetlaufer procedure (Perkins, 1986; Table 5.1).

5.2.3. X-ray data collection

X-ray scattering data were obtained for serum IgA1 in one session using the camera at Station 8.2 and in two further sessions using the camera at Station 2.1, both at the SRS Daresbury Laboratory, Warrington, U.K (Towns-Andrews *et al.*, 1989; Worgan *et al.*, 1990). X-ray data for PTerm455 were obtained in two independent sessions using the camera at Station 2.1. Experiments were performed with beam currents in a range of 144 to 286 mA and a ring energy of 2.0 GeV. Samples were measured for ten minutes in ten equal time frames at concentrations between 1.0 and 2.9 mg/ml. The Q range was calibrated using fresh, slightly stretched rat tail collagen, based on a diffraction spacing of 67.0 nm. Samples were held in Perspex cells of sample volume 20 µl, contained within mica windows of thickness between 10 and 15 µm, and cooled at 15°C. Buffers and samples were measured in alternation for equal times to minimise background subtraction errors. Data were accepted only if the subsequent Guinier plots were linear and reproducible in repeated measurements. Each of the ten time frames was individually checked using Guinier analysis to check for the absence of time-dependent radiation-damage effects. Data reduction was performed using the standard Daresbury software package OTOKO (Boulin *et al.*, 1986). Curves were

Table 5.1. Composition of human IgA1 and its fragments

IgA1 fragment	Amino acids	O-linked CHO	N-linked CHO	Molecular weight	Dry volume (nm^3)	Neutron spheres N (dry)	Hydrated volume (nm^3)	X-ray spheres N (hydrated)	Absorption coefficient (1%, 1 cm, 280 nm)
Fab (4 domains)	437	0	0	47,000	59.9	=	79.2	=	=
Hinge	23	5	0	5,500	6.4	=	8.7	=	=
Fc (4 domains)	422	0	2	50,000	63.4	=	83.9	=	=
Tailpiece	18	0	2	4,000	4.8	=	6.4	=	=
Method 1 (12 domains, no hinge, no tailpiece)	1296	0	2	144,000	183.3	1224	242.2	1618	=
Method 2, PTerm455 (12 domains, no tailpiece)	1342	10	2	155,000	196.2	1310	259.6	1734	13.2
Method 3, serum IgA1 (12 domains)	1378	10	4	164,000	205.7	1374	272.5	1820	12.7

normalised using an ion chamber monitor positioned after the sample for at least 5.9 hours using a uniform ^{55}Fe radioactive source. Reduced curves were calculated by subtraction of the buffer runs from those of the samples.

5.2.4. Neutron data collection

Neutron scattering data for serum IgA1 were obtained in three different sessions and those for PTerm455 were obtained in a single session on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot U.K. (Heenan & King, 1993). The pulsed neutron beam was derived from proton beam currents of 167 to 200 μA. Monochromatisation was achieved using time-of-flight techniques. A ^{3}He ORDELA wire detector was employed to record intensities at a fixed sample-to-detector distance of 4.3 m. The samples and $^{2}H_2O$ buffers were measured in 2 mm thick rectangular Hellma cells positioned in a thermostated rack at 15°C. Data acquisitions were for fixed totals of 40-300 × 10^6 monitor counts in runs lasting 0.5 to 3.5 h each for protein concentrations of 0.8 to 12.2 mg/ml. Spectral intensities were normalised relative to the scattering from a partially deuterated polystyrene standard (Wignall & Bates, 1987). Transmissions were measured for all samples and backgrounds. Reduction of the raw data collected in 100 time frames of 64 × 64 cells utilised the standard ISIS software package COLETTE (Heenan *et al.*, 1989). Scattered intensities were binned into individual diffraction patterns based on the wavelength range 0.2 to 1.0 nm, and were corrected for linear wavelength-dependence of the transmission measurements. The patterns were merged to give the full curve $I(Q)$ in a maximal Q range of 0.07 to 2.38 nm^{-1}. The Q range was based on 0.04% logarithmic increments, which was optimal both for Guinier analyses at low Q, and for better signal-noise ratios at large Q.

5.2.5. Analysis of reduced X-ray and neutron data

In a given solute-solvent contrast, the radius of gyration R_G is a measure of structural elongation if the internal inhomogeneity of scattering densities has no effect. Guinier analyses at low Q give the R_G, and the forward scattering at zero angle $I(0)$ (Glatter & Kratky, 1982):

$$\ln I(Q) = \ln I(0) - R_G^2 Q^2/3.$$

This expression is valid in a $Q.R_G$ range up to 1.5. The relative $I(0)/c$ values (c = sample concentration) for samples measured in the same buffer during a data session gives the relative molecular weight of the proteins when referenced against a suitable standard (Kratky, 1963; Wignall & Bates, 1987). A regression analysis of 13 neutron $I(0)/c$ values measured on LOQ for proteins in the molecular weight range 27,000 to 166,000 showed a linear relationship with molecular weight that can be used to derive these (Chapter 2; Figure 2.10). If the structure is elongated, the mean radius of gyration of cross-sectional structure R_{XS} and the mean cross-sectional intensity at zero angle $[I(Q)Q]_{Q\to 0}$ (Hjelm, 1985) can be obtained from:

$$\ln [I(Q)Q] = [I(Q)Q]_{Q\to 0} - R_{XS}^2 Q^2/2.$$

For immunoglobulins, the cross-sectional plot exhibits two regions, a steeper innermost one and a flatter outermost one (Pilz *et al.*, 1973), and the two analyses are identified by $R_{XS\text{-}1}$ and $R_{XS\text{-}2}$ respectively. The R_G and $R_{XS\text{-}1}$ analyses lead to the triaxial dimensions of the macromolecule. If the structure can be represented by an elliptical cylinder, $L = \sqrt{12(R_G^2 - R_{XS\text{-}1}^2)}$, where L is its length (Glatter & Kratky, 1982). Data analyses employed an interactive graphics program SCTPL5 (A.S. Nealis, A.J. Beavil and S.J. Perkins, unpublished software) on a Silicon Graphics 4D35S Workstation.

Indirect transformation of the scattering data $I(Q)$ in reciprocal space into real space to give $P(r)$ was carried out using the GNOM program (Semenyuk & Svergun, 1991)

$$P(r) = \frac{1}{2\pi^2} \int_0^\infty I(Q)\, Qr\, \sin(Qr)\, dQ$$

$P(r)$ corresponds to the distribution of distances r between volume elements. This offers an alternative calculation of R_G and $I(0)$ which is now based on the full scattering curve, and also gives the maximum dimension L. This was used with the X-ray IgA1 $I(Q)$

curve in the Q range between 0.10 to 2.20 nm^{-1} and the neutron PTerm455 $I(Q)$ curve in the Q range between 0.13 to 2.20 nm^{-1}. A range of D_{max} values was tested, and the final choice of D_{max} was based on three criteria: (i) $P(r)$ should exhibit positive values; (ii) the R_G from GNOM should agree with the R_G from Guinier analyses; (iii) the $P(r)$ curve should be stable as D_{max} was increased beyond the estimated macromolecular length.

5.2.6. Homology modelling of human IgA1 α-chain domains and Fab and Fc fragments

For predictions and modelling, a multiple sequence alignment of 27, 22 and 21 α-chain sequences for the C_H1, C_H2 and C_H3 domains respectively was generated using sequences from the SWISSPROT database (codes P01876, P01877 and P01877 for human IgA1, IgA2m(1) and IgA2m(2)) and the World Wide Web version of the Kabat database (Johnson *et al.*, 1996; http://immuno.bme.nwu.edu) for other mammalian α-chain domains (codes 013507-013511, 013536, 013577, 013578, 013580, 013582, 013609-013621). That for human IgA2(n) was taken from Chintalacharuvu *et al.* (1994). The sequences were aligned using CLUSTALW (Thompson *et al.*, 1994). From the three alignments, averaged secondary structure predictions were performed using the Chou-Fasman, GORI, GORIII, PHD and SAPIENS methods, and likewise the mean hydropathy and solvent accessibilities were predicted using the Eisenberg, PHD and SAPIENS methods (summarised in Edwards & Perkins, 1996).

Homology models for the three α-chain domains of human IgA1 were created using INSIGHT II 95.0 molecular graphics software with the BIOPOLYMER, DISCOVER, DISCOVER3, HOMOLOGY and DELPHI modules (Biosym/MSI, San Diego, CA, USA) on Silicon Graphics INDY Workstations. Each model was built using a reference crystal structure using the rigid-body fragment assembly method in HOMOLOGY. The IgA1 C_H1 domain was modelled on the C_H1 domain from the crystal structure of murine IgA J539 Fab fragment (Suh *et al.*, 1986; PDB code 2fbj). The IgA1 C_H2 and C_H3 domains were modelled on the C_H2 and C_H3 domains from the crystal structure of human IgG1 Fc fragment (Deisenhofer, 1981; PDB code 1fc1-A). In the modelling, the secondary structure of the crystal structures were identified using

DSSP (Kabsch & Sander, 1983). The structurally conserved regions of each domain was defined as β-strands that were conserved in both the IgA secondary structure prediction and the crystal structure. Loop structures in each domain were modelled either directly as a designated loop using the reference crystal structure, or indirectly as a searched loop using a database of Brookhaven loop fragments if sequence insertions or deletions occurred (Hobohm & Sander, 1994; Hobohm *et al.*, 1992). After this, all the sidechains were mutated to those in the human IgA1 sequence. Energy refinements using DISCOVER were performed at the loop splice junctions, the sidechain and mainchain atoms of the loop residues, and the sidechain atoms of the structurally conserved regions. The stereochemical validity of the homology models was confirmed using PROCHECK (Laskowski *et al.*, 1993).

Homology models for each of the full IgA1 Fab and Fc fragments in IgA1 (Figure 5.1) were now constructed. The Fab fragment was built using the crystal structure of the TR1.9 Fab fragment (Chacko *et al.*, 1996) for the V_L, C_L and V_H domains. The homology model of the IgA1 C_H1 domain was superimposed onto that in TR1.9 using the α-carbon atoms of the conserved β-strands. The Fc fragment was built using the crystal structure of the human IgG1 Fc fragment (above) by the superimposition of the IgA1 C_H2 and C_H3 models onto the C_H2 and C_H3 domains in the IgG1 Fc structure using the α-carbon atoms of the conserved β-strands. A second Fc fragment model was created which included four disulphide bridges in the Fc fragment next to the hinge. For this, Cys241-Cys242-His243 were added to each N-terminus of the first Fc model, then the two C_H2 domains were rotated equally towards each other to bring the two Cys299 residues into proximity with each other. The two Cys241 residues were subject to a distance-restrained energy minimisation to bring these together, then further minimisation was performed to create the four Cys241-Cys241, Cys299-Cys299 and 2 × Cys242-Cys301 bridges. In order to assess buried sidechains at the domain interfaces in the models and crystal structures, the sidechain solvent accessibilities were calculated for the six isolated V_L, C_L, V_H, C_H1, C_H2 and C_H3 domains, and again for the intact Fab and Fc fragments, and for adjacent pairs of domains within the Fab and Fc fragments. This employed a 0.14 nm radius probe using the Lee & Richards (1971) method incorporated within the COMPARER program (Šali

& Blundell, 1990).

5.2.7. Automated modelling of PTerm455 and IgA1

Three different automated search strategies were used to model the IgA1 structure. All used a vertical twofold axis of symmetry centred on the Fc fragment (Figure 5.1) to define the position of both Fab fragments relative to the Fc fragment. Each IgA1 model was structurally defined by the distance between the centres of mass of the two Fab fragments and that between the Fab and Fc fragment. Equivalent planar or pyramidal structures for IgA1 were not therefore distinguished. The hinge length of each model was defined as the distance between the α-carbon atoms of Cys220 and Pro244. The preparation of each set of models and their scattering curve fits (below) took about 2-3 weeks of computer processor time each on an INDY Workstation. The three approaches are now summarised:

Method 1. The long axes of the homology models for the Fab and Fc fragments of IgA1 were positioned relative to each other in the same orientation used by Mayans *et al.* (1995) to model bovine IgG1 and IgG2 (Figure 5.3). Holding this orientation fixed, the C-terminal α-carbon atom (Cys220) of the C_H1 domain was superimposed on the N-terminal α-carbon atom (Pro244) of the corresponding C_H2 domain (Figure 5.3). An N-linked oligosaccharide model was added to Asn263 on each C_H2 domain in an extended conformation. Within INSIGHT II, the moment of inertia of each Fab fragment was used to define its X-, Y- and Z-axes, where the X-axis corresponds to the longest dimension of the Fab fragment, the Y-axis is approximately parallel to the twofold symmetry axis of the Fc fragment in Figure 5.1, and the α-carbon atom of Cys220 is the origin (Figure 5.3). Using an INSIGHT script file, the two Fab fragments were translated in 0.5 nm steps along their X-axes between 0 nm and 10 nm, along their Y-axes between -10 nm and 10 nm, and along their Z-axes between -10 nm and 10 nm (Figure 5.3). These translations produced $21 \times 41 \times 41 = 35{,}301$ models that systematically explored the three-dimensional space about the Fc fragment, in analogy to the two-dimensional search used for bovine IgG1 and IgG2 (Mayans *et al.*, 1995).

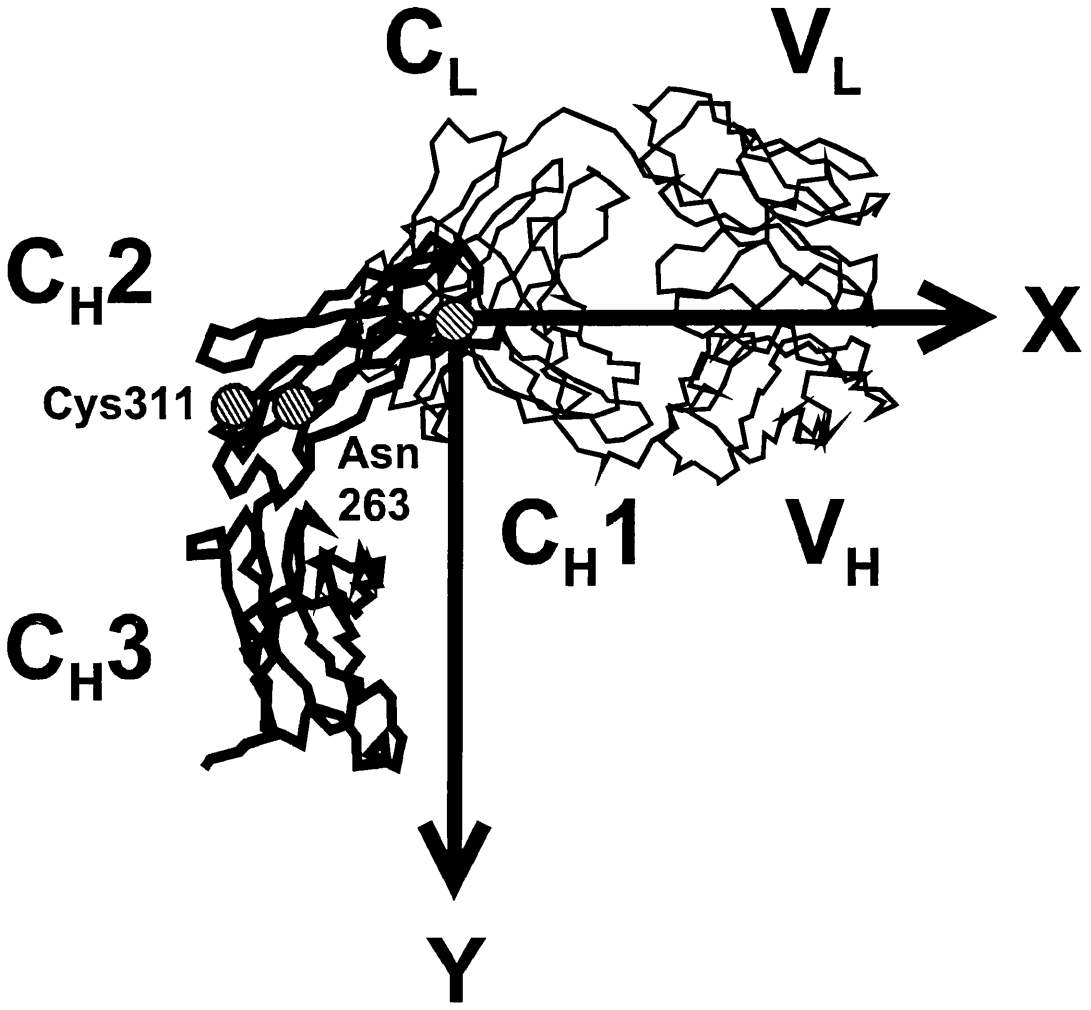


Figure 5.3. Starting model used to generate PTerm455 and IgA1 models by translations of the Fab fragments relative to a fixed Fc fragment (Method 1). Half the starting model is shown in the plane of the X- and Y-axes, with the Z-axis perpendicular to the plane of the Figure. This is comprised of a Fab fragment (V_H, C_H1, V_L, C_L; thin lines) and half an Fc fragment (C_H2, C_H3; thick lines). The C-terminal α-carbon (Cys220) of the C_H1 domain in the Fab fragment and the N-terminal α-carbon (Pro244) of the C_H2 domain in the Fc fragment are set to be coincident at the origin of the X- and Y-axes (shaded circle). The position of Asn263 and Cys311 on the outer surface of the Fc fragment is shown by further shaded circles.

Method 2. To form PTerm455, the homology models for the Fab and Fc fragments of IgA1 were combined with 12,000 structures for a 25-residue hinge peptide Cys220-Pro244. This peptide was set up as a β-strand of maximum length 8.75 nm using BIOPOLYMER. A molecular dynamics simulation using DISCOVER3 created random hinge structures from this starting model. In order to generate a wide range of structures, the simulation was performed at a temperature of 773 K. First, the peptide structure was subjected to energy minimisation for 300 iterations. After a temperature equilibration step of 5,000 fs, the simulation was run for 500,000 fs. The hinge structure was saved every 100 fs to produce 5,000 models. In further simulations, the hinge length was constrained to be greater than either 2 nm, 3 nm, 4 nm, 5 nm, 6 nm, 7 nm or 8 nm, and the simulations were run at 773 K for 100,000 fs to produce 7 × 1,000 = 7,000 further models. PTerm455 models were produced by a fragment assembly method. Holding the Fc fragment fixed, two copies of each hinge model were added to it by superimposing the four mainchain atoms of Pro244 in the hinge onto those of Pro244 in the Fc fragment. Likewise, the Fab fragments were now added to the hinge peptides by superimposing the mainchain atoms of Cys220 in the hinge and the Fab fragment. This meant that the Fab fragments displayed all orientations relative to the Fc fragment, unlike that of Method 1 above. The duplicate Cys220 and Pro244 residues were deleted from the PTerm455 model. An N-linked oligosaccharide structure was added to Asn263 as above. An O-linked oligosaccharide model was added in an extended conformation to each of the ten glycosylation sites in the hinge.

Method 3. To form serum IgA1, the best PTerm455 curve-fit model from Method 2 was used with 16,000 tailpiece structures created using molecular dynamics. A 19-residue tailpiece peptide corresponding to Lys454-Tyr472 was built in a β-strand conformation of length 6.65 nm using BIOPOLYMER. A molecular dynamics simulation performed at 773 K for 500,000 fs generated 5,000 tailpiece structures. Further structures were created by setting the distance between the α-carbon atoms of Lys454 and Tyr472 to be longer than 1 nm, 2 nm, 3 nm, 4 nm, 5 nm or 6 nm, and these simulations were run at 773 K for 100,000 fs to produce 6 × 1,000 = 6,000 models. Since the best tailpiece models were approximately 4 nm in length, a final set of 5,000 models was generated whose lengths were constrained between 3.5 nm and 5.5 nm.

Two copies of each tailpiece model were added to the PTerm455 model to form IgA1 by superimposing the mainchain atoms of Lys454 in both the tailpiece and the Fc fragment, followed by deletion of the duplicated residue. An N-linked oligosaccharide model was added in an extended conformation to Asn459 on each tailpiece peptide.

5.2.8. Debye scattering curve calculations from sphere models of PTerm455 and IgA1

Each PTerm455 and IgA1 model was used to calculate X-ray and neutron scattering curves for comparison with the experimental curves. Each set of atomic coordinates for a model was placed within a three dimensional grid of cubes. A sphere of equal volume to the cube was placed at the centre of each cube if a user-specified cutoff for the minimum number of atoms contained within a cube was satisfied. For both PTerm455 and IgA1, a cube side length of 0.531 nm in combination with a cutoff of 4 atoms consistently produced sphere models within 2% of the total dry volume of each protein calculated from its composition (Table 5.1). The optimal total of dry spheres for IgA1 and PTerm455 were 1374 and 1310 respectively. Since the hydration shell is detected around glycoproteins by X-ray scattering, the sphere models were accordingly adapted by adding spheres to the surface of the dry models using the HYPRO procedure of Ashton *et al.* (1997). HYPRO is better suited to protein structures with large void spaces such as that within the Fc fragment than our previous hydration method which performed a simple expansion of the sphere model (Smith *et al.*, 1990). The optimal totals of hydrated spheres for IgA1 and PTerm455 are 1820 and 1734 spheres respectively (Table 5.1).

The X-ray and neutron scattering curve $I(Q)$ was calculated assuming a uniform scattering density for the spheres using the Debye equation as adapted to spheres (Perkins & Weiss, 1983):

$$\frac{I(Q)}{I(0)} = g(Q)\,(n^{-1} + 2n^{-2} \sum_{j=1}^{m} A_j \frac{\sin Qr_j}{Qr_j})$$

$$g(Q) = (3(\sin QR - QR \cos QR))^2 / Q^6R^6$$

where $g(Q)$ is the squared form factor for the sphere of radius r, n is the number of spheres filling the body, A_j is the number of distances r_j for that value of j, r_j is the distance between the spheres, and m is the number of different distances r_j. For sphere models based on cubes with a side length of 0.531 nm, the minimum value of r_j is much less than the nominal resolution of $2\pi / Q_{max}$ of the scattering curves (3.1 nm for Q_{max} = 2.0 nm^{-1}). The method has been tested with known crystal coordinates (Smith *et al.*, 1990; Perkins *et al.*, 1993; Ashton *et al.*, 1997). X-ray curves were calculated from the hydrated sphere models without corrections for wavelength spread or beam divergence as these are considered to be negligible for synchrotron X-ray data. Neutron curves were calculated from the dry sphere models and corrected using a 10% wavelength spread for a nominal λ of 0.6 nm and a beam divergence of 0.016 radians (Ashton *et al.*, 1997). The number of spheres N in the dry and hydrated models after grid transformation was used to assess steric overlap between the Fab and Fc fragments, where models showing less than 95% of the optimal total were discarded. The modelled scattering curves were assessed by calculation of the R_G, R_{XS-1} and R_{XS-2} values in the same Q ranges used in the experimental Guinier fits, where models giving values that exceeded specified ranges (5.6 nm < R_G < 6.4 nm) were discarded. Models that passed both filters were then ranked using a goodness-of-fit R-factor defined by analogy with protein crystallography and based on the experimental curves in the Q range extending to 2.0 nm^{-1} (denoted as $R_{2.0}$; Smith *et al.*, 1990; Beavil *et al.*, 1995).

5.3. Results and discussion

5.3.1. X-ray and neutron Guinier analyses of PTerm455 and IgA1

The three-fragment arrangement of IgA1 in solution was characterised using synchrotron X-ray scattering data $I(Q)$ for recombinant PTerm455 and serum IgA1 at concentrations between 1.0 and 2.9 mg/ml. The X-ray Guinier regions of PTerm455 and IgA1 were stable in time-frame analyses, and showed that both were resistant to X-ray-induced aggregation, which is contrary to the aggregation observed for immunoglobulins M and E (Perkins *et al.*, 1991; Beavil *et al.*, 1995). At the lowest Q

Figure 5.4. (Overleaf) Guinier R_G and R_{XS} plots for human PTerm455 and IgA1. Those for PTerm455 were measured at concentrations of 1.2 mg/ml (X-rays) and 2.8 mg/ml (neutrons). Those for IgA1 were measured at concentrations of 2.2 mg/ml (X-rays) and 2.0 mg/ml (neutrons). In each panel, the lower plot is PTerm455 and the upper plot is IgA1. The filled circles between the QR_G and QR_{XS} ranges (arrowed) show the data points used to determine R_G, R_{XS-1} and R_{XS-2} values. The Q range used for the R_G data was 0.13 to 0.28 nm^{-1}, and those for the R_{XS-1} and R_{XS-2} data were 0.28 to 0.51 nm^{-1} and 0.56 to 1.04 nm^{-1} respectively.

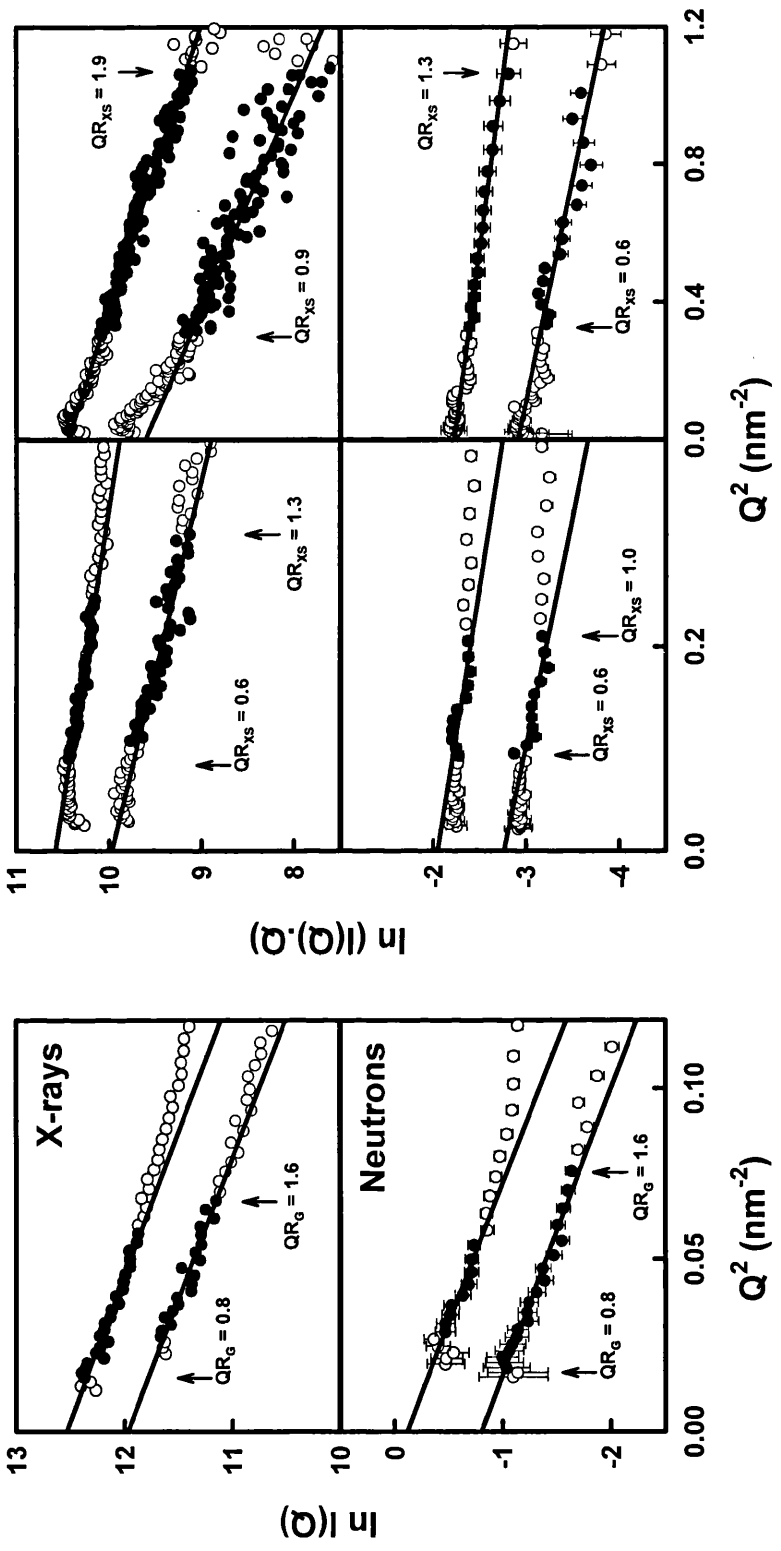


Figure 5.4. Guinier R_G and R_{XS} plots for human PTerm455 and IgA1 (legend on page 200).

Table 5.2. Scattering analyses of human IgA1 and related IgG structures

Protein	Data	Guinier analyses				P(r) analyses			
		R_G (nm)	R_{XS-1} (nm)	R_{XS-2} (nm)	L (nm)	R_G (nm)	L (nm)	M1 (nm)	M2 (nm)
PTerm455	X-ray	6.16 ± 0.12 (4)	2.56 ± 0.12 (4)	1.62 ± 0.18 (4)	19.4	5.95	20.0	4.0	8.7
	Neutron	5.84 ± 0.18 (2)	2.03 ± 0.14 (2)	1.22 ± 0.05 (2)	19.0	=	=	=	=
Serum IgA1	X-ray	6.20 ± 0.13 (5)	2.20 ± 0.26 (7)	1.56 ± 0.16 (7)	20.1	6.12	21.0	3.7	9.1
	Neutron	6.11 ± 0.18 (4)	2.17 ± 0.23 (4)	1.18 ± 0.12 (4)	19.8	=	=	=	=
Bovine IgG1 [1]	Neutron	5.64 ± 0.28 (4)	2.38 ± 0.09	1.02 ± 0.07	17.7 ± 1.2 17.8 ± 1.8	=	15.6 ± 2.1	5.4-5.7	=
Bovine IgG2 [1]	Neutron	5.71 ± 0.51 (3)	2.41 ± 0.12	0.98 ± 0.06	17.9 ± 2.1 18.2 ± 1.6	=	=	=	=
Porcine IgG [1]	Neutron	5.28 ± 0.45 (3)	2.27 ± 0.06	1.22 ± 0.06	16.5 ± 1.8 17.4 ± 1.1	=	=	=	=
Rabbit IgG [1]	Neutron	5.75	2.16	0.98	=	=	=	=	=

[1] Taken from Mayans *et al.* (1995).

values, Guinier analyses resulted in linear plots in three distinct regions of the $I(Q)$ curves from which the R_G, R_{XS-1} and R_{XS-2} values were obtained within satisfactory $Q.R_G$ and $Q.R_{XS}$ limits between 0.8-1.6 (Figure 5.4; Table 5.2). PTerm455 exhibited an X-ray R_G value of 6.16 ± 0.12 nm and IgA1 had one of 6.20 ± 0.13 nm. It is significant that these values are similar. Anisotropy ratios R_G/R_O (where R_O is the R_G value of the sphere with the same volume as the hydrated glycoprotein; Table 5.1) were calculated from the X-ray R_G values of PTerm455 and IgA1. Thus R_G/R_O was 2.01 for PTerm455 and 1.99 for IgA1. Comparison of these anisotropy ratios to that of 1.28 for typical globular proteins (Perkins, 1988) showed that both proteins have extended structures in solution. Since the only major difference between the two proteins is the absence and presence of the N-glycosylated tailpiece (Figure 5.1), the tailpiece would have to be folded up against the C_H3 domain of IgA1 in order not to increase the R_G value and the anisotropy ratio of IgA1 relative to PTerm455.

The X-ray Guinier analyses lead to length determinations for PTerm455 and IgA1. Their R_{XS-1} values were 2.56 ± 0.12 nm and 2.20 ± 0.26 nm respectively. Their R_{XS-2} values were 1.62 ± 0.18 nm and 1.56 ± 0.16 nm respectively. The consistency of these pairs of values again showed no major difference between PTerm455 and IgA1. The combination of the R_G and R_{XS-1} values resulted in the overall length L of the two proteins (Methods). These were similar at 19.4 nm for PTerm455 and 20.1 nm for IgA1.

PTerm455 and IgA1 contain 7% and 9% carbohydrate (w/w) which has a higher scattering density than that of the protein (Perkins, 1986). Neutron scattering data was obtained on PTerm455 and IgA1 as controls for possible X-ray-induced aggregation, correct molecular weights and possible contrast-dependence effects, and also monitors the hydration shell which is visible by X-ray scattering but not by neutron scattering. Contrast effects were monitored by neutron data collection using 100% ^2H_2O buffers, which corresponds to a high negative protein-solvent scattering contrast, and is the opposite of the high positive protein-solvent contrast observed in X-ray scattering. Neutron data were collected at concentrations between 0.8 and 12.2 mg/ml in 100% ^2H_2O buffers. PTerm455 gave a R_G value of 5.84 ± 0.18 nm, and IgA1 had one of 6.11 ± 0.18 nm. The good agreement with the X-ray values (Table 5.2) showed that IgA1

was stable during X-ray irradiation. The small decrease in the neutron R_G values compared to the X-ray values is attributable to the invisible hydration shell (Perkins, 1986; Svergun *et al.*, 1998), and also showed that no significant contrast effects were present. The neutron Guinier $I(0)/c$ values were 0.156 ± 0.004 for PTerm455 and 0.172 ± 0.009 for IgA1. A regression analysis showed that PTerm455 had a molecular weight of 140,000 ± 15,000 and IgA1 had one of 150,000 ± 15,000 (Methods; Chapter 2; Figure 2.10). These agree with values of 155,000 and 164,000 respectively calculated from their compositions (Table 5.1), and confirmed that both proteins were monomeric in solution as expected.

The neutron length determinations L of PTerm455 and IgA1 supported the X-ray data. The neutron $R_{XS\text{-}1}$ values were 2.03 ± 0.14 nm and 2.17 ± 0.23 nm respectively, and the neutron $R_{XS\text{-}2}$ values were 1.22 ± 0.05 nm and 1.18 ± 0.12 nm respectively. These R_{XS} values were mostly significantly less than the X-ray values, in agreement with similar observations for other proteins, and this difference was attributed to the non-detection of the hydration shell by neutron scattering. The values of L calculated from the R_G and the $R_{XS\text{-}1}$ values is 19.0 nm for PTerm455 and 19.8 nm for IgA1. The similarity of these values again supported an IgA1 structure in which the tailpiece was packed in a compact structure against the Fc fragment.

Since both IgA and IgG have 12-domain structures (Figure 5.1), it was of interest to compare their sizes. Table 5.2 showed that human IgA consistently showed slightly increased neutron R_G values compared to several mammalian IgG molecules, while the comparison of the neutron $R_{XS\text{-}1}$ and $R_{XS\text{-}2}$ values was more variable. IgA1 was seen to be longer at about 19.5 nm compared to IgG for which L was about 17.5 nm. This showed that the arrangement of the two Fab and Fc fragments in IgA1 is more extended than that in these IgG molecules.

5.3.2. X-ray and neutron distance distribution analyses *P(r)* of PTerm455 and IgA1

The distance distribution function $P(r)$ provided complementary information on the structures of PTerm455 and IgA1. The maxima *M1* and *M2* in these curves

Figure 5.5. (Overleaf) Distance distribution functions $P(r)$ for PTerm455, human IgA1 and bovine IgG2. (*a*) PTerm455 using neutron data measured at 2.4 mg/ml, for which *M1* and *M2* occur at 4.0 nm and 8.7 nm respectively, and L is 20 nm. (*b*) IgA1 using X-ray data measured at 2.1 mg/ml, for which the peaks *M1* and *M2* occur at 3.7 nm and 8.9 nm respectively, and the length L is 21 nm. (*c*) Bovine IgG2 using neutron data (Mayans *et al.*, 1995), for which *M* occurs at 5.4 to 5.7 nm, and L is 16 nm.

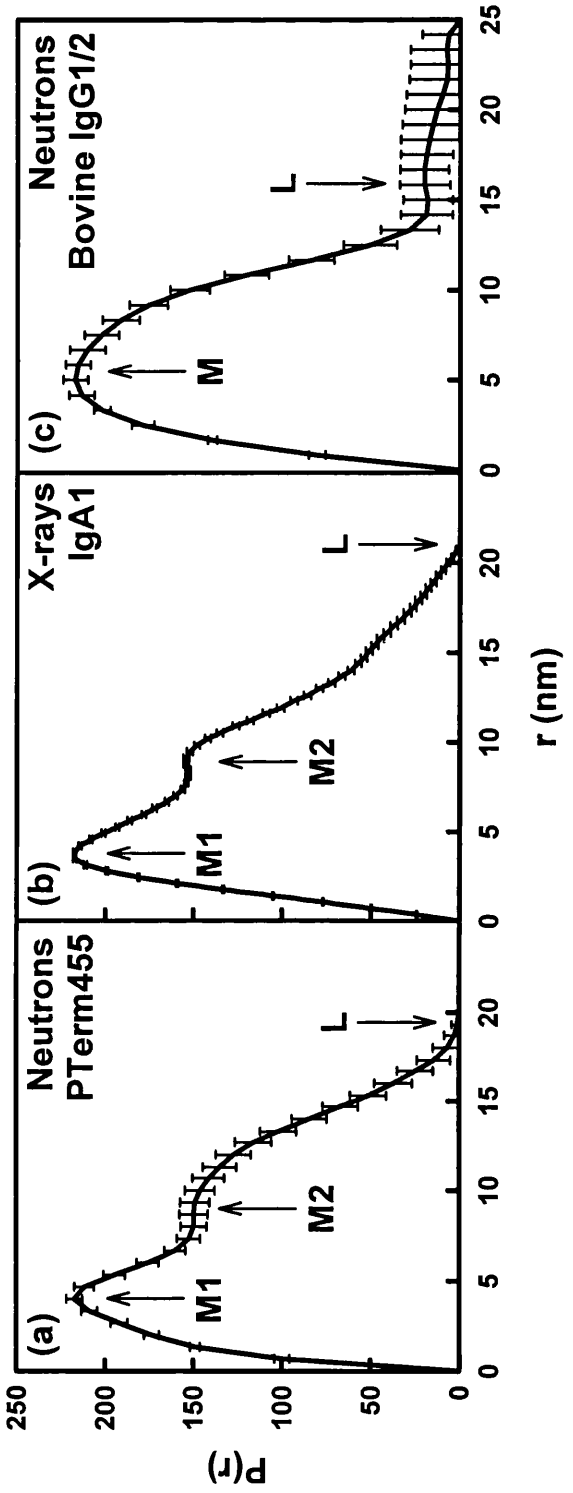


Figure 5.5. Distance distribution functions $P(r)$ for PTerm455, human IgA1 and bovine IgG2 (legend on page 205).

correspond to two sets of frequently occurring interatomic distances within the structure. The neutron and X-ray $P(r)$ curves in Figures 5.5(a) and 5.5(b) showed that the structures of PTerm455 and IgA1 were similar, with two peaks *M1* and *M2* at $r = 3.7$ to 4.0 nm and $r = 8.7$ to 9.1 nm respectively. The R_G values determined from the $P(r)$ curves were 5.95 nm for PTerm455 and 6.12 nm for IgA1. The lengths L of PTerm455 and IgA1 were determined to be 20.0 nm and 21.0 nm respectively. These R_G and L values were in agreement with the Guinier values (Table 5.2).

The observation of two peaks *M1* and *M2* in the $P(r)$ curves of PTerm455 and IgA1 is in contrast with the observation of a single broad peak *M* for bovine IgG1/2 at $r = 5.4$-5.7 nm (Figure 5.5c). This indicated a different arrangement of the Fab and Fc fragments in IgA1 compared to that in bovine IgG1/2. The peak *M1* at 3.7-4.0 nm can be assigned to the most commonly occurring distance within a single Fab or Fc fragment as these are approximately 8 nm long. The peak *M2* is about half the overall length L, and can be assigned to the most common distance over the whole IgA1 structure. That distinct *M1* and *M2* peaks were observed implies that the hinge peptides of IgA1 are relatively rigid and extended, and hold the Fab and Fc fragments apart as distinct structural entities. This explanation is consistent with the composition of the hinge peptides, where the high incidence of O-glycosylation and Pro residues lead to extended structures (Gerken *et al.*, 1989; Shogren *et al.*, 1989; Williamson, 1994). The corresponding hinge in bovine IgG1/2 lacks O-glycosylation sites, and is expected to be more flexible in its structure. The flexibility of the IgG hinge would mean that the two Fab and Fc fragments would adopt a diverse range of structures, and the separate peaks *M1* and *M2* would coalesce into a single peak *M* (Figure 5.5c).

5.3.3. Structure predictions for the three IgA1 α-chain domains

While a crystal structure of a murine IgA Fab fragment is known, and can be directly used for homology modelling, no crystal structure is known for the IgA Fc fragment. IgG crystal structures were therefore used for the homology modelling of the human IgA1 α-chains. This was validated using comparison with five averaged secondary structure predictions for the human α-chain (Figures 5.6 and 5.7). The predictions were calculated from multiple sequence alignments using 27 C_H1 sequences,

Figure 5.6. (Overleaf) Sequence alignment and structure prediction for mammalian IgA C_H1, C_H2 and C_H3 domains, the hinge and the tailpiece. The human IgA α-chain sequences (IgA1 and the three allotypic variants of IgA2) are identified by their SWISSPROT codes except for IgA2(n) (Chintalacharuvu *et al.*, 1994). The α-chain sequences for the other mammalian IgA proteins are identified by their Kabat database codes. The mean secondary structure predictions for the alignment using five methods are denoted by: A, α-helix; B, β-strand; T, turn; C, coil; l, loop; i, buried coil; o, exposed coil. The β-strands are indicated by arrows and are labelled A to G. The consensus solvent accessibilities predicted using three methods are denoted by: e, exposed; b, buried; *, indeterminate.

Figure 5.6. (*a*) Sequence alignment and structure prediction for mammalian IgA C_H1 domain and the hinge (legend on page 208).

Figure 5.6. (*b*) Sequence alignment and structure prediction for mammalian IgA C_H2 domain (legend on page 208).

Figure 5.6. (*c*) Sequence alignment and structure prediction for mammalian IgA C_H3 domain and the tailpiece (legend on page 208).

Figure 5.7. (Overleaf) Sequence alignment and structure prediction for the IgA1 C_H1, C_H2 and C_H3 domains, the hinge and the tailpiece. This is based on the human IgA1 α-chain, the three allotypic variants of human IgA2, and the two crystal structures used for homology modelling. The sequences are identified by their SWISSPROT codes except for IgA2(n) (Chintalacharuvu *et al.*, 1994). The crystal structures are identified using their Brookhaven and SWISSPROT or Kabat codes. Cys residues are identified by vertical lines, the known disulphide bridges are connected, and the unknown disulphide links are marked (?). The N-linked and O-linked glycosylation sites are in bold and underlined.

The mean secondary structure predictions using five methods (Figure 5.6) are denoted by: A, α-helix; B, β-strand; T, turn; C, coil; l, loop; i, buried coil; o, exposed coil. The observed secondary structure in the crystal structures is denoted by: E, β-strand; B, single residue β-ladder; T, turn; S, bend; H, α-helix; G, 3_{10} helix. The β-strands are indicated by arrows and are labelled A to G.

The consensus solvent accessibilities predicted using three methods (Figure 5.6) are denoted by: e, exposed; b, buried; *, indeterminate. The observed solvent accessibilities were denoted by e and b. Bolded letters correspond to residues at the interface between domains on the same polypeptide chain, and bolded and underlined letters corresponds to residues at the interface between domains on different chains.

Figure 5.7. (*a*) Sequence alignment and structure prediction for the IgA1 C_H1 and the hinge (legend on page 212).

Figure 5.7. (*b*) Sequence alignment and structure prediction for the IgA1 C_H2 domain (legend on page 212).

Figure 5.7. (*c*) Sequence alignment and structure prediction for the IgA1 C_H3 domain and the tailpiece (legend on page 212).

22 C_H2 sequences and 21 C_H3 sequences (Methods; Figure 5.6). The predicted secondary structures were correlated with those observed in the murine IgA and human IgG1 C_H1 crystal structures and the human IgG1 C_H2 and C_H3 crystal structures (Figure 5.7). Consensus β-strands were identified if they occurred in the majority of the predictions, and were labelled A to G in accordance with the DEBA|GFC(C') nomenclature for C1-set Ig folds (Bork *et al.*, 1994; Chothia & Jones, 1997). Figure 5.7 showed that the predictive accuracies were good at 86%, 74% and 68% for the C_H1, C_H2 and C_H3 domains respectively. Most of the β-strands were predicted and were used to define the homology modelling (Methods). The exceptions were β-strands A and D in the C_H2 domain and β-strand G in the C_H3 domain. Since the C_H2 β-strands A and D correspond to two edges of a β-sheet, this location may account for these failed predictions (Figure 5.7b). The mispredicted C_H3 β-strand G has high sequence similarity with IgG, but may be influenced by its proximity to the tailpiece (Figure 5.7c). Interestingly, no regular secondary structure was predicted for the hinge between the C_H1 and C_H2 domains (Figure 5.7a), and the tailpiece at the C-terminus of the C_H3 domain was predicted to contain a β-strand (Figure 5.7c).

Using the alignments of Figure 5.7, three averaged solvent accessibility predictions for each of the three IgA C_H domains were performed to correlate the α-chain sequences with the IgG crystal structures. These were compared with accessibilities calculated for each of the isolated domains outside the Fab or Fc crystal structures. For all three C_H domains, the predictions showed good accuracies of 76%, 76% and 70% for the C_H1, C_H2 and C_H3 domains respectively. They were generally successful in identifying exposed loop residues and the alternation of buried and exposed residues in β-strands. Many specific differences between the predicted and observed accessibilities correspond to the effect of domain packing within the Fab and Fc structures, and are discussed below.

5.3.4. Homology modelling of the IgA1 Fab fragment

Homology models were used for constrained scattering curve fits in order to enhance the utility of these scattering models for IgA1. For the Fab fragment (Figure 5.1), the crystal structure of the human IgG1 TR1.9 Fab fragment (Chacko *et al.*, 1996)

was used to model those in human IgA1 because it contains a κ-light chain, which is more common than λ-light chains in human antibodies, and a κ-light chain is also found in the murine IgA J539 Fab crystal structure. To verify this procedure, the sidechain solvent accessibilities at the C_L-C_H1 interface were examined. Extensive contacts are made between the DEBA faces of the two Ig folds. Of 18 C_L exposed residues that became partially or fully buried at the C_L-C_H1 interface in both crystal structures, 15 were identical between the murine IgA and human IgG1 sequences. Of the 13 C_H1 exposed residues at this interface, 6 were identical between the murine IgA and human IgG1 sequences, and 7 were identical between the human IgA1 and human IgG1 sequences (Figure 5.7a). Indeed 8 of these 13 residues in human IgA1 were predicted to be buried even though they were exposed at the domain surface. This conservation indicated that the IgG1 TR1.9 Fab structure provided a good basis for modelling the human IgA1 Fab structure. Further support for this conclusion involves Cys133 in the C_H1 domain at the loop between strands A and B, which is bridged with the C-terminal cysteine of the C_L domain (Biewenga & van Run, 1992; Chintalacharuvu & Morrison, 1996). This disulphide bridge was readily created by a slight energy refinement of the A-B loop of the C_H1 domain and the four C-terminal residues of the C_L domain.

5.3.5. Homology modelling of the IgA1 Fc fragment

The homology model of the human IgA1 Fc fragment was based on using the human IgG1 Fc crystal structure (Deisenhofer, 1981) as a template on which the IgA1 C_H2 and C_H3 homology models were superimposed (Methods). The two IgG1 C_H3 domains interacted through their DEBA faces in a manner analogous to the C_L-C_H1 interaction. The predicted and observed accessibilities were used to verify the buried residues at the domain interface (Figure 5.7c). Of the 20 exposed residues on each C_H3 domain that became buried at the C_H3-C_H3 interface (Figure 5.7c), 5 were identical to residues in the human IgA1 sequence, and 9 were predicted to be buried. This justified the use of the human IgG1 crystal coordinates for the homology modelling of the C_H3 domains.

Differences were observed for the C_H2 domains. In the human IgG1 Fc crystal structure, the two C_H2 domains contain a void space between them which is occupied

by carbohydrate at Asn299 in the numbering of Figure 5.7(b). In IgA1, the N-linked glycosylation site at Asn263 replaces that at Asn299 in IgG1, and Asn263 is now on the outside of the Fc fragment (Figure 5.3). In the sequence alignment, a one-residue deletion occurred at the link between the C_H2 and C_H3 domains of IgA1 compared to IgG1 (Figure 5.7b). The consensus accessibility prediction for the IgA1 C_H2 domain suggested that the D-E loop and most of β-strand E are buried, which is contrary to the observed sidechain accessibilities in the IgG1 Fc crystal structure. Eight Cys residues (Cys241, Cys242, Cys299, Cys301) are in proximity to each other at the junction of the hinge with the C_H2 domain, however different disulphide bridge assignments have been proposed. While there is agreement on the Cys241-Cys241 bridge, mutually inconsistent Cys242-Cys299, Cys242-Cys301, Cys299-Cys299, Cys301-Cys301 and Cys299-Cys301 bridges have been proposed (Putnam *et al.*, 1979; Yang *et al.*, 1979; Biewenga & van Run, 1992). If Cys241-Cys241 and Cys242-Cys301 or Cys242-Cys299 are bridged, this can be accommodated without major change in the IgA1 Fc model, although Cys 299 or Cys301 will be left free. If either Cys299-Cys299 or Cys301-Cys301 are bridged as well, a significant conformational rearrangement of the C_H2 domains is required. It was decided to use the domain arrangement observed in the IgG1 Fc fragment for the scattering curve fits, as this involved the fewest assumptions. Nonetheless curve fits based on the inclusion of a presumed Cys299-Cys299 bridge in a second IgA1 Fc model were performed as a control (Methods).

5.3.6. Translational search for an IgA1 solution structure (Method 1)

The optimisation of the positions of the IgA1 Fab and Fc fragments against the scattering data was firstly based on a translational search method used previously for bovine IgG (Mayans *et al.*, 1995). This employed the experimental neutron data for PTerm455 and X-ray data for IgA1. The three-dimensional search systematically explored the space around the Fc fragment to indicate whether curve fits were possible for a given arrangement of Fab fragments that was fixed in orientations observed in IgG (Figure 5.3). The two Fab fragments were positioned symmetrically relative to the Fc fragment in order to simplify the modelling.

Table 5.3. Three modelling searches for the IgA1 structure

Modelling procedure	Filter	Models	Hinge length (nm)	Data	Spheres	R_G (nm)	R_{XS-1} (nm)	R_{XS-2} (nm)	R-factor (%)
Method 1	None	35,301	0 -17	Neutron	1010 - 1258	3.3 - 7.3	0.0 - 3.1	0.1 - 2.2	5.0 - 27
Fab translations			(full range)	X-ray	1293 - 1678	3.4 - 8.9	0.0 - 4.1	0.0 - 2.3	4.8 - 37
	N > 1163 $5.6 < R_G < 6.4$	2,562	8.5 ± 2.6 (mean)	Neutron	1222 ± 10	5.8 ± 0.2	2.0 ± 0.4	1.2 ± 0.2	8.2 ± 5.0
	N > 1534 $5.6 < R_G < 6.4$	2,562	8.5 ± 2.6 (mean)	X-ray	1612 ± 13	6.2 ± 0.1	2.0 ± 0.6	1.3 ± 0.3	13.6 ± 7.6
Method 2 : Hinge	None	12,000	0.4 - 8.2 (full range)	Neutron	811 - 1346	2.7 - 6.8	0.0 - 2.9	0.7 - 2.5	5.0 - 28
molecular dynamics				X-ray	997 - 1816	2.8 - 7.8	0.0 - 3.3	0.1 - 2.8	4.8 - 29
	N > 1245 $5.6 < R_G < 6.4$	867	6.8 ± 1.0 (mean)	Neutron	1299 ± 11	5.7 ± 0.1	2.1 ± 0.3	1.3 ± 0.1	6.6 ± 0.7
	N > 1647 $5.6 < R_G < 6.4$	867	6.8 ± 1.0 (mean)	X-ray	1727 ± 17	6.2 ± 0.1	2.1 ± 0.4	1.5 ± 0.1	9.4 ± 2.2
Method 3 : Tailpiece	None	16,000	7.0	Neutron	1319 - 1379	5.7 - 5.9	1.9 - 2.1	1.4 - 1.5	6.3 - 7.1
molecular dynamics				X-ray	1763 - 1860	6.1 - 6.4	1.8 - 2.0	1.5 - 1.6	5.4 - 7.1
	N > 1305 $5.6 < R_G < 6.4$	16,000	7.0	Neutron	1361 ± 6	5.8 ± 0.4	2.0 ± 0.1	1.5 ± 0.1	6.6 ± 0.1
	N > 1729 $5.6 < R_G < 6.4$ $R < 6.3$	2,243	7.0	X-ray	1816 ± 14	6.2 ± 0.1	1.9 ± 0.1	1.6 ± 0.1	6.1 ± 0.1

[1] The hinge length is defined as the distance between the α-carbon of Cys220 and the α-carbon of Pro244.

The search created 35,301 models in a translational search of each Fab fragment within a cube of 10 nm × 20 nm × 20 nm relative to a fixed Fc fragment. The scattering curves were calculated using previously calibrated methods (Perkins *et al.*, 1998a). The hinge lengths in the 35,301 models ranged from 0 to 17 nm, compared to the theoretical maximum hinge length of 8.4 nm (Table 5.3). The 35,301 calculated neutron and X-ray R_G values ranged between 3.3 and 8.9 nm, the $R_{XS\text{-}1}$ values ranged between 0.0 to 4.1 nm, and the $R_{XS\text{-}2}$ values ranged from 0.0 to 2.3 nm (Table 5.3). These encompassed the experimental PTerm455 and IgA1 values in Table 5.2 as desired. The *R*-factor values assess the quality of the curve fit between the modelled and experimental curves, and these ranged from 5.0% to 27% for the neutron models, and from 4.8% to 37% for the X-ray models.

The 35,301 models were evaluated in contour plots for the neutron and X-ray curve fits and showed that the search was unsuccessful (Figure 5.8). Each model was characterised by the distance between the centres of mass of its two Fab fragments and that between the centres of mass of a Fab fragment and the Fc fragment. Figure 5.8 was interpreted as follows:

(i) The starting point of the models involved the steric overlap of the Fab and Fc fragments (Figure 5.3). These overlapping models were identified from the lower contours of Figures 5.8(a) and 5.8(d) where the Fab-to-Fab and Fab-to-Fc separations were too low at 7 nm or less, and these were rejected when the filters for steric overlap (*N*) were applied. It was impossible to produce models that have a large Fab-to-Fab distance and a small Fab-to-Fc distance, and this accounts for the distribution of models seen in Figure 5.8(g).

(ii) The neutron and X-ray R_G contour plots showed the lowest R_G values at small Fab-to-Fab and Fab-to-Fc distances, and these R_G values increased as these distances were increased (Figures 5.8b and 5.8e). The best-fit R_G contours correspond to two curves of similar-fit models. The joint application of filters for both the neutron and X-ray *N* and R_G values left 2562 models (7% of the starting total) with the two distributions shown in Figure 5.8(h). One of these ran close to the R_G contours of Figures 5.8(b) and 5.8(e), while the other corresponded to hinge lengths greater than the maximum allowed of 8.4 nm and was disallowed.

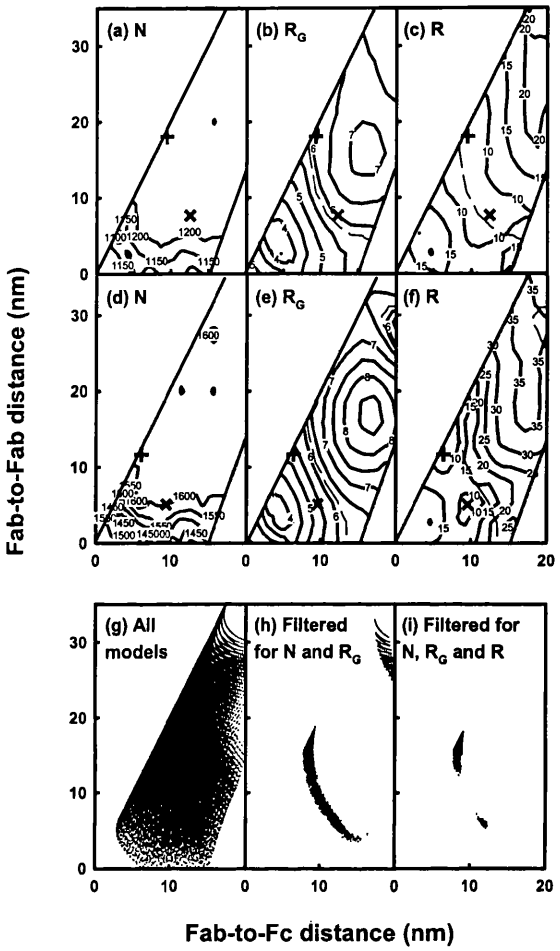


Figure 5.8. Outcome of the translational search of the Fab fragments relative to the Fc fragment using Method 1. The contour maps show the outcome of the neutron fits for dry PTerm455 in (*a*, *b*, *c*) and that for the X-ray fits for hydrated IgA1 in (*d*, *e*, *f*) as a function of the distance between the centres of mass of the two Fab fragments and that between the Fab and Fc fragments. The distribution is shown in (*g*, *h*, *i*) of all 35,301 models as a function of these two distances, and the 2,562 and 958 models that remain after the simultaneous application of filters for the neutron and X-ray curve fits (second column of Table 5.3). The models shown in (*i*) have an *R*-factor less than 10%.

(*a*, *d*) The number of spheres *N* in the dry and hydrated models, where the target values are 1224 and 1618 spheres respectively (Table 5.1), and steric overlap is visible at the bottom.

(*b*, e) The radius of gyration R_G for the dry and hydrated models, where the target values are 5.84 nm for dry PTerm455 and 6.20 nm for hydrated IgA1 (Table 5.2). The dotted line indicates the contour for the target R_G values.

(*c*,*f*) The *R*-factor for the dry and hydrated modelling curve fits show two best-fit minima for each of the dry and hydrated models that correspond to the + and × symbols.

(iii) The neutron and X-ray *R*-factor contours each exhibited two shallow minima at values of 6-7% in Figures 5.8(c) and 5.8(f). The minima occurred close to the R_G contour line in both cases, and confirmed the outcome of the search. At this point, limitations in the translational search became evident. The joint application of a 10% filter for the neutron and X-ray *R*-factor left 958 models at two best-fit minima and not one (Figure 5.8i). The application of a joint 7% filter left only 25 models, even though many more good-fit models would have been expected. The separate neutron and X-ray analyses should have resulted in common positions for the *R*-factor minima, however these were seen to be different in Figures 5.8(c) and 5.8(f). Despite the success of this search method for bovine IgG1 and IgG2 (Mayans *et al.*, 1995), it was concluded that improvements were required for modelling the structure of human IgA1.

5.3.7. Molecular dynamics search for a PTerm455 solution structure (Method 2)

Method 1 was limited by holding the two Fab fragments in a fixed orientation in the IgA1 models and not allowing for the presence of the hinges. To allow for these, molecular dynamics simulations were used to produce a large number of random structures for the 23-residue hinge peptide. This enabled the Fab and Fc fragments to be connected to the hinge peptide in any orientation, and this gave a complete model for PTerm455. The first simulation generated 5,000 hinge models with lengths between 0.4 nm and 5.6 nm. Since these lengths were significantly below the maximum hinge length of 8.4 nm, further simulations were performed in which its minimum length was set to be either 2, 3, 4, 5, 6, 7 or 8 nm. This gave 12,000 hinge models with a full range of lengths (Figure 5.9). These were combined with the Fab and Fc fragments to create PTerm455 models for the calculation of the neutron and X-ray curves.

The 12,000 models gave the lowest *R*-factor values for the most extended hinges of lengths 6 nm to 8 nm (Figure 5.9). The evaluation of the 12,000 models in the contour plots of Figure 5.10 showed an improved outcome compared to that of Figure 5.8, and is summarized as follows:

(i) The distribution of the models seen in Figure 5.10(g) is more limited for reason of the constraint imposed by the length of the hinge peptide in every model. Sterically overlapping models occurred again at the bottom of Figures 5.10(a) and

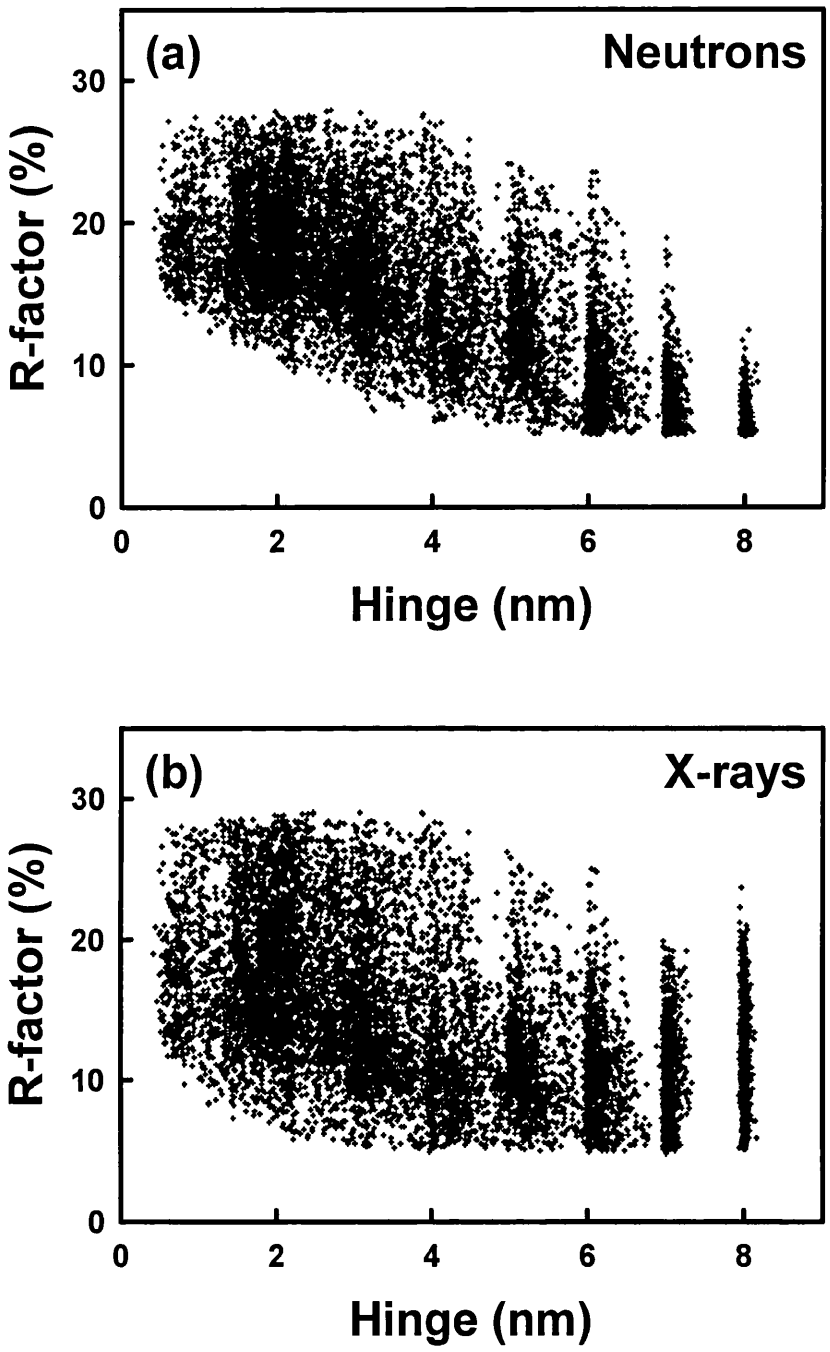


Figure 5.9. Distribution of the 12,000 models generated by molecular dynamics simulation of IgA1 hinge structures. The R-factor of the 12,000 models is shown as a function of hinge length, where the maximum length of the 23-residue hinge is 8.4 nm. (*a*) and (*b*) correspond to the curve fits for dry PTerm455 models and hydrated IgA1 models.

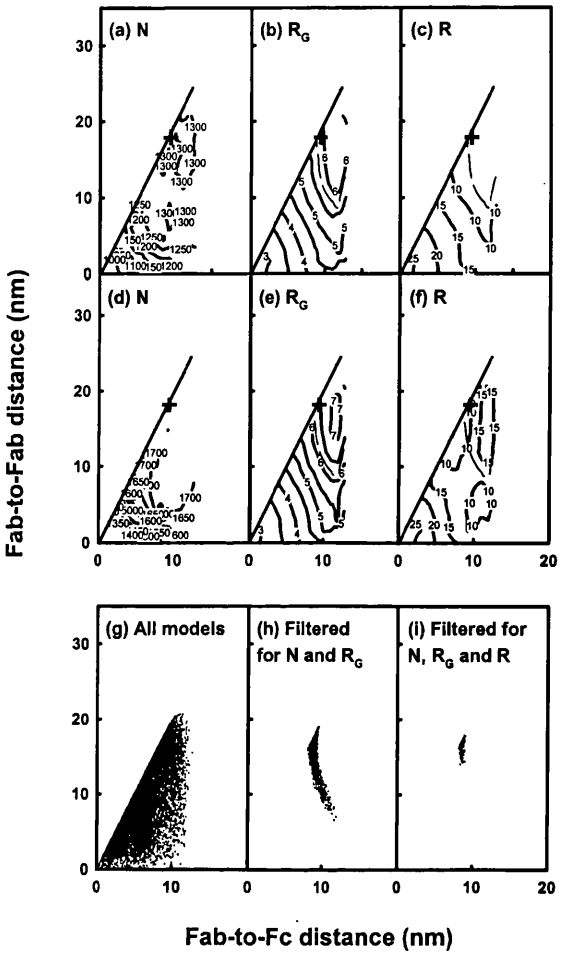


Figure 5.10. Outcome of the molecular dynamics searches of hinge structures connecting the Fab and Fc fragments using Method 2. The contour maps show the outcome of the neutron fits for dry PTerm455 in (*a*, *b*, *c*) and that for the X-ray fits for hydrated IgA1 in (*d*, *e*, *f*) as a function of the distance between the centres of mass of the two Fab fragments and that between the Fab and Fc fragments. The distribution is shown in (*g*, *h*, *i*) of all 12,000 models as a function of these two distances, and the 867 and 104 models that remain after the simultaneous application of filters for the neutron and X-ray curve fits (second column of Table 5.3). The models shown in (*i*) have an *R*-factor better than 7%, and are found at one minimum and not to the two seen in Figure 5.8.

(*a*, *d*) The number of spheres *N* in the dry and hydrated models, where the target values are 1310 and 1734 spheres respectively (Table 5.1), and steric overlap is visible at the bottom.

(*b*, *e*) The radius of gyration R_G for the dry and hydrated models, where the target values are 5.84 nm for dry PTerm455 and 6.20 nm for hydrated IgA1 (Table 5.2). The dotted line indicates the contour for the target R_G values.

(*c*, *f*) The *R*-factor for the dry and hydrated modelling curve fits show a single best-fit minimum that is coincident in both the dry and hydrated models and denoted by the + symbol.

Table 5.4. Comparison of modelling curve fits for human IgA1 and bovine and murine IgG

Model	Reference	Hinge length (nm)	Fab-Fab distance (nm)	Fab-Fc distance (nm)		Modelled R_G (nm)	Source of data	Observed R_G (nm)	R-factor (%)
Direct modelling									
PTerm455	This work (Method 2)	7.0, 7.0	16.9	8.9, 8.9	Neutron	5.69	PTerm455	5.84	6.4
		(23 residues)			X-ray	6.17	=	6.16	=
Serum IgA1	This work (Method 3)	7.0, 7.0	16.9	8.9, 8.9	Neutron	5.69	=	6.11	=
		(23 residues)			X-ray	6.15	IgA1	6.20	5.4
Curve comparison									
Bovine IgG1/2	Mayans *et al.* (1995)	3.5, 3.6	8.1	8.4, 8.0	Neutron	5.03 *	IgG1/2	5.64-5.71	5.3 **
		(12, 19 residues)			Neutron	5.03 *	PTerm455	5.84	10.8
					X-ray	5.24	IgA1	6.20	9.5
Murine IgG1	Harris *et al.* (1998a)	3.7, 3.8	7.3	6.1, 7.4	Neutron	4.48	IgG1/2	=	8.8
	(Code 1igy)	(15 residues)			Neutron	4.48	PTerm455	5.84	14.1
					X-ray	4.61	IgA1	=	11.8
Murine IgG2a	Harris *et al.* (1997)	3.8, 4.9	9.4	7.4, 8.2	Neutron	5.05	IgG1/2	=	5.3
	(Code 1igt)	(21 residues)			Neutron	5.05	PTerm455	5.84	10.5
					X-ray	5.31	IgA1	=	9.9

* These were recalculated using the Q ranges employed for IgA1 here for reason of consistency.

** This was erroneously reported as 1.2% in Mayans *et al.* (1995) and Perkins *et al.* (1998a).

5.10(d), and these were removed when the steric overlap parameter N was applied as a filter.

(ii) In the neutron and X-ray R_G contour plots (Figures 5.10b and 5.10e), the two best-fit R_G contour lines were closer to each other than those seen in Figure 5.8. The joint application of filters for both the neutron and X-ray N and R_G values left 867 best-fit models (7% of the starting total) within a single distribution shown in Figure 5.10(h).

(iii) The neutron and X-ray R-factor contours again exhibited two shallow minima in Figures 5.10(c) and 5.10(f), but these now occupied the same position unlike Figure 5.8, and showed that the neutron and X-ray analyses were now self-consistent. This time, the joint application of a neutron and X-ray 7% filter for the R-factor left a higher number of 104 good fit models at a single narrower minimum in Figure 5.10(i) compared to Figure 5.8(i). The mean Fab-to-Fab distance of 17 nm and mean Fab-to-Fc distance of 9 nm showed that T-shaped models were favoured, not Y-shaped models.

As an illustration of the 104 best-fit models, the model giving an R-factor of 6.3-6.4% (Table 5.4) for both the neutron and X-ray curves gave the curve-fit in Figure 5.11(a) which extended out to Q of 1 nm^{-1}. The difference beyond Q of 1 nm^{-1} is attributable to a flat background from incoherent scattering in the protein sample. The superposition of all 104 best-fit models showed that the Fab fragments were positioned in an extended T-shaped arrangement relative to the Fc fragment (Figures 12a and 12b). There, the Fab fragments formed a clear circle of possible fits about the central location of the Fc fragment, in which the range of hinge conformations is visible at the centre and a single unique structure could not be identified from those shown. The mean Fab-to-Fab distance was 17 nm, and the mean Fab-to-Fc distance was 9 nm. It was noticeable that there were not any Y-shaped structures among the 104 best-fit models in Figure 5.12.

5.3.8. Molecular dynamics search for an IgA1 solution structure (Method 3)

In order to create the full IgA1 structure, two tailpieces with 58 amino acid and carbohydrate residues (5% of IgA1 w/w) were added to the best-fit PTerm455 model from Method 2. Molecular dynamics was used to generate 16,000 tailpiece models (Methods). The addition of these to the PTerm455 model gave IgA1 structures with a

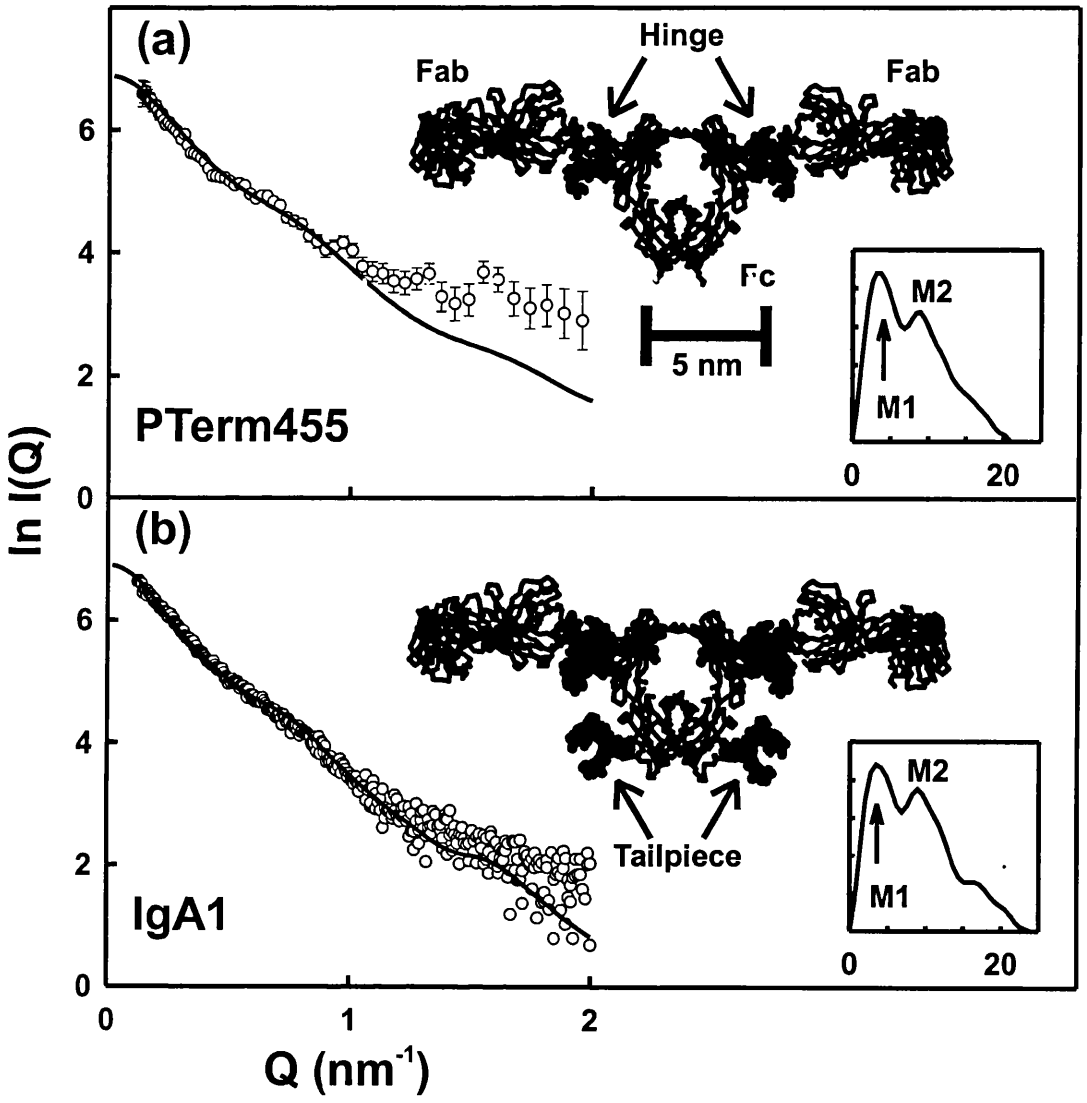


Figure 5.11. Final curve fits for (*a*) the neutron model of PTerm455 and (*b*) the X-ray model of serum IgA1. The modelled curves (continuous lines) are compared with neutron data for PTerm455 in 100% ^2H_2O from Instrument LOQ and X-ray data for IgA1 from Station 2.1 (open circles). Experimental error bars are shown when significant. The *P*(*r*) curves are shown as insets for comparison with Figure 5.5. The α-carbon traces of both models are shown, with the N-linked and O-linked oligosaccharide chains shown in bold.

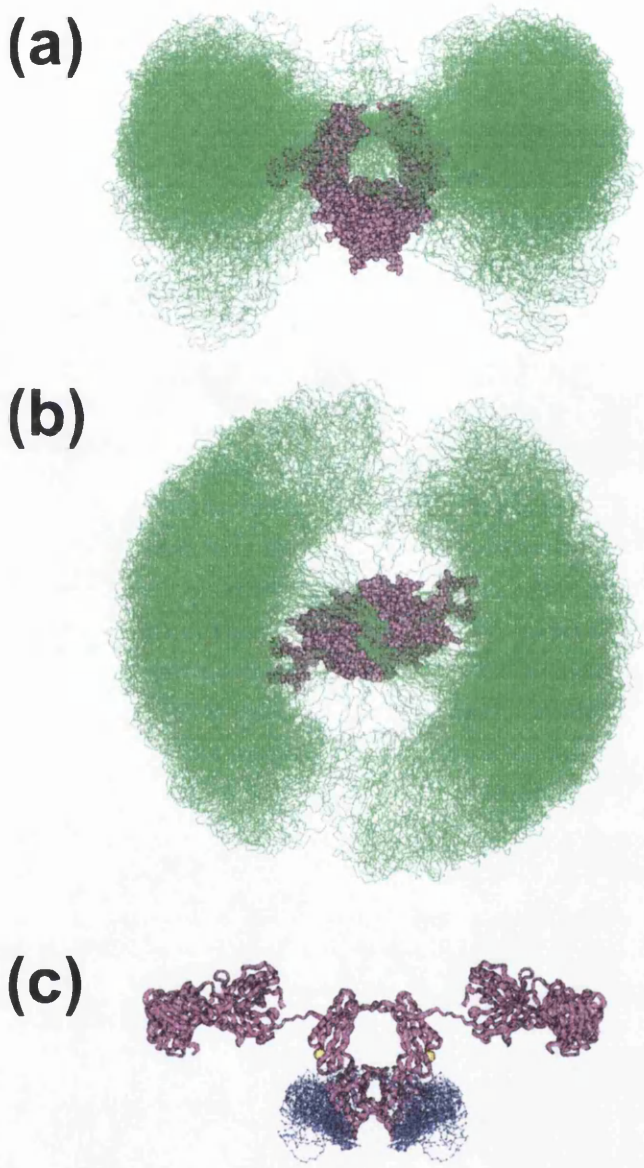


Figure 5.12. The best-fit models for PTerm455 and IgA1. For PTerm455, two orthogonal views of the superimposition of the 104 best-fit models for IgA1 are shown in (*a*) and (*b*). The Fc fragment is shown in purple at the centre, while the 104 pairs of Fab fragments are shown in green. For IgA1, a best-fit model for PTerm455 is shown in (*c*) in purple with the 82 best-fit tailpiece structures shown in blue.

wide range of tailpiece conformations.

Table 5.3 showed that all 16,000 models satisfied the generous neutron and X-ray filters used for Method 2. Since the X-ray curve corresponded to the intact serum IgA1 molecule, and the best-fit PTerm455 model had given an *R*-factor value of 6.3% with this in Method 2, an *R*-factor filter of 6.3% was applied to the 16,000 models to leave 2,243 models (14% of the starting total). Slightly improved curve fits were obtained, where the mean *R*-factor of 9.4% for the 867 models of Method 2 was reduced to 6.1% for the 2,243 models of Method 3. The distribution of the 82 tailpiece conformations giving the lowest *R*-factors of 5.4-5.7% relative to the Fc fragment is shown in Figure 5.12. These were generally positioned folded back close to the surface of the Fc fragment. The mean tailpiece length of the final 82 models was 2.9 ± 1.0 nm and that for the 2,243 models was 4.9 ± 1.5 nm. These folded-back models would account for the observed similarity in the Guinier and $P(r)$ analyses for PTerm455 and IgA1, although it should be noted that the difference in R_G values between the most extended and most compact tailpiece models is only 0.3 nm (Table 5.3).

Control calculations were performed. The positions of the Fab and Fc fragments in the best-fit model (Figures 9a and 10c) were used to guide the replacement of the IgG-based Fc homology model with a second Fc model that contained four possible disulphide bridges in IgA1 (not shown), in which the hinge peptides made contact with both the Fab and Fc fragments (Methods). This had little effect on the curve fits, where the *R*-factor changed by less than 0.7%, and showed that the modelling did not depend on the presumed disulphide bridges within the Fc fragment of IgA1. In another control, the O-linked oligosaccharides at the hinge were replaced with those recently reported in Mattu *et al.* (1998) (Methods). This also had little effect on the fits, where the *R*-factor changed by less than 0.2%.

Given that Cys311 on the C_H2 domain and Cys471 on the tailpiece were free (Figure 5.1), the IgA1 scattering models were analysed to see whether Cys311 and Cys471 may form a bridge. In the homology model, Cys311 is surface exposed at the base of the C_H2 domain, and is separated by 5.6 nm from its neighbouring Cys311 which

is too distant to form a bridge as proposed by Yang *et al.* (1979). However Cys311 is 4.5 nm away from Lys454 at the C-terminus of the Fc fragment and potentially within range of Cys471 if the tailpiece is maximally 6.6 nm in length (Figure 5.3). A β-strand was predicted in the tailpiece (Figure 5.7a), which if present may be associated with the edge of a β-sheet in the C_H3 domain to bring the tailpiece into a suitable position to permit formation of a Cys311-Cys471 bridge. The scattering analyses revealed a distribution of best-fit folded-back tailpiece conformations (Figure 5.12) which would be consistent with a Cys311-Cys471 bridge (Prahl *et al.*, 1971) and the absence of free Cys residues in IgA1 (Biewenga & van Run, 1992).

5.3.9. Comparison of the IgA1 and IgG solution structures

The IgA1 model of Figures 5.11 and 5.12 is distinct from that for bovine IgG1 and IgG2 in Mayans *et al.* (1995), and those for two crystal structures for murine IgG1 and IgG2a (Harris *et al.*, 1998b). Even though the hinge lengths are similar between 12-23 amino acid residues in these proteins, the arrangements of the Fab and Fc fragments in these models are significantly different (Table 5.4). The Fab-to-Fab distance for IgA1 is 16.9 nm, which is double those of 7.3-9.1 nm seen for the IgG structures. The Fab-to-Fc distance for IgA1 is 8.9 nm which is at the upper limit of those of 6.1-8.4 nm seen for the IgG structures.

The dissimilarity between the IgA1 and IgG structures was further tested by comparing the calculated scattering curves from the crystal structures of murine IgG1 and IgG2a with the experimental curves for bovine IgG1/2, PTerm455 and IgA1 (Figure 5.13). The IgG2a model gave good agreement with the bovine IgG1/2 neutron data, giving similar neutron R_G values and a low *R*-factor of 5.3%, however the $P(r)$ curve gave two peaks *M1* and *M2*. The IgG1 model gave poorer agreement with the bovine IgG neutron data, as this had too compact a structure, but the $P(r)$ curve now gave a single peak. Visual inspection of both these curve fits with the bovine IgG1/2 data showed reasonable but not exact agreement. This is as expected, since the two crystal structures correspond to two snapshots of single discrete conformations that have been trapped by the crystal lattice from many that almost certainly exist in solution (Harris *et al.*, 1998b), while the bovine IgG1/G2 solution data corresponds to an average of all

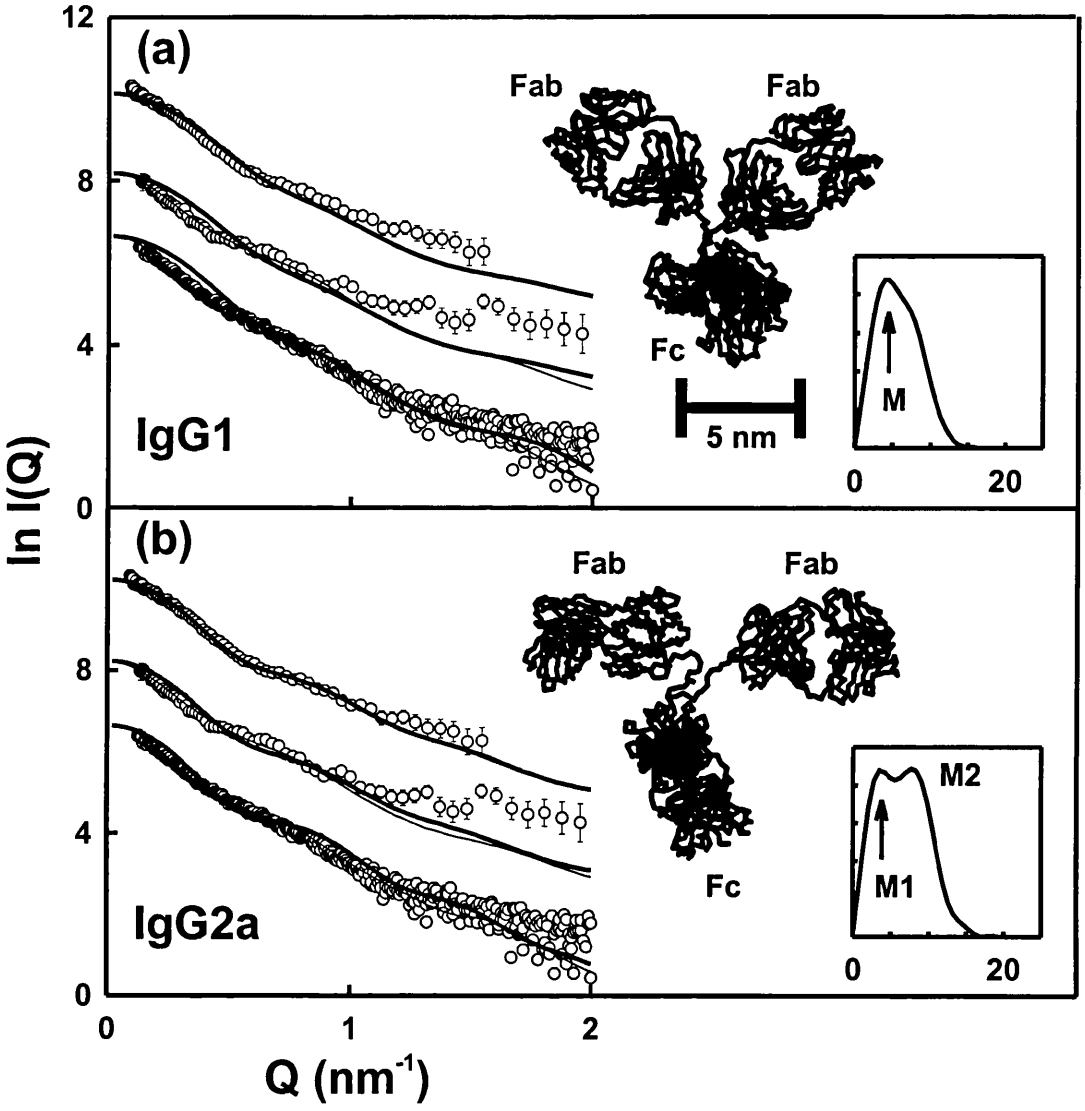


Figure 5.13. Curve fits based on (*a*) the murine IgG1 and (*b*) the murine IgG2a crystal structures. The calculated scattering curves (bold continuous line) correspond to the dry neutron and hydrated X-ray models for these structures. These are compared with the experimental neutron curve for bovine IgG1/2 (Mayans *et al.*, 1995) and human PTerm455, and the experimental X-ray curve for serum IgA1. The thin continuous lines correspond to the best-fit curves for these three models. The $P(r)$ curves are shown as insets for comparison with Figure 5.5.

conformations. In distinction to these, the comparisons of the IgG models with the PTerm455 and IgA1 data gave much worsened curve fits. The *R*-factors were higher at 9.9-14.1%, and clear curve-fit discrepancies occurred at the highest curve intensities at the lowest Q values that reflected the increased R_G values of PTerm455 and IgA1 in comparison to murine IgG1 and IgG2a. These comparisons showed that the IgG and IgA1 structures did not show reasonable similarities to each other, in support of the modelling of IgA1 in Figures 11 and 12.

5.4. Conclusions

The combination of solution scattering analyses and tightly constrained modelling have provided new insights on the solution structure of IgA1. Most importantly, even though IgA1 and IgG1/2 shared a common 12-domain protein structure, this study showed that the arrangement of the Fab and Fc fragments within these structures are different in the two antibody classes. The Fab fragments in the IgA1 structure are widely separated compared to that in IgG. This had previously not been recognised, and is the first evidence that IgA immunoglobulins form a distinct structure from that seen for IgG, in much the same way that immunoglobulins E and M constitute structurally distinct classes from that of IgG by the insertion of an extra pair of domains to replace the hinges in IgG. The experimental basis for this result is based on the larger R_G values for IgA1 compared to IgG1/2 (Table 5.2) and more so by the distinctive $P(r)$ curves for IgA1 and IgG1/2 (Figure 5.5). Difficulties were experienced in modelling IgA1 based on a Fab arrangement seen in IgG (Figures 5.8 and 5.13), while an improved search based on independently positioned Fab fragments gave self-consistent X-ray and neutron modelling analyses and structures that were distinct from those for IgG (Modelling Methods 2 and 3; Figures 5.10, 5.11 and 5.13; Table 5.4). Even though solution scattering is unable to identify unique structures because of rotational averaging, it is able to rule out incorrect structures such as those based on IgG. The basis for the different structures in IgA1 and IgG1/2 is attributable to the O-linked glycosylation of the hinge peptide in IgA1, together with the disulphide arrangement at the top of the Fc fragment.

The IgA1 model leads to an appreciation of the functional properties of IgA1 by

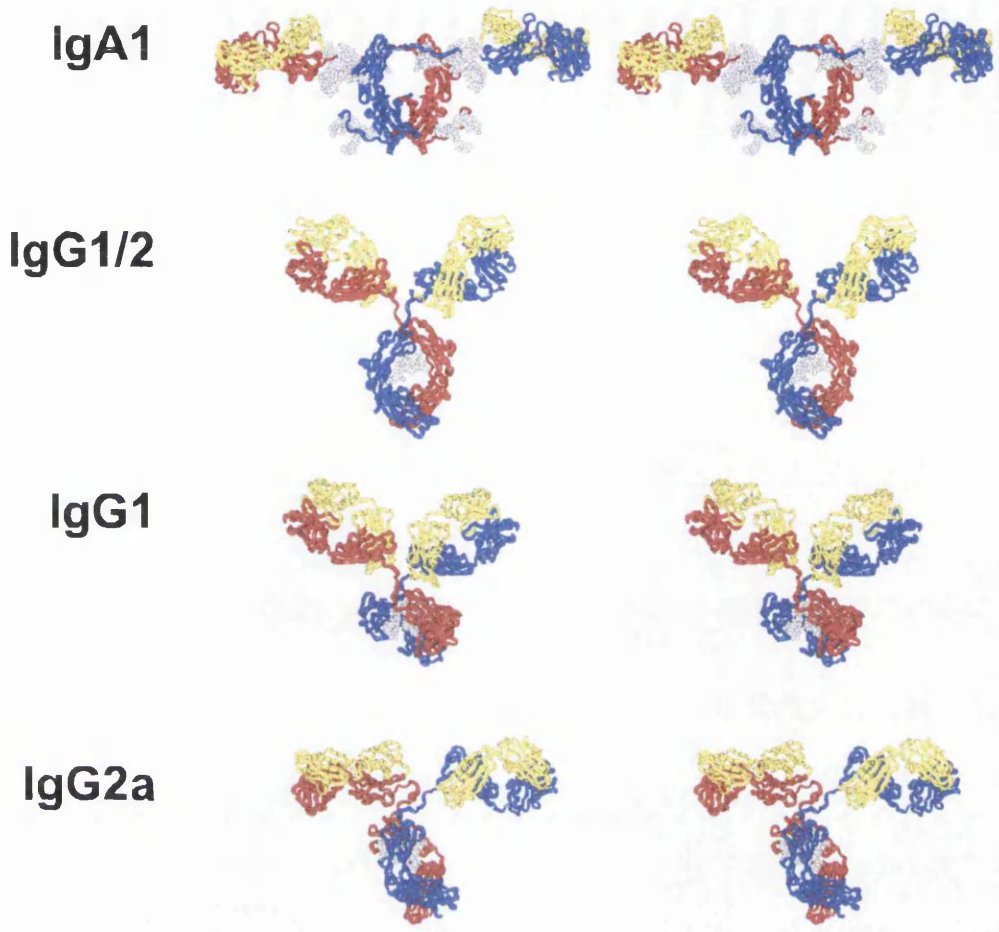


Figure 5.14. Stereoview ribbon representations for a best fit IgA1 model, the neutron model of bovine IgG1/2 and the murine IgG1 and IgG2a crystal structures. The light chains are shown in yellow. The heavy chains are shown in blue and red. The oligosaccharides are shown in grey.

considerations of the antigen sites at the tips of the Fab fragment, the site for the IgA receptor FcαR on the Fc fragment, and the exposure of the hinge region. The views of IgA1 in Figures 5.11, 5.12 and 5.14 show that all three regions are accessible, and this is consistent with the known link between antigen-binding and effector functions. This linkage does not occur through conformational changes within the antibody molecule; rather it takes place via Fc-Fc interaction or by the association of multiple antigen-complexed antibodies (Harris *et al.*, 1998b). In relation to the Fab fragments, tip-to-tip separation between the two Fab antigen-binding sites in the IgA1 model (Figure 5.14) is high at 23 nm, and this may be important for facilitating IgA1 interactions with a class of widely-separated epitopes on foreign particles. This Fab tip-to-tip separation is closer together at 13-16 nm for the IgG structures (Figure 5.14). IgA1 may therefore possess an immune advantage in comparison to IgG in that its structure would permit a greater diversity of immune targeting. This may account for the relative abundance of IgA1 in serum in comparison to IgA2 where this hinge is missing to bring the two Fab fragments closer together. IgA1 also shows the same potential advantage in comparison to immunoglobulin M, in which the tip-to-tip separation between its Fab fragments is about 13 nm (Perkins *et al.*, 1991). In relation to the Fc fragment, the myeloid FcαR site in IgA1 is found at the C_H2-C_H3 interface (Carayannopoulos *et al.*, 1996) and this is similar to the binding site in IgG for bacterial proteins A and G (Deisenhofer, 1981; Sauer-Eriksson *et al.*, 1995), rheumatoid factor (Corper *et al.*, 1997) and rat neonatal Fc receptor (Burmeister *et al.*, 1994). In the IgA1 models of Figures 5.12(a) and 5.12(b), many Fab-Fc orientations are shown, in which a large proportion will make the FcαR sites at the C_H2-C_H3 interface accessible for interactions with FcαR. Finally, in relation to the IgA1 hinge, their extended glycosylated structures in the models of Figure 5.12 may offer an advantage over the nonglycosylated hinge of IgG in that they offer protection from proteolytic attack while maintaining the ability to reach widely-separated epitopes. It is known that jacalin binds to the IgA1 hinge, and the hinge models from this work are seen to be extended and fully accessible to permit this interaction to take place (Kerr, 1990). These extended IgA1 hinge structures are consistent with the association of O-linked glycosylation sites and Pro-rich sequences with extended polypeptide structures (Gerken *et al.*, 1989; Shogren *et al.*, 1989; Williamson, 1994).

In combination with the scattering modelling, the homology modelling of the Fab and Fc fragments provide additional insights. The structure predictions for the IgA α-chain showed that the C_H1 and C_H3 domains could be closely modelled by homology with known crystal structures, but this is less so with the C_H2 domains. The C_H2 modelling is affected by different possible disulphide bridge assignments, nonetheless the oligosaccharide at Asn263 is seen to be at the surface of the Fc fragment. The IgA1 scattering model shows that this may be in proximity to the O-linked glycosylated hinge, as this might lead to carbohydrate-carbohydrate interactions that could stabilise the IgA1 structure. However, mutagenesis of the Asn263 site to prevent glycan attachment does not significantly perturb biological activities such as FcaR binding (Mattu *et al.*, 1998) or pIgR binding (Chaung & Morrison, 1997), which argues against such an interaction. The C_H2 modelling also showed that free Cys311 residues existed on the surface of the Fc fragment (Figure 5.3). Their location supports a potential Cys311-Cys471 bridge with the tailpiece, but is in conflict with a proposed Cys311-Cys311 bridge (Prahl *et al.*, 1971; Yang *et al.*, 1979). The tailpiece conformations deduced from the IgA1 scattering curve fits would be consistent with a labile Cys311-Cys471 bridge in monomeric IgA1 (Figure 5.12c). In dimeric IgA1, Cys471 is involved in dimer formation and disulphide bridges with the J chain (Bastian *et al.*, 1992; Atkin *et al.*, 1996). In secretory IgA1, Cys311 forms bridges with secretory component (Fallgreen-Gebauer *et al.*, 1993). The scattering models indicate that the surface of IgA1 is sufficiently accessible for these complexes to be formed.

The present study of IgA1 represents a new extension of automated search methods for scattering curve fits constrained by known crystal structures. Previous approaches had employed systematic translations and/or rotations of domains to generate a sufficient number of models for curve fits (Perkins *et al.*, 1998a, 1998b). For IgA1, this approach gave inconsistencies during the search for possible IgA1 structures (Modelling Method 1; Figure 5.8), as it had assumed that the Fab fragments in IgA1 were in a similar arrangement to that in IgG. Nonetheless the modelling was successfully improved by using the known sequence of the IgA1 hinge to generate a wide range of hinge structures using molecular dynamics. The combination of these structures with the Fab and Fc fragments gave models for which only a small fraction

fitted the data (Figure 5.10). This indicates that constrained scattering curve-fit methods are versatile and can be applied to a wide range of structural modelling studies (Perkins *et al.*, 1998a, 1998b).

Chapter 6

The Structure of MFE-23, an Anti-CEA Single-chain Antibody Fragment, by X-ray Crystallography

6.1. Introduction

Immunoglobulins, or antibodies, are produced by the vertebrate immune system in response to stimulation by foreign molecules, or antigens. An antibody functions by binding specifically to its antigen and then activating effector mechanisms to neutralize the antigens or the invading microorganisms that produced them. Examples of effector functions include the initiation of phagocytic cells and the activation of the complement system. Efficient antibody production is therefore dependent upon the generation of an immensely diverse population of antigen-binding specificities and the initiation of proper effector functions. A primary aim in designing an antibody for performing specific targeting functions, such as MFE-23 targeted against tumours, is the selection of a molecule that has the desired antigen-binding characteristics, and it is for this reason that the remainder of this account will focus on the immunoglobulin antigen-binding site. An insight into the effector functions of immunoglobulins has been given in Chapter 5, in which the complete structure of the human IgA1 molecule was examined.

6.1.1. Overview of antibody structure

The structure and function of antibodies have been reviewed extensively (e.g. Burton & Woof, 1992; Padlan, 1994). Much of the current understanding of immunoglobulins is focussed on two systems; the human immunoglobulins for reason of their obvious clinical interest and those from mice which represent a good model of the human and other mammalian systems. In humans and mice, there are five principal classes of immunoglobulin, IgA, IgD, IgE, IgG and IgM, which were identified by their reactivities to specific antisera. In fact, the DNA coding regions for certain immunoglobulin classes have been duplicated, giving rise to different isotypes for these molecules. Therefore, in humans the full complement of immunoglobulins is IgA1, IgA2, IgD, IgE, IgG1, IgG2, IgG3, IgG4 and IgM, while in mice it is IgA, IgD, IgE, IgG1, IgG2a, IgG2b, IgG3 and IgM. IgG is the predominant class of serum immunoglobulin, and for reason of its relative ease of study more is known about IgG than the other classes and it is typically considered as the prototype immunoglobulin (Figure 6.1). An IgG molecule comprises four polypeptide chains, two identical light chains (κ or λ) of approximately 250 residues and two identical heavy chains of approximately 450 residues, and the four chains are held together by means of

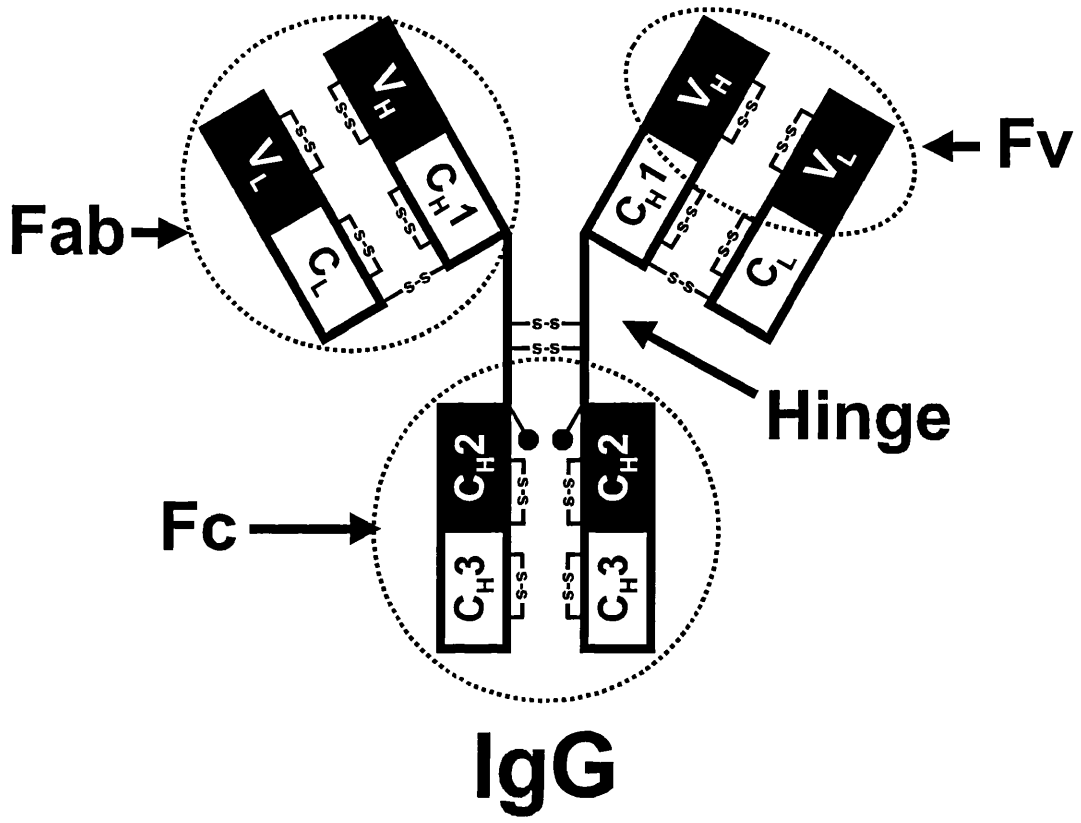


Figure 6.1. Schematic diagram of the twelve domain structure of human IgG1. Each rectangle represents and Ig-fold domain. IgG1 has two heavy chains which each contain a V_H, C_H1, C_H2 and C_H3 domain, and two light chains which each contain a V_L and a C_L domain. The twelve domains are arranged as two four-domain Fab fragments, which are both linked to a four-domain Fc fragment by a hinge region. Disulphide bridges are designated S-S. All domains have the conserved internal disulphide bridge between β-strands B and F. The disulphide bridge that links a heavy chain to a light chain is formed between a cysteine in the hinge region and the C-terminal cysteine on the light chain. The heavy chains are covalently joined by two inter-chain bridges which are formed by two cysteines in the hinge region of each heavy chain, bonding with their corresponding residue on the other heavy chain. The filled circles (●) designate N-linked oligosaccharides. There is an N-linked site on both of the C_H2 domains, and the oligosaccharide located at this position occupies the cavity in the Fc region. A V_H and a V_L domain associate to form an Fv fragment, the smallest antibody fragment to retain a complete antigen-binding site.

disulphide bridges and non-covalent interactions between complementary surfaces. The four chains fold up into a total of 12 Ig-fold domains (Chapter 1). Each light chain contains an N-terminal V-set domain (V_L) and a C1-set domain (C_L). Each heavy chain contains an N-terminal V-set domain (V_H) followed by three C1-set domains (C_H1, C_H2 and C_H3). The C_H1 and C_H2 domains are separated by a long flexible hinge. IgG is classically divided into two Fab fragments and a single Fc fragment according to its proteolytic cleavage with papain, where the papain cleavage site lies between the C_H1 domain and the hinge disulphide bridges. A Fab fragment contains four domains which are the two domains from a light chain and the V_H and C_H1 domains from a heavy chain. The Fc is also a four-domain fragment, containing the C_H2 and C_H3 domains from both heavy chains. A Fab fragment is so-named because it possesses antigen-binding function. The binding site is localized to the Fv fragment that comprises a V_H and a V_L domain.

6.1.2. The immunoglobulin antigen-binding site

Sequence analyses of their V-set domains revealed how the immunoglobulin structure is able to adapt to bind different antigens (Wu & Kabat, 1970). In each V-set domain there are three regions that exhibit sequence hypervariability. It was correctly predicted that these hypervariable regions confer antigen-binding specificities and they are therefore also referred to as complementarity-determining regions (CDRs). In contrast, the regions outside of the hypervariable regions are mostly conserved and are termed the framework regions of the V-set domains. Examination of the DEBA|GFCC'C" V-set fold structure (Figure 1.2; Chapter 1) shows that the first CDR occurs at the loop between β-strands B and C, the second CDR is at the loop between strands C' and C" and the third CDR is at the loop between strands G and F. All three CDRs are therefore located at the N-terminal side of the V-set fold, resulting in their exposure. Also, all three CDRs are attached to strands in the GFCC'C" β-sheet. In the Fv structure, the V_H and the V_L domains are closely associated by means of non-covalent interactions between their GFCC'C" sheets to form a compact structure. This association brings the three CDRs from the two domains into close proximity to each other so that together the six CDRs form a continuous surface that covers an area of approximately 2800 $Å^2$ (Padlan, 1994). Crystal structures of antibody-antigen

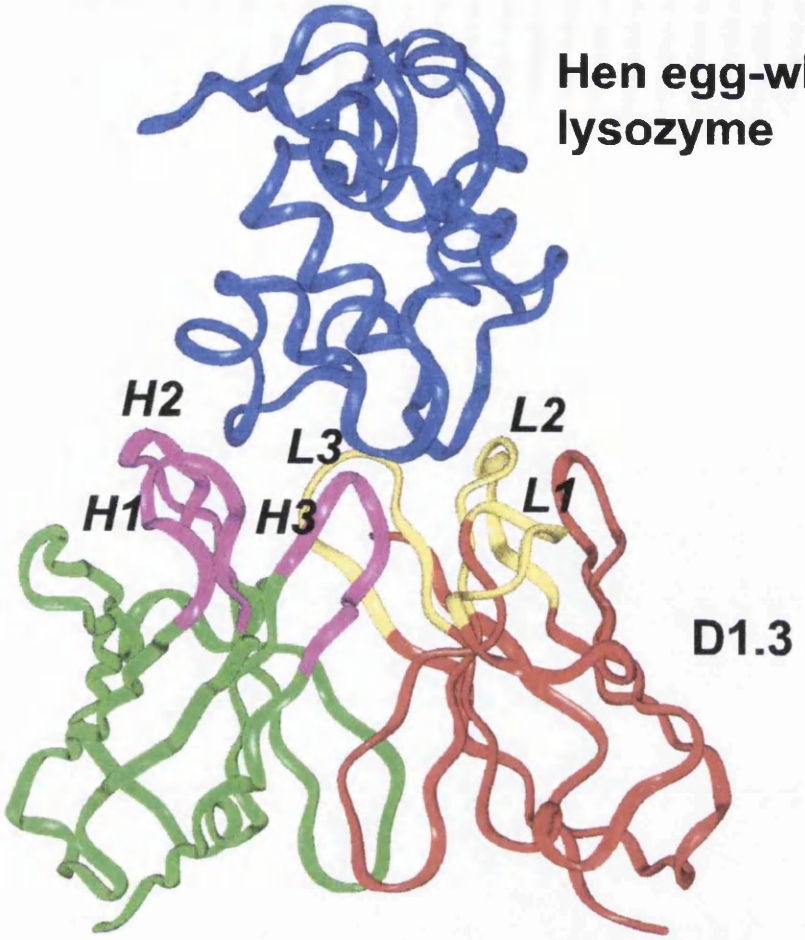


Figure 6.2. The association of the D1.3 antibody Fv fragment with its antigen hen egg-white lysozyme. The coordinates are from PDB entry 1vfb (Bhat *et al.*, 1991). The framework of the V_H domain is shown in green and its CDRs are magenta. The framework of the V_L domain is shown in red and its CDRs are yellow. For clarity, the six CDRs (*H1*, *H2*, *H3*, *L1*, *L2* and *L3*) are labelled.

complexes confirmed that the immunoglobulin antigen-binding site is formed mainly by the CDRs and an example of antibody binding to its antigen is shown in Figure 6.2.

In order to appreciate the diversity of the antigen-binding site, it is useful to consider the complex genetic mechanisms that generate sequence hypervariability and consequently lead to the potential for more than 10^8 different antibody specificities in a single mammalian immune system (Kuby, 1996). The immunoglobulins are encoded by three multigene families, the κ and λ light chain families and the heavy chain family, each of which contains several coding sequences (gene segments) separated by non-coding regions. Functional immunoglobulin genes are formed by the rearrangement and fusion of these gene segments. The κ and λ light chain families contain V, J and C gene segments; the rearranged VJ segments encode the V_L domain and the C segment encodes the C_L domain. The heavy chain family contains V, D, J and C gene segments; the rearranged VDJ gene segments encode the V_H domain and the C segment encodes its C1-set domains, which will be different for each immunoglobulin class or subclass. The rearrangements of these gene segments and other somatic events are responsible for creating antibody diversity. In mice, seven mechanisms have so far been shown to generate diversity, and similar ones are thought to exist in humans. (1) There are multiple copies of V and J segments in the κ and λ light chain families and V, D and J segments in the heavy chain family. (2) Random rearrangements of the different copies of these segments amplifies diversity, e.g. for the 300 to 1000 V_H gene segments, 13 D_H segments and 4 J_H segments in mice there is a minimum of $300 \times 13 \times 4 = 1.6 \times 10^4$ combinations. (3) The joining of V, D and J coding sequences is often flexible, producing alternative amino acids at each coding joint. Further variations may occur at the coding joints due to the enzymatic addition of nucleotides during the recombination of gene segments. (4) Nucleotide addition by enzymes repairing single-stranded DNA at the ends of gene segments. (5) Addition of up to 15 nucleotides by the terminal deoxynucleotide transferase enzyme to the VD and DJ joints of heavy chains. Significantly, the VJ coding joint in the V_L domain and the VD and DJ coding joints in the V_H domain all fall within CDR-3, and consequently particular sequence diversity is present in these loops (Kabat *et al*, 1991). (6) In the V-set domain coding regions, somatic hypermutation of nucleotides occurs at a rate one million times greater than the

spontaneous mutation rate in other genes. Somatic mutations are largely random over the whole V_H or V_L domain sequences. However the selection of antibody-producing B-cells is dependent upon their antigen-binding specificity so they mainly cluster within the CDRs. In particular, as the immune response proceeds, hypermutations tend to be located in CDR-1 and CDR-2 of both domains (Berek & Milstein, 1987). (7) Finally, random combinations of light and heavy chains considerably increase the diversity of the binding-site.

A recent review of known antibody structures highlights the effects of these somatic events (Rees *et al.*, 1996). As expected, CDR-*H3* displays the greatest variability in length, containing between 4 and 27 residues compared to *H1* which has 5 or 7 residues, *H2* which has 9 to 12 residues, *L1* which has 10 to 17 residues, *L2* which has 7 residues and *L3* which has 7 to 11 residues. It is apparent from X-ray crystal structures that, despite the diversity of the binding site, certain common features are maintained. The association of the V_H and the V_L domains in the Fv arranges the CDRs so that *H3* and *L3* are in the centre of the binding-site, surrounded by the four other CDRs (Figure 6.2). Although all six CDRs have been observed to make contact with antigen in crystal structures, they do so to varying degrees. Unsurprisingly, given the greater sequence variability of CDR-*H3* and CDR-*L3* and their position at the centre of the binding-site, these two CDRs along with CDR-*H2* play a prominent role in binding antigen. On average, these three CDRs account for more than 70% of the interactions between antibody and antigen (Wilson & Stanfield, 1993). In some structures, the occasional framework residue is found to interact with antigen (e.g. Amit *et al.*, 1986; Sheriff *et al.*, 1987). The region of the immunoglobulin binding-site that contacts antigen is termed the paratope, while the region of the antigen bound by the antibody is termed the epitope, and the paratope and the epitope are observed to possess complementary surfaces in crystal structures. It is for this reason that, in antibodies which bind small haptens, the surface of the binding-site commonly has a deep pocket, while those which bind "linear" antigens such as peptides or polynucleotides typically possess a groove in the binding-site, and antibodies that bind globular proteins typically have large flat binding-sites. In the latter case, complementary electrostatic areas are especially prominent in determining specificity, where it is generally found that

hydrophobic patches on the paratope interact with hydrophobic patches on the epitope, polar atoms in the paratope interact with oppositely-charged atoms on the epitope, and hydrogen bonds are formed between proton donors and acceptors (Braden *et al.*, 1998).

The immunoglobulin binding-site appears to be adapted, not just to maximize the energies of interactions between the antibody and antigen, but also to minimize the loss of conformational entropy upon complexation. Thus analyses of antibody-antigen complex structures and CDR sequences reveal a preference for certain types of residues to make contact with antigen (Kabat, 1977; Janin & Chothia, 1990; Padlan, 1990; Mian *et al.*, 1991; Padlan, 1994; Lea & Stuart, 1995). There is a high incidence of surface-exposed aromatic amino acids, particularly tyrosine residues, in the CDRs relative to the framework regions and they are frequently observed to make contact with antigen. Reasons for the utilization of aromatic residues include: their large sizes can contribute significantly to the hydrophobic effect; their large polarizabilities contribute to van der Waals interactions; they are able to form hydrogen bonds through their aromatic rings and through polar atoms in their sidechains if they possess them; and their sidechains have relatively few degrees of freedom due to the rigid nature of aromatic rings which results in a smaller loss of conformational entropy. Polar amino acids, notably histidine and asparagine, are also common binding-site residues. However, it is unusual to observe apolar aliphatic valine, isoleucine and leucine residues in the binding-site, and this is attributed to the many rotational degrees of freedom in their sidechains and their ability to contribute to antigen-binding only through the hydrophobic effect and van der Waals interactions.

It has also been observed that certain sequence positions within the CDRs are preferentially utilized for contacting antigen (Padlan, 1994; Padlan *et al.*, 1995). These so-called specificity-determining residues are characterized by an even higher degree of sequence variability than that observed for the other residues within the hypervariable regions. SDRs are observed in all of the CDRs, but they are particularly abundant in CDR-*H3*. A possible explanation of SDRs is that they occur at exposed positions within the CDRs and, importantly, at positions that are not essential for determining the mainchain conformations of the CDRs. This explanation has particular relevance when

it is considered that for five of the CDRs (*H1*, *H2*, *L1*, *L2* and *L3*) only a limited number of mainchain conformations have been observed in known antibody structures (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Al-Lazikani *et al.*, 1997). These are termed canonical structures, and their existence implies that the generation of antibody diversity is not a wholly random process. Each canonical structure can be identified from its sequence by the residues that determine its mainchain conformation, either through hydrogen bonds, sidechain packing or preferred mainchain torsion angles. The known canonical structures have gained an important application in the modelling of antibody-binding sites, even though it is not possible to achieve a complete definition of the binding site because CDR-*H3* has a highly variable structure.

An example of a specific antibody-antigen complex is that formed between the antibody D1.3 and its antigen hen egg-white lysozyme (HEL; Bhat *et al.*, 1991) (Figure 6.2). This association represents a general model of antibody binding to a globular protein antigen, and a recent review characterizes some of its important features and highlights the intricacies underlying the general features of the immunoglobulin binding-site (Braden *et al.*, 1998). Comparison of the free D1.3 Fv structure (Bhat *et al.*, 1990) with its complex with HEL (Bhat *et al.*, 1991) shows that the solvent accessible surface area buried upon complexation is approximately 1234 $Å^2$ in total, and for both D1.3 and HEL approximately 50% of the buried area corresponds to hydrophilic atoms. Interactions occur between 17 D1.3 residues and 16 HEL residues and both sidechains and mainchains are involved. These interactions include 17 hydrogen bonds, many van der Waals interactions but no salt bridges. Upon complexation, there is a slight reorientation of the V_H and V_L domains, and the many contacts made between the antigen and CDR-*H3* significantly restrict the mobility of this loop. However, the sidechains of residues in the binding site have very similar conformations in the free and complexed Fv structures. Water molecules make an important contribution to the association. There are about 50 water molecules located around the antibody-antigen interface, well-ordered water molecules contribute a total of 10 hydrogen bonds to the stability of the complex, and water molecules fill cavities that are formed by the paratope and the epitope not having exactly complementary surfaces and which could have a destabilizing effect if left unfilled. However, the

immobilization of water molecules at the antibody-antigen interface could result in a decrease in the entropy of the system and therefore subtract from the total binding energy.

Mutation analyses of contact residues in both antibody and antigen reveal that only a small number of contact residues make a significant contribution to the energetics of binding. This has lead to the hypothesis of "functional" or "energetic" paratope and epitope within the full interaction sites (Braden *et al.*, 1998). Five of 17 D1.3 contact residues and 4 of the 16 HEL contact residues make a major contribution to the interaction and the important residues in the two opposite protein surfaces are juxtaposed to each other. An implication of the "functional" paratope and epitope is that the binding site is tolerant to mutations, which may be important for allowing an antibody to cope with mutations in its antigen. Mutation analyses also indicate that hydrogen bonds formed between the antibody and antigen do not necessarily make a significant contribution to the binding energy, instead they merely act to neutralize the energetic contributions from the hydrogen bonds between protein and water molecules that they replace. Mutations that affect the complementarity of the antibody-antigen interacting surfaces appear to be corrected by water molecules.

The immunoglobulin antigen-binding site represents an elegant solution to the complex problem faced by immunoglobulins, that is having to adapt to recognise whatever antigens the immune system is exposed to. Its diversity is generated by a series of somatic events that produce sequence variations within the six CDRs of the antibody. Although there is a degree of structural conservation to the effects of these sequence variations, as the example of the D1.3-HEL complex illustrates, the physical realization of these conserved structural features in the binding of an antibody to a globular protein involves an intricate series of interactions between both protein and water molecules, all of which have differing degrees of effect on complex formation.

6.1.3. The design and production of MFE-23 for tumour targeting

As more becomes known about the immunoglobulin antigen-binding site, its adaptation for clinical targeting applications is typically becoming a rational design

process (for review, see Chester & Hawkins, 1995). Accordingly, the design of MFE-23, a single-chain Fv fragment (scFv), for targeting colon tumours was achieved after several important design issues had been addressed. Foremost, it was required that MFE-23 has binding specificity for the tumour marker carcinoembryonic antigen (CEA). For tumour targeting, there is evidence that antibodies with high affinity for antigen confer an advantage over those with lower affinity (Schlom *et al.*, 1992). A phage-display library was therefore used to produce MFE-23 because this enabled a vast amount of antibodies to screened for high-affinity binding to CEA. The tumour-penetrating ability of an antibody has a major impact on its clinical efficiency and it has been demonstrated that the smallest antigen-binding fragment, the Fv, penetrates tumours more rapidly and more deeply than Fab fragments or whole IgG (Yokota *et al.*, 1992). Another advantage of using the Fv fragment for clinical targeting is that for antibodies of mouse origin the Fv, for reason of its small size, will be less immunogenic than other antibody molecules when injected into human patients. It is desired that a human immune response is avoided because it may prevent tumour localization, especially with repeated use of an antibody molecule, and lead to toxic side-effects. It is possible to improve the targeting efficiency of an Fv molecule by synthesizing it as a single-chain molecule in which the V_H and V_L domain are joined by a flexible polypeptide linker. The reason for this is that unlike the two polypeptide chains in a Fab fragment, which associate non-covalently through four domains and covalently by means of a disulphide bridge, the two domains in the Fv fragment can dissociate and diffuse apart from each other. A polypeptide linker between the two domains serves to prevent their separation (Huston *et al.*, 1988). Finally, the characteristic biodistribution of scFvs correspond to a rapid clearance from the blood via the kidneys and thus scFvs offer a means for early imaging of tumours.

Figure 6.3 illustrates the procedure used to produce MFE-23 (Chester *et al.*, 1994). A cDNA library was produced from mRNA isolated from the lymphocytes of mice immunized with CEA. Immunoglobulin V_L and V_H domain sequences were then amplified separately by the polymerase chain reaction (PCR). The relative sequence invariability of the N- and C-terminal sequences of the V_H and the V_L domains facilitates the design of a set of PCR primers to specifically amplify V-set domains, and the PCR

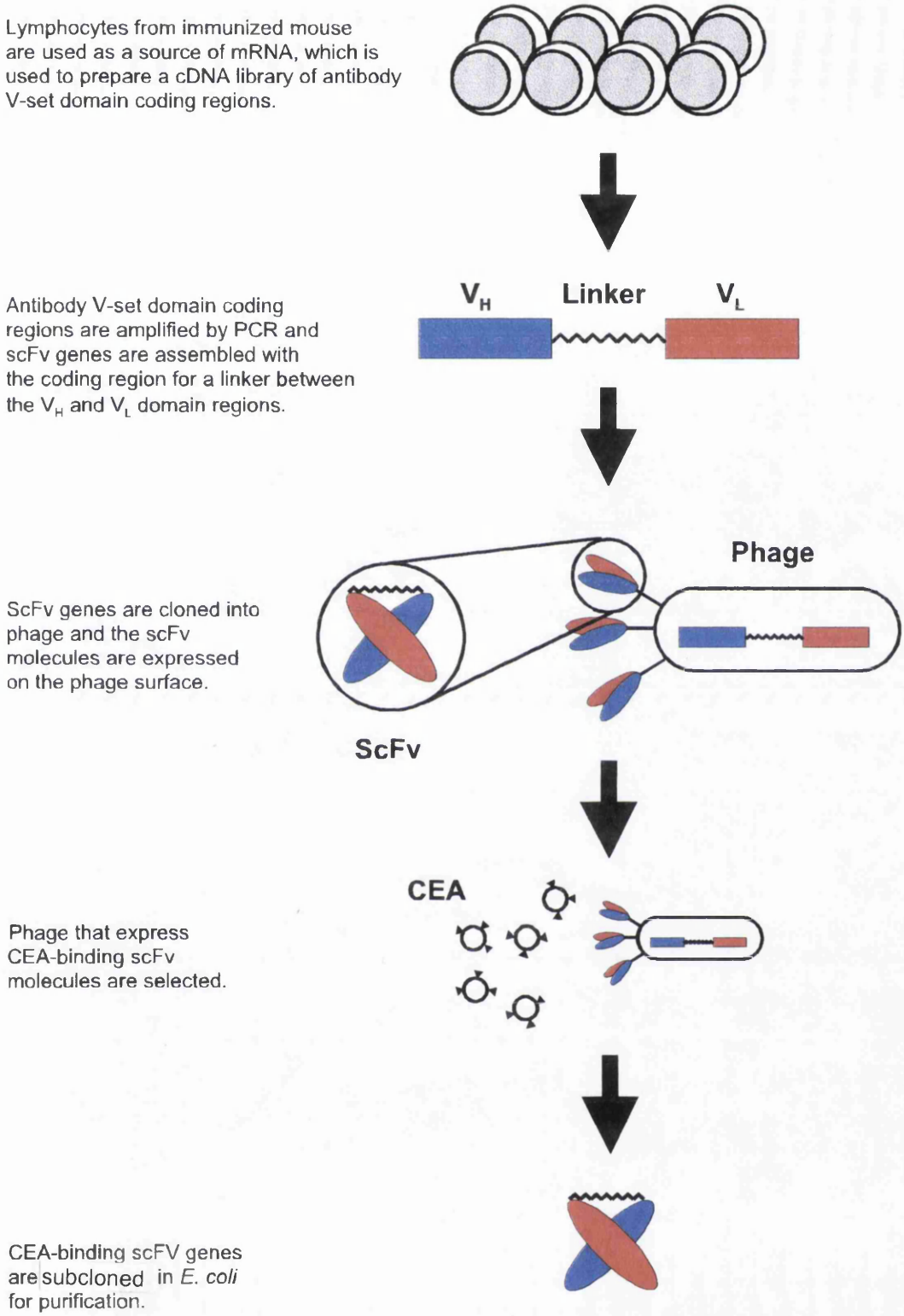


Figure 6.3. Production of MFE-23 from a phage display library (Adapted from Chester *et al.*, 1994).

reaction also allows the introduction of restriction enzyme cleavage sites into the sequences for cloning procedures (Orlandi *et al.*, 1989). V_H and V_L domain coding sequences were assembled to encode scFv molecules, each with an N-terminal V_H domain joined to a V_L domain by a $(Gly_4\text{-}Ser)_3$ linker peptide (Huston *et al.*, 1988), and these scFv genes were cloned into bacteriophage vectors. The protein product of each bacteriophage vector is an scFv molecule fused to the N-terminal region of the phage gene III protein, which signals its expression on the surface of the phage without the loss of its antigen-binding function (McCafferty *et al.*, 1990). A random V_H-V_L combinatorial phage library of 10^7 members was screened for specificity to CEA using biotinylated CEA and a streptavidin-capture system. After successive rounds of decreasing biotinylated CEA concentrations, MFE-23 was selected for its high affinity binding to CEA. The use of phage expression libraries to produce antibody fragments has inherent theoretical advantages over other techniques. The random combinatorial approach of phage libraries mimics the random association of light and heavy chains and could give rise to antigen-binding sites not present in the repertoire of the immunized mice. Additionally, the ease of functionality screening and subcloning serve to increase the likelihood of isolating a molecule with the desired characteristics.

MFE-23 was subcloned into a pUC119 plasmid for expression as a soluble scFv in *E. coli*. The vector encoded an N-terminal *pelB* signal sequence, for secretion of the MFE-23 into the bacterial periplasm (Pluckthun, 1990). For the initial characterization of MFE-23, it was cloned with an 11-residue c-myc peptide tag at its C-terminus, which facilitated its identification during purification and localization studies in tissues (Chester *et al.*, 1994; Verhaar *et al.*, 1995). Subsequently, it has been cloned with a six-histidine tag at its C-terminus for easy purification by immobilized metal affinity chromatography (Casey *et al.*, 1995).

Against a range of normal human tissues, MFE-23 displays only a weak reactivity and that is towards normal colon (Chester *et al.*, 1994). MFE-23 has been shown to have better tumour targeting properties than an scFv derived from A5B7, a clinically important anti-CEA monoclonal antibody produced by hybridoma technology (Verhaar *et al.*, 1995). This may be due to the difference in CEA-affinities between the

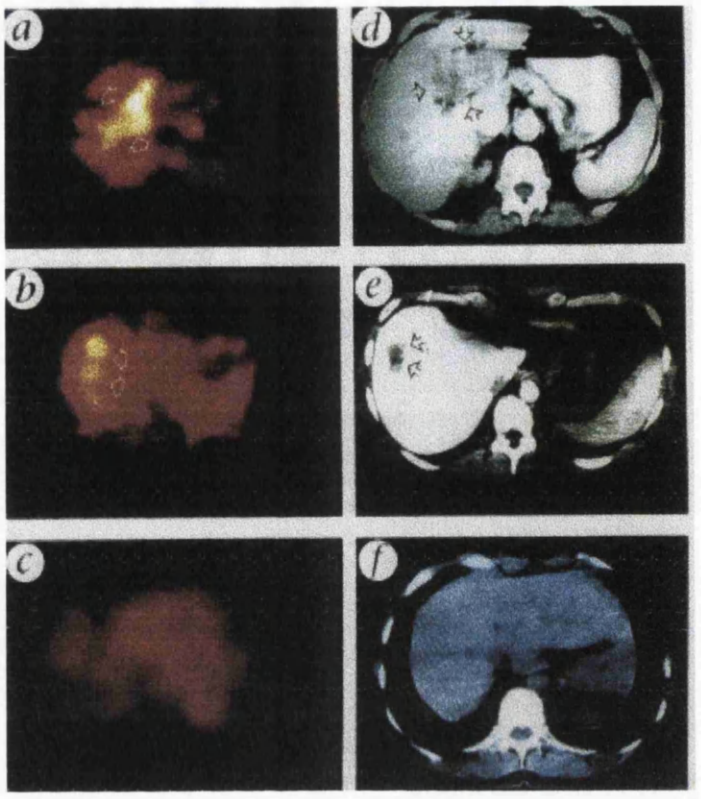


Figure 6.4. Tumour localization of MFE-23. (*a*) and (*b*) Single-photon emission computerized tomography (SPECT) gamma camera images show localization of ^{123}I-labelled MFE-23 to liver metastases in two patients with colon carcinoma 1 and 4 h after injection respectively. (*d*) and (*e*) Corresponding X-ray CT illustrating anatomy at the same level and confirming the presence of tumour. In (*e*) the metastasis was only visible after intra-arterial injection of contrast material directed to reach the hepatic portal vein (CT portography). Tumour deposits are indicated with arrows. (*c*) For comparison, SPECT images taken 4 h after injection are shown in a patient with no tumour. The uniform distribution of antibody in normal tissues is demonstrated. (*f*) The corresponding CT image confirms the absence of tumour (Adapted from Begent *et al.*, 1996).

two molecules (MFE-23 has a Kd of 2.5 nM compared to a Kd of 25 nM for the A5B7 Fab), which are a direct result of the methods used for their production (Verhaar *et al.*, 1995). MFE-23 labelled with iodine-123 has been demonstrated to be localized at the site of tumours in patients with liver metastases by direct visualization using gamma camera imaging (Figure 6.4; Begent *et al.*, 1996). Even though in this case MFE-23 labelled with ^{123}I retains CEA-binding and has the desired localization in patients, it is recognised that there is a potential problem with this commonly-used technique. Radioiodination is sited at surface-exposed tyrosines, however the majority of such residues in antibodies occur within the CDRs and are instrumental in antigen-binding. Iodination can therefore lead to a decrease in the affinity of antigen-binding (Nikula *et al.*, 1995). To overcome this problem, a cysteine residue has been incorporated into the C-terminal tag of MFE-23 (Verhaar *et al.*, 1996). The cysteine presents a free thiol group for radiolabelling with Technetium-99m. This method has several advantages over radioiodination, namely radiolabelling occurs distal to the binding site, the gamma energy emission of ^{99m}Tc is relatively low compared to ^{131}I which is beneficial for patient trials, and the short half-life of ^{99m}Tc (6 hours compared to 8 days for ^{131}I) is on a similar scale to the time taken for MFE-23 to clear from circulation in patients and is thus suitable for early imaging of tumours.

The importance of MFE-23 is not restricted to its tumour detection properties. It has already been noted that genetic-engineering techniques can be applied to fuse various polypeptide tags to the C-terminal of MFE-23. Utilizing a similar strategy, more ambitious employments of its tumour-targeting properties are being developed. Antibody-directed enzyme prodrug therarpy (ADEPT) is potentially a very important means for treating tumours (Bagshawe, 1989). The principle of ADEPT is that a tumour-targeting antibody is coupled to an enzyme that will activate a cytotoxic agent from an inactive prodrug specifically at the site of a tumour. For this purpose, a recombinant form of MFE-23 fused to carboxypeptidase G2 has been produced (Michael *et al.*, 1996). The carboxypeptidase cleaves a glutamate residue from an inert benzoic acid mustard molecule to produce the active drug form (Chester & Hawkins, 1995). ADEPT clinical trials have shown encouraging responses in patients with colorectal cancer (Bagshawe *et al.*, 1995).

Another potential means for treating tumours is targeted gene therapy. Genetic engineering techniques also enable scFvs to be inserted into retroviral envelope proteins for altering the target cells of the virus (Russell *et al.*, 1993). Accordingly, MFE-23 has been incorporated into the envelope of Moloney murine leukaemia virus for targeting this retrovirus to human colon cancer cells (Konishi *et al.*, 1998). The aim of this approach is to use the targeted retrovirus to introduce suicide genes into tumour cells.

6.1.4. Structure-based approaches for improving MFE-23

It is always desired that existing clinical targeting strategies be improved. Aside from the fusion of antigen-binding function with various novel "effector" functions described above, a worthwhile approach for improvement is the modification of the actual MFE-23 Fv molecule. There are two major ways to modify an Fv molecule, either based on altering the antigen-binding properties, or improving its "humanisation". In "humanization", the V_H and V_L domain framework residues that are not essential for the structure of the binding-site are converted to the equivalent residues from a human antibody with the aim of making the molecule less immunogenic in patients, therefore increasing its lifetime. Two methods have been developed for "humanising" an antibody. One method is "CDR-grafting", in which genetic engineering techniques are used to insert the CDR sequences from the antibody of interest (typically of mouse origin) into the corresponding positions on a human Fv fragment (Jones *et al.*, 1986; Reichmann *et al.*, 1988). The second method is "resurfacing", in which the Fv fragment of interest is masked by converting its surface-accessible framework residues to the equivalent residues from a suitable human Fv structure (Padlan, 1991; Roguska *et al.*, 1994).

In order to carry out useful modifications to the Fv molecule, sequence data alone may fail to identify important residues and it is therefore important to obtain information on the three-dimensional structure. There are several reasons for this. The six CDRs form a continuous surface area and their close proximity means that mutational changes to one CDR may be propagated to adjacent CDRs. There are contacts between the CDRs and framework regions, therefore the alterations of certain CDR residues may induce conformational changes in the framework and vice-versa.

When mutating binding-site residues, it is desirable to avoid the residues that are important for determining the conformations of the CDR loops, but to focus instead on the residues whose sidechains are freely-exposed and thus able to make contact with antigen. Because of the large number of antibody structures in the Brookhaven protein databank, homology modelling of antibodies is more advanced than for any other protein (for recent reviews see, Rees *et al.*, 1996 and Padlan, 1994). In particular, the identification of a limited number of canonical structures for CDRs-*H1*, *H2*, *L1*, *L2* and *L3* in the known structures has greatly improved the reliability of these models, since it is now possible to predict these canonical structures from their sequences (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Al-Lazikani *et al.*, 1997). However, there are some significant limitations to homology models of antibodies. Firstly, for the five CDRs that possess canonical structures, not all sequences can be classified using the rules derived so far, suggesting either that the rules are incomplete as they stand or that there are as yet unidentified canonical structures. Secondly, CDR-*H3* is too variable in its sequence to exhibit canonical structures. Although some progress has been made in identifying families of CDR-*H3* structures (Rees *et al.*, 1996), no sequence rules have been determined that enable these loop structures to be predicted with confidence. Thirdly, it is possible that certain canonical structures are incompatible with each other, leading to conformational changes in one or more loops. Such a phenomenon can not be discounted from the structures that currently exist. Fourthly, the association of the V_H and the V_L domains varies between different Fv structures (Padlan, 1994), which will affect the relative orientations of the CDRs from the two domains. Although analyses of known structures can be used to identify conserved interface residues and hence select reasonable models for the association of the Fv module (Chothia *et al.*, 1985; Rees *et al.*, 1996), the observation of domain rearrangements on binding to antigen (e.g. Braden *et al.*, 1998) adds an element of doubt to any such models.

The molecular design of MFE-23 as an anti-tumour agent would be better understood by knowledge of its crystal structure, and would facilitate the creation of improved targeting molecules. The MFE-23 crystal structure presented here is the first one for an anti-tumour scFv molecule derived from phage technology, and is of great interest for reason of its use in patient trials (Begent *et al.*, 1996). Previously, loop

structures for five of the six antigen-binding loops *H1*, *H2*, *L1*, *L2* and *L3* could be predicted by database approaches (Chothia *et al.*, 1989; Pedersen & Rees, 1993; Read *et al.*, 1995; Al-Lazikani *et al.*, 1997). Past experience of loop predictions show however that the sixth loop *H3* is less straightforward to predict in the absence of a crystal structure. While well over 100 crystal structures have been reported for Fab fragments (Figure 6.5; Wilson & Stanfield, 1994), only five previous crystal structures correspond to the scFv molecules Se155-4, NC10, L5MK16, CC49/212 and C219 (Zdanov *et al.*, 1994; Kortt *et al.*, 1994; Perisic *et al.*, 1994; Raag & Whitlow, 1995; Hoedemaeker *et al.*, 1997). The interest of the MFE-23 scFv structure lies in its antigenic specificity for the immunoglobulin folds of CEA, as this raised the theoretical possibility that its antigen-binding loops may form lattice contacts with its own immunoglobulin fold, unlike those seen with the other scFv molecules. This actually turned out to be the case, and provided insight into how MFE-23 may interact with CEA. Evidence for the significance of the observed lattice contacts was provided using mutagenesis data on the MFE-23 interaction with CEA (Read *et al.*, 1995) and from prediction studies (Padlan *et al.*, 1995). Crystal structures have been reported for the murine Fab fragment A5B7 and a humanised form of this, both of which also bind to human CEA, albeit more weakly with K_d values of 25 nM and 38 nM respectively (Verhaar *et al.*, 1995; Banfield *et al.*, 1996, 1997). A comparison of their antigen-binding loops with those in MFE-23 was used to verify the significance of the MFE-23 combining sites for the same antigenic target. The lifetime of murine MFE-23 in human circulation would be improved by the design of a less immunogenic scFv structure. The MFE-23 crystal structure permitted the identification of closely related human Fab fragment, and this enabled the "humanisation" of MFE-23 to be evaluated.

6.2. Materials and methods

6.2.1. Cloning, expression and purification of MFE-23

The cloning, expression and purification of the murine scFv molecule MFE-23 is described by Verhaar *et al.* (1995). MFE-23 was purified by Dr P. A. Keep from the Clinical Oncology Department at the Royal Free and University College Medical School. MFE-23 contained a *pelB* leader peptide at its N-terminus to permit its secretion from bacterial cells (after which this leader is cleaved), and a C-terminal 11-

Figure 6.5. (Overleaf) Sequence numbering and secondary structure alignment of MFE-23 with the crystal structures of heavy and light chains from Fab fragments. The Fab fragments of murine D1.3, D44.1, and J539 represent the three V_H classes 1, 2 and 3, and are shown with the known human Fab crystal structures (New, TR1.9, 3D6, HuIgM, Kol and Hil). The MFE-23 sequence numbering and its Kabat numbering are shown, together with the Chothia structural and Kabat hypervariable definitions of the antigen-binding loops *H1-H3* and *L1-L3* and the labelling of 10 β-strands A to G (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Kabat *et al.*, 1991). The six MFE-23 antigen-binding structural loops, the $(Gly_4Ser)_3$ linker sequence between the V_H and the V_L domains, and the C-terminal myc-tag sequence are underlined. The 29 residues in MFE-23 to be converted into those of TR1.9 in order to humanise the MFE-23 structure are denoted by * if they are chemically dissimilar and + if they are similar. Residues in bold correspond to consensus β-strand residues used to superimpose the V_H and V_L domains for structural comparisons (Table 6.7). The V_H domains are labelled with their framework structural class 1, 2 or 3 (Saul & Poljak, 1993). The secondary structure elements identified by the DSSP program are labelled as follows: E, β-strand; B, single residue β-ladders; T, turns; S, bends; G, 3_{10}-helix.

Figure 6.5. Sequence numbering and secondary structure alignment of MFE-23 with the crystal structures of heavy and light chains from Fab fragments (legend on page 255).

residue myc-tag for identification during purification. A pUC119 plasmid containing this construct was used to transform *E.coli* cells. MFE-23 was purified from cultures in 2×TY media containing ampicillin (100 μg/ml) and glucose (0.1% w/v), to which 1 mM isopropyl β-D thiogalactoside was added to induce expression. From the culture, a filtrate was collected by centrifugation, filtered (0.2 μm filter, Nalgene), concentrated 5-fold using a spiral cartridge with a 10 kDa cut-off (S1Y10, Amicon), and dialyzed against Dulbecco's phosphate-buffered saline (PBS;137 mM NaCl, 0.5 mM $MgCl_2$, 2.7 mM KCl, 8.1 mM Na_2HPO_4, 1.5 mM KH_2PO_4, pH 7.4) containing 0.02% NaN_3 w/w. After refiltration, MFE-23 was purified by affinity chromatography using CEA covalently attached to Sepharose-4B gel (Pharmacia), eluting it with 0.05 M diethylamine (pH 11), and buffering immediately with 1 M phosphate (pH 7.5). The MFE-23 fractions were pooled, dialyzed overnight against PBS and 0.02% NaN_3, concentrated by ultrafiltration (PM10 membrane, Amicon), subjected to gel filtration on Sephacryl-S100 (Pharmacia), and concentrated (YM10 membrane, Amicon). Samples used for crystallization were dialyzed against 20 mM Tris-HCl (pH 6.5). The concentration of MFE-23 was determined from an absorption coefficient of 20.0 calculated from its sequence (1%, 280 nm, 1 cm path; Perkins, 1986), which is 40% higher than the value of 14.3 used by Verhaar *et al.* (1995).

6.2.2. Crystallization and data collection

Crystals of MFE-23 were grown by the hanging-drop vapour diffusion method at 18°C by Mr Jeremy Thornton (Clinical Oncology Department, Royal Free and University College Medical School) and Dr Maninder Sohi (Dr Brian Sutton's group, Randall Institute, King's College, London). Briefly, protein (2 mg/ml) was mixed 1:1 with 100 mM Tris-HCl (pH 6.5) buffer containing 45% saturated ammonium sulfate. A 10 μl droplet of this mixture was equilibrated against 0.5 ml 100 mM Tris-HCl (pH 6.5) in 45% saturated ammonium sulphate. Diffraction data was collected by Dr Tommy Wan (Dr Brian Sutton's group, Randall Institute, King's College, London). Crystals were mounted in quartz glass capillary tubes, and their diffraction patterns were recorded using a precession camera and an R-AXIS-IIC image plate mounted on an RU200 rotating anode X-ray source. Data from a single crystal was used to solve the MFE-23 structure (Table 6.1). Autoindexing and integration of reflections were

Table 6.1. Summary of data collection and structure refinement statistics for MFE-23

Data collection (Table 6.3)	
Resolution	15.0-2.4 Å
No. of unique reflections	11,537
R_{merge} [1]	8.2%
Redundancy[2]	5.3
Completeness (15.0-2.4 Å)[3]	99.6%
No. of crystals	1
Structure refinement (Table 6.4)	
Resolution	8.0-2.4 Å
No. of unique reflections	11,204
No. of protein atoms with occupancy of 1	1,697
After simulated annealing: R_{factor} [4]	25.5%
R_{free}	29.9%
No. of unique reflections used for R_{free} [5]	909
Final R_{factor}	19.0%

[1] $R_{merge} = 100 \times \Sigma_{hkl} \Sigma_n | \langle I \rangle - I_n | / \Sigma_{hkl} \Sigma_n I_n$ which is summed over all reflections, where $\langle I \rangle$ is the mean intensity of the reflection *hkl*, and I_n is the intensity of *n* observations of a reflection *hkl*

[2] Redundancy = number of measurements/number of independent reflections

[3] Completeness = 100 × number of independent reflections measured/theoretical maximum number

[4] $R_{factor} = 100 \times \Sigma | F_o - F_c | / \Sigma F_o$ where F_o and F_c are the observed and calculated structure factor amplitudes within the set of reflections used for refinement.

[5] $R_{free} = 100 \times \Sigma | F_o - F_c | / \Sigma F_o$ calculated for a randomly selected 8% set of structure factors throughout the resolution range and not used in refinement

performed using DENZO to 2.4 Å resolution (Otwinowski, 1993). Using the CCP4 suite of programs (Collaborative Computational Project Number 4, 1994), the initial processing of the data with SCALA and AGROVATA showed that the crystal belonged to a trigonal space group with unit cell dimensions a = 61.70 Å, b = 61.70 Å, c = 128.00 Å, $\alpha = 90°$, $\beta = 90°$, $\gamma = 120°$. Calculation of the Matthews constant (V_m = unit cell volume / n × molecular weight; where n is the number of protein molecules in the unit cell) indicated that six MFE-23 molecules (V_m = 2.24 Å^3/Da) best satisfied the protein content normally observed for protein crystals (Matthews, 1968). The solvent content of the MFE-23 crystal was estimated as 43% of the unit cell volume.

6.2.3. Crystal space group and structure determination by molecular replacement

Systematic absences in the data set along *00l* indicated the presence of a screw axis in the crystal lattice. The data was processed in the lowest trigonal space group P3, and molecular replacement was used to determine which of the two alternative screw axes 3_1 or 3_2 was correct. Using the AMORE program (Navaza, 1994), the Fv fragment from a murine IgA Fab molecule J539, initially solved at 4.5 Å and subsequently refined at 1.95Å (Suh *et al.*, 1986; accession code: 2fbj), was used in cross-rotation function and translation function searches. The J539 search model contained residues equivalent in position to the first 120 residues of the MFE-23 V_H domain and the first 106 residues of the MFE-23 V_L domain. The J539 and MFE-23 sequences were identical at 59/120 positions in the V_H domains, and at 84/106 positions in the V_L domains (Figure 6.5). Cross-rotation function and translation function searches were carried out over the resolution range 15.0 to 3.5Å. The cross-rotation function was performed on the data processed in space group P3 using an integration radius of 24 Å to identify two clear solutions (Table 6.2). The two peaks for space group P3 was consistent with the estimate of six molecules per unit cell. The two best solutions from the rotation search were used to perform translation searches in the space groups P3, $P3_1$ and $P3_2$ to test alternative trigonal rotation axes (Table 6.2). Each one produced a single solution for translation searches in the space group $P3_2$ but failed to produce single solutions in the space groups P3 and $P3_1$ and this demonstrated that the crystal system has a 3_2 screw axis. Data reduction showed that the higher symmetry space group $P3_221$ with only one molecule in the asymmetric unit was the correct space group. Data was processed using

Table 6.2. Molecular replacement searches for MFE-23

	Rotation Search[1]					
	Solution 1		Solution 2		Solution 3[2]	
P3	15.8		15.7		10.0	
	Translation Search[1]					
	Solution 1a	Solution 1b	Solution 2a	Solution 2b	Solution 3a	Solution 3b
P3	14.9	14.3	14.8	14.3	10.1	9.3
$P3_1$	15.5	15.2	15.3	15.3	10.1	9.8
$P3_2$	21.6	-	21.9	-	10.5	10.1

[1] The correlation coefficients from the AMORE program are shown.

[2] Solution 3 is shown for comparison with the two correct solutions.

Table 6.3. Data reduction statistics for MFE-23 using AGROVATA

Resolution (Å)	Observed reflections	Unique reflections	R_{merge} (%)[1]	$I/\sigma(I)$
15.00-6.84	2424	510	7.1	3.8
6.84-5.11	3668	739	6.6	6.4
5.11-4.26	4588	901	7.3	6.0
4.26-3.72	5348	1039	7.2	6.7
3.72-3.35	6035	1153	7.2	7.3
3.35-3.07	6618	1257	8.6	6.3
3.07-2.85	7340	1361	9.8	5.7
2.85-2.67	7864	1418	12.2	5.2
2.67-2.53	8573	1546	15.4	4.5
2.53-2.40	8292	1613	18.9	3.2
Overall	60750	11537	8.2	3.2

[1] $R_{merge} = 100 \times \Sigma_{hkl} \Sigma_n | \langle I \rangle - I_n | / \Sigma_{hkl} \Sigma_n I_n$ which is summed over all reflections, where $\langle I \rangle$ is the mean intensity of the reflection *hkl*, and I_n is the intensity of *n* observations of a reflection *hkl*

11,537 unique reflections in the resolution range 15 to 2.4 Å with an overall R_{merge} of 8.2%, and an average multiplicity of 5.3 (Table 6.3). The molecular replacement analysis was repeated with space group $P3_221$, and a single solution in the rotation and translation searches confirmed that this was the correct space group. The rigid-body refinement function of AMORE was used for data between 8 and 2.8Å to refine the positional parameters of the molecular replacement solution and generate the initial model with 1724 atoms. The *R*-factor was 51.3% for data between 8.0 and 2.4 Å with $F/\sigma(F) \geq 2$.

6.2.4. Crystallographic model building and refinement

The initial model was manually rebuilt using O and refined using X-PLOR version 3.1 (Jones *et al.*, 1991; Brünger, 1992a). The crystal packing showed that pairs of MFE-23 molecules were closely associated by the two-fold symmetry axis (Figure 6.6). The 3_2 trigonal screw axis produces infinite helices of scFv pairs with large solvent pores between the helices. In the rigid-body refinement protocol of X-PLOR, only reflections with $F/\sigma(F) \geq 2$ were used. The model was treated as separate V_H and V_L domains (H and L residues respectively). Atoms in the putative antigen-binding loop residues H26 to H35, H50 to H59, H97 to H107, L24 to L33, L49 to L55, and L88 to L96 were set to zero occupancies, and no atoms for the linker or the myc-tag were included. After rigid-body refinement, $2F_o$-F_c and F_o-F_c electron density maps contoured at 1σ and 3σ respectively were visualized in O. Atoms were set to zero occupancies if they did not correspond to atoms in MFE-23 or did not fit into the electron density, a total of 630 in all. The model contained mainly the conserved residues and the ten β-strands that form the DEBA|A'GFCC'C" β-sandwich of V-set Ig domains. Two further rounds of rigid-body refinement and manual inspection resulted in an *R*-factor of 46.4% for 1094 atoms. After defining and removing 8.5% of unique reflections for the calculation of R_{free}, cycles of positional refinement and model rebuilding in O were used to convert the J539 sequence into that of MFE-23 and to build the antigen-binding loops as density became apparent, adding a total of 584 atoms. The 1678 atoms in this model gave an *R*-factor of 26.2%. Next, the slowcool simulated annealing protocol in X-PLOR was used to refine the model against reflections between 8 and 2.4 Å using an initial temperature of 4,000 K and a weight of 100,000, and a

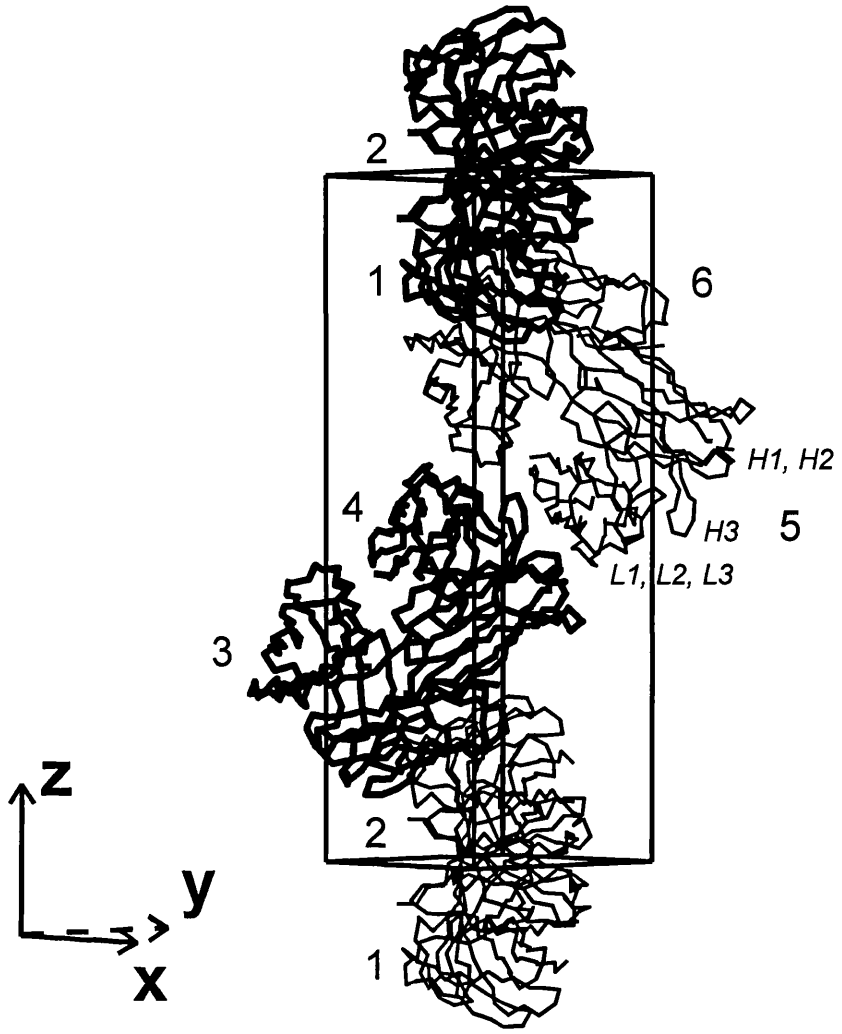


Figure 6.6. Crystal packing of MFE-23 within a single unit cell. In this view, the X-axis points out of the plane of the paper and the Y-axis points into the plane of the paper. Eight MFE-23 structures are shown, numbered 1-8, six of which are found per unit cell. The six loops *H1-H3* and *L1-L3* are identified in molecule 5. The contacts between *H1-H3* of molecule 4 and the EBA face of the V_L domain of molecule 5 is visible.

Table 6.4. Refinement statistics for the final MFE-23 crystallographic model

Shell (Å)	Number of reflections	Cumulative *R*-factor (%)[1]	Number of reflections	Cumulative R_{free} (%)
8.00-4.81	1088	20.4	93	29.1
4.81-3.97	1031	17.1	115	24.6
3.97-3.51	1070	16.9	71	23.0
3.51-3.21	1005	17.3	98	23.5
3.21-3.00	1059	17.8	77	23.6
3.00-2.83	982	18.4	104	24.2
2.83-2.69	1035	18.9	73	25.0
2.69-2.58	1012	19.4	93	25.3
2.58-2.48	1006	19.9	100	26.2
2.48-2.40	1007	20.5	85	26.7

[1] Cumulative values correspond to all reflections in the resolution range from 8 Å to the smaller value specified in the first column.

further 18 atoms were added; this gave good correlation between the rates of reduction in the *R*-factor and R_{free} to values of 26.1% and 30.9% respectively (Kleywegt & Jones, 1997). *B*-refinement lowered these values to 25.5% and 29.9% respectively. Forty-nine water molecules were now included using the PEAKMAX and WATPEAK programs using peaks (> 3.5σ) found in the F_o-F_c map at distances of 0.1 to 3.5 Å from the protein surface. A model with 49 water oxygen atoms resulted in an *R*-factor of 20.5% and an R_{free} of 26.7%, where the distributions of *R* and R_{free} against resolution are shown in Table 6.4. Next, using 90 water molecules, the use of positional and *B*-refinements against all reflections between 8 and 2.4 Å resulted in the final model with 1697 atoms and an *R*-factor of 19.0%. The secondary structure was assigned using the DSSP program (Kabsch & Sander, 1983). The model was stereochemically verified using the PROCHECK program (Laskowski *et al.,* 1993). Solvent accessibilities were calculated using a probe of 1.4 Å in the COMPARER program (Lee & Richards, 1971; Šali & Blundell, 1990). The electrostatic surface charge was calculated using INSIGHT II 95.0 and DELPHI software (Biosym/MSI, San Diego, U.S.A.) on INDY Workstations (Silicon Graphics, Reading, U.K.).

6.3. Results and discussion

6.3.1. Refined molecular structure of MFE-23

A high quality electron density map for MFE-23 was derived using X-ray diffraction data at 2.4 Å resolution from a single well-formed crystal of MFE-23 (Methods). The $P3_221$ space group of the crystal was identified and the structure of MFE-23 was solved by molecular replacement as a single molecule. This was based on the Fv fragment from the murine IgA Fab molecule J539 as the starting model (Figure 6.6; Methods). The MFE-23 molecules are seen as dimers in one set of crystal lattice packing contacts, as exemplified by the pairs numbered as 1,2 and 3,4 in Figure 6.6. The initial model was successfully refined to a final model with an *R* value of 19.0% (Methods). Good visual agreement was seen with a $2F_o$-F_c electron density map contoured at 1σ, which is exemplified by views of the *H3* antigen-binding surface loop in Figure 6.7. This was supported by a mean correlation coefficient of 0.9 for the real-space fit of the model against this map. Generally, the V_L domain with a mean *B*-factor of 21.2 $Å^2$ fitted the electron density better than the V_H domain with a mean *B*-factor of

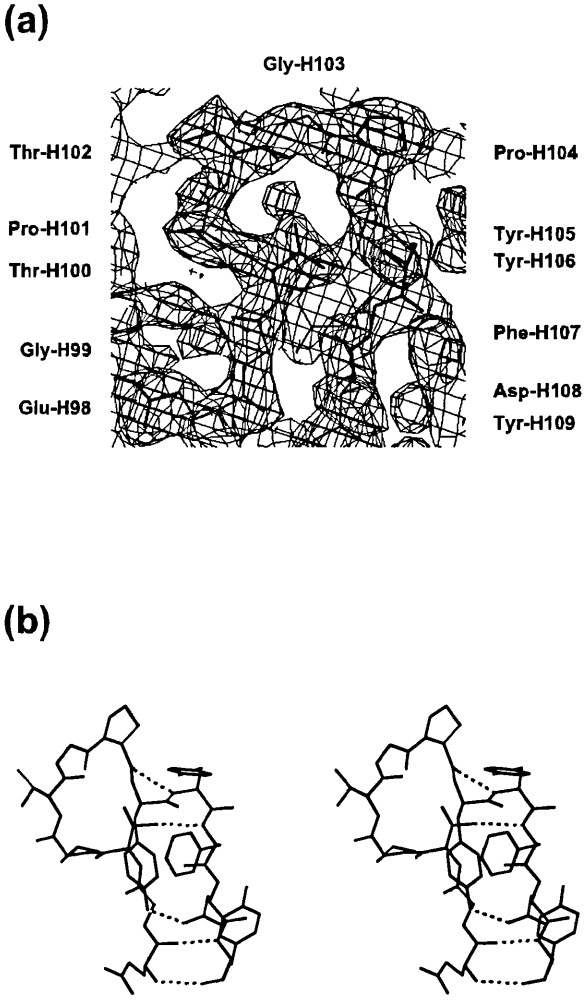


Figure 6.7. Structure of the *H3* loop of MFE-23. (a) The $2F_o$-F_c electron density map phased by the final MFE-23 model at 2.4 Å and contoured at 1σ is shown with electron density for residues H98 to H109 of CDR-*H3*. (b) Stereoview of residues H99 to H109 in the hypervariable *H3* loop. The sequence shown in a clockwise direction is Glu^{H98}-Gly-Thr-Pro-Thr-Gly-Pro-Tyr-Tyr-Phe-Asp-Tyr^{H109}. Hydrogen bonds (broken lines) were assigned if the angle between the acceptor and donor is greater than or equal to 120° and their separation within 2.5 Å for H-N or H-O bonds or 3.0 Å for N-O or O-O bonds.

26.9 Å^2. The MFE- 23 secondary structure displayed the typical V-set β-strand topologies with two β- sheets DEBA and A'GFCC'C" in both the V_H and V_L domains. Quantitative analyses using the DSSP secondary structure analysis program showed excellent correlations with the β-sheet structures of other murine and human Fv structures (Figure 6.5).

The correlation coefficients of the real-space fit of the model to the electron density and to the *B*-factor values was fully compatible with the secondary structure of MFE-23 determined using DSSP (Figure 6.8). Note that the MFE-23 sequence numbering in Figure 6.5 is used throughout unless the Kabat database numbering is specified. The *B*-factors were lowest in the regions occupied by β-strands. The N-terminal, C-terminal and loop residues were generally more flexible and fitted the density less well than the β-strand residues. Within the V_H domain, these included the sidechains of residues H40 to H44 on the loop between β-strands C and C' which form the intermolecular crystal lattice contacts at the MFE-23 dimer interface. The crystal lattice also included a second set of contacts that comprise regions close to the six antigen binding loops *H1, H2, H3, L1, L2* and *L3* (Figures 6.5 and 6.6). These loops, in particular those for *H1, H2* and *H3*, are easily visualised in molecules 4 and 5 in Figure 6.6, and the interaction between *H1, H2* and *H3* in molecule 4 and the DEBA β-sheet of the V_L domain in molecule 5 is likewise easily visualised in Figure 6.6. The reduced *B*-factors of the 9 residues of *H3* compared to the five other loops in Figure 6.8 may be explained by the well-defined nature of its interaction with the DEBA β-sheet.

In the final MFE-23 model, several regions in the MFE-23 map were partially disordered and their occupancies were set to zero. In particular, it was not possible to model either the 15 linker residues between residues H120 and L1 or the 15 C-terminal residues after residue L106 which includes the myc-tag, and these were presumed to be flexible and solvent-exposed. The final MFE-23 model contained 1696 atoms, which is 98% of those present in the V_H and V_L domains, and 93% of the total in MFE-23. A total of 90 water molecules was added only when the density was completely unambiguous. Small fragments of spurious density were apparent, which may indicate contacts between the linker, the myc-tag and the Fv structure, and these may correspond

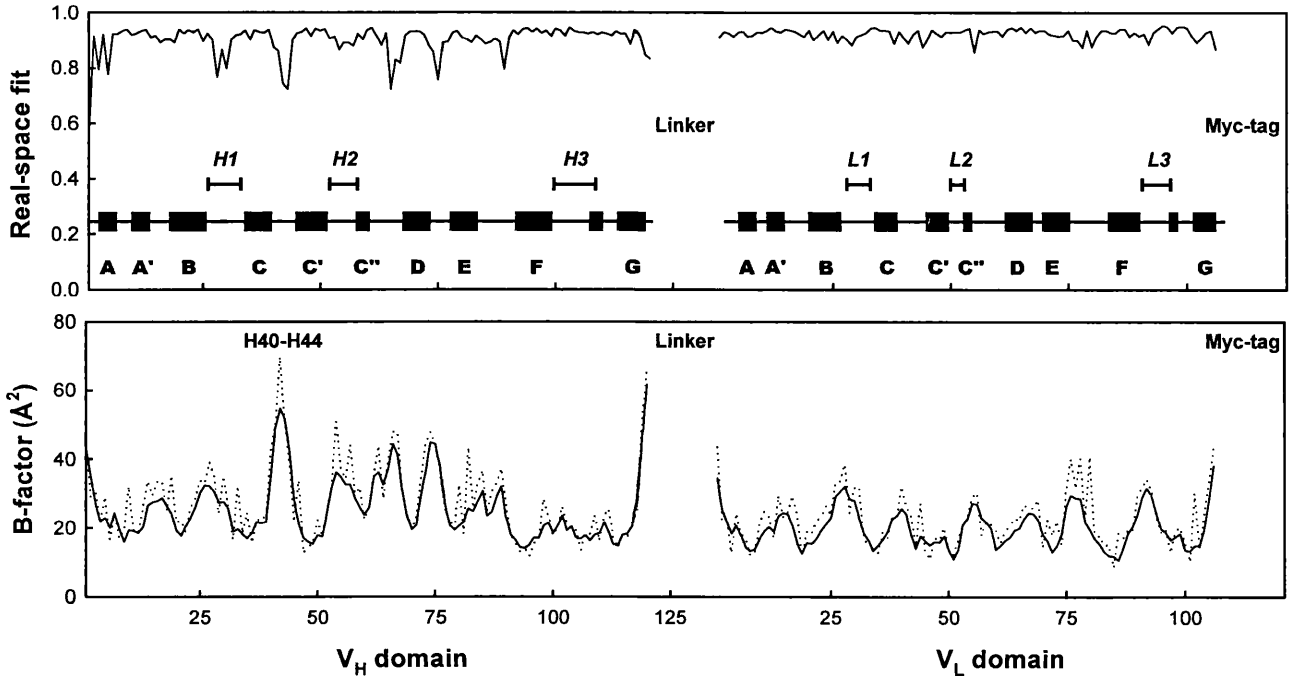


Figure 6.8. Profiles of the final MFE-23 model to evaluate the quality of the structure on a sequential basis. (*a*) Real-space electron density fit correlation coefficient against a $2F_o$-F_c map for each residue. Inside this panel, the β-strand structure of MFE-23 is represented as blocks drawn to scale against residue numbers, together with the structural *H1*, *H2*, *H3*, *L1*, *L2* and *L3* loops identified using the DSSP program (Kabsch & Sander, 1983). (*b*) Average *B*-factor (Å^2) for the mainchain (continuous line) and the sidechain (broken line) of each residue.

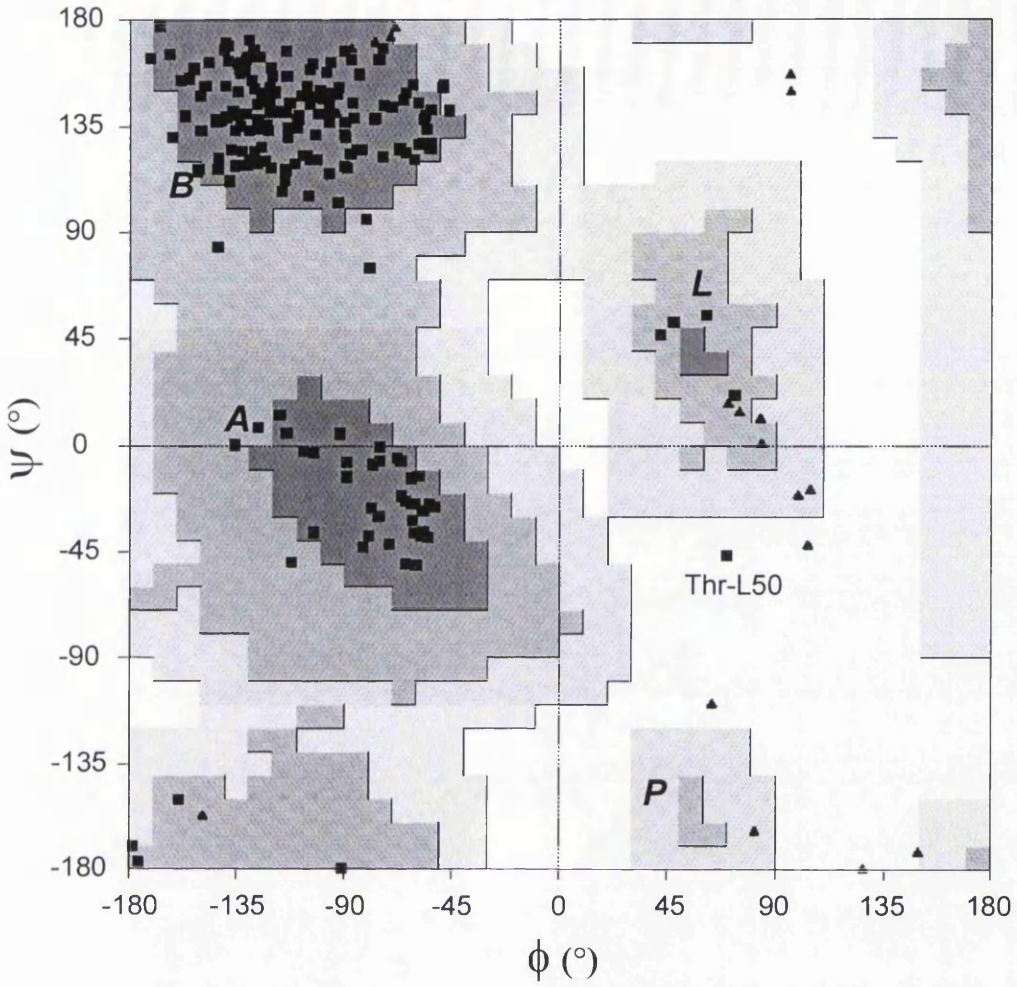


Figure 6.9. A Ramachandran plot of the mainchain torsion angles ϕ and ψ of the final MFE-23 model. The plot was generated using PROCHECK. Allowed regions are shown in grey for α-helix (*A*), β-sheet (*B*), left-handed helix (*L*) and poly-proline (*P*) mainchain conformations, with the most favoured conformational region indicated by increased darkness. Glycine residues are indicated by triangles and all other residues are shown as squares. Thr-L50 is the only non-glycine residue that occupies a disallowed region of the plot.

to unmodelled protein atoms. The final MFE-23 model yielded a good Ramachandran plot of the main chain torsion angles ϕ and ψ (Figure 6.9) (Ramachandran & Sassiekharan, 1968). The only outlier was Thr-L50 with $\phi = 68°$ and $\psi = -46°$ in a γ-turn even though this was well defined in the electron density map, and this is in common with many other Fab and Fv structures (Banfield *et al.*, 1996, 1997).

6.3.2. Crystallographic dimer in the MFE-23 structure

The crystal packing arrangement of MFE-23 was analysed to see whether MFE-23 formed monomers or dimers. Well-defined back-to-back intermolecular contacts were observed such as those seen between molecules 1 and 2 or molecules 3 and 4 pairs in Figure 6.6. At the centre of the putative dimer, Pro-H41 in the two adjacent V_H domains form hydrophobic contacts with each other, and Pro-L39 and Gly-L40 in the two adjacent V_L domains do so likewise. In fact, the Pro-H41 contacts correspond to partial localised disorder in the MFE-23 structure, and this is reflected in the highest *B*-factors that were observed in the H40-H44 surface loop in the MFE-23 structure (Figure 6.8). At the surface of the putative dimer, other reciprocal contacts between the two MFE-23 molecules occur. These include hydrogen bonds between Thr-H115 and Glu-L104 on the β-strands G of a pair of opposing V_H and V_L domains, and hydrophobic contacts between Ala-H9, Leu-H11, Ala-L9 and Ile-L10 on the β-strand A' of a pair of opposing V_H and V_L domains.

The $(Gly_4Ser)_3$ linker between the V_H and V_L domains is not visible in the MFE-23 crystal structure. If fully extended, this linker would be 5.3 nm in length. This is compatible with the 3.0 nm separation between the α-carbon atoms of Ser-H120 at the V_H C-terminus and Glu-L1 at the V_L N-terminus within a MFE-23 monomer if this is formed by the standard interaction between two GFC faces in an antibody Fv fragment (Chothia *et al.*, 1985). It is also compatible with the 2.1 nm separation required to form a diabody scFv structure, in which the V_H and V_L domains within a single scFv molecule are first separated from each other, then reassociate through their GFC faces with a second scFv molecule to form a dimer (see the right or left pairs in Figure 7.1; Chapter7). This showed that the presence of a monomer or dimer of MFE-23 in the crystal structure could not be distinguished.

The total accessible sidechain surface area lost by the interaction between the V_H and V_L domains is 80 Å^2 per domain out of a total accessible sidechain area of 1300 Å^2 per domain. That lost by the interaction between a standard pair of V_H and V_L domains in an Fv fragment is 230 Å^2 per domain. The magnitudes of these surface area changes make it more likely that two independent monomers of scFv have associated to form the crystallographically-observed dimer. That MFE-23 is monomeric in solution was shown using neutron scattering (Chapter 7).

6.3.3. The six CEA-binding loops of MFE-23

The six antigen-binding loops or complementarity-determining regions *H1*, *H2*, *H3*, *L1*, *L2* and *L3* are defined by slightly different residues that depend on whether they correspond to structural variability (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Al-Lazikani *et al.*, 1997) or sequence hypervariability (Wu & Kabat, 1970). These are denoted as structural and hypervariable loops respectively (Figure 6.5). *H3* does not exhibit canonical structures, while the five structural loops *H1*, *H2*, *L1*, *L2* and *L3* have restricted mainchain conformations known as canonical structures (Table 6.5; Figures 6.7, 6.10 and 6.11). The hypervariable definition is used to discuss the specificity-determining residues (SDRs) for antigen binding, in which loop residues that have a higher relative variability compared to others in the loops are correlated with residues that are important in determining antigen specificity (Padlan *et al.*, 1995).

In antibody-antigen complexes, *H3* arguably plays the most important role in antigen binding (Padlan, 1994). In MFE-23, its sequence shows that it is largely hydrophobic. In the crystal structure, *H3* had clear density in a $2F_o$-F_c map contoured at 1σ (Figure 6.7a). Most interestingly, it forms crystal lattice contacts at one end of the DEBA face of the MFE-23 L chain (Figure 6.6). It possessed a β-hairpin structure which is looped to one side at its tip (Figures 6.7b and 6.11), and contains 9 structural residues H100 to H108 and 11 hypervariable residues H99 to H109. Its most notable feature is that, atypically of other structural *H3* loops, the position of the α-carbon atom of Gly-H99 before the loop deviated by 3.2 ± 0.4 Å from those found in 9 representative Fab crystal structures (see below). Eight contiguous *H3* residues H99 to H106 occupy SDR positions. Thr-H100 is buried and is centrally important for the *H3* structure. Its

Table 6.5. Solvent accessibilities of the antigen-binding loops in the crystal structure of MFE-23 and in two homology models

MFE-23 *H3* residues	Residue sidechain accessibility (%) Crystal structure	Joint canonical-database model	Database model
Gly-H99	4	23	0
Thr-H100	5	73*	11
Pro-H101	32	15	36
Thr-H102	98	75	83
Gly-H103	80	95	95
Pro-H104	68	95*	71
Tyr-H105	6	75*	32*
Tyr-H106	63	77	36*
Phe-H107	30	25	1*
Asp-H108	0	54*	60*

MFE-23 residues	Number of residues in MFE-23	R.m.s. deviation of α-carbon positions (Å) Joint canonical-database model	Database model
All residues (H2 to H118, L1 to L106)	223	1.49	1.24
Framework	186	0.91	0.95
Structural *H1* (H26-H32)	7	0.67	0.77
Structural *H2* (H52-H57)	6	0.32	0.44
Structural *H3* (H100-H108)	9	2.73	2.36
Structural *L1* (L26-L31)	6	1.22	1.36
Structural *L2* (L49-L51)	3	0.10	0.09
Structural *L3* (L90-L95)	6	0.41	0.54

* Changes greater than 25% between the crystal structure and its modelling are asterisked.

Figure 6.10. (Overleaf) Ribbon views of the antigen binding loops of MFE-23 (left) and A5B7 (right: PDB code 1clo), both of which bind to CEA. The V_H and V_L domains are shown in green and cyan respectively, and the *H1*, *H2*, *H3* and *L1*, *L2*, *L3* loops are shown in magenta and yellow respectively. These are shown using the structural and hypervariable definitions(Figure 6.5), except that only residues H50-H60 of *H2* are shown. Beneath the ribbon views is a sequence alignment of the A5B7 antigen-binding loops with the full MFE-23 sequence and its β-strands and loops, in which the structural Chothia loops are underlined, and the hypervariable Kabat loops are in bold (Figure 6.5). Residue similarities are indicated by vertical strokes.

Figure 6.10. Ribbon views of the antigen binding loops of MFE-23 (left) and A5B7 (right: PDB code 1clo), both of which bind to CEA (legend on page 273).

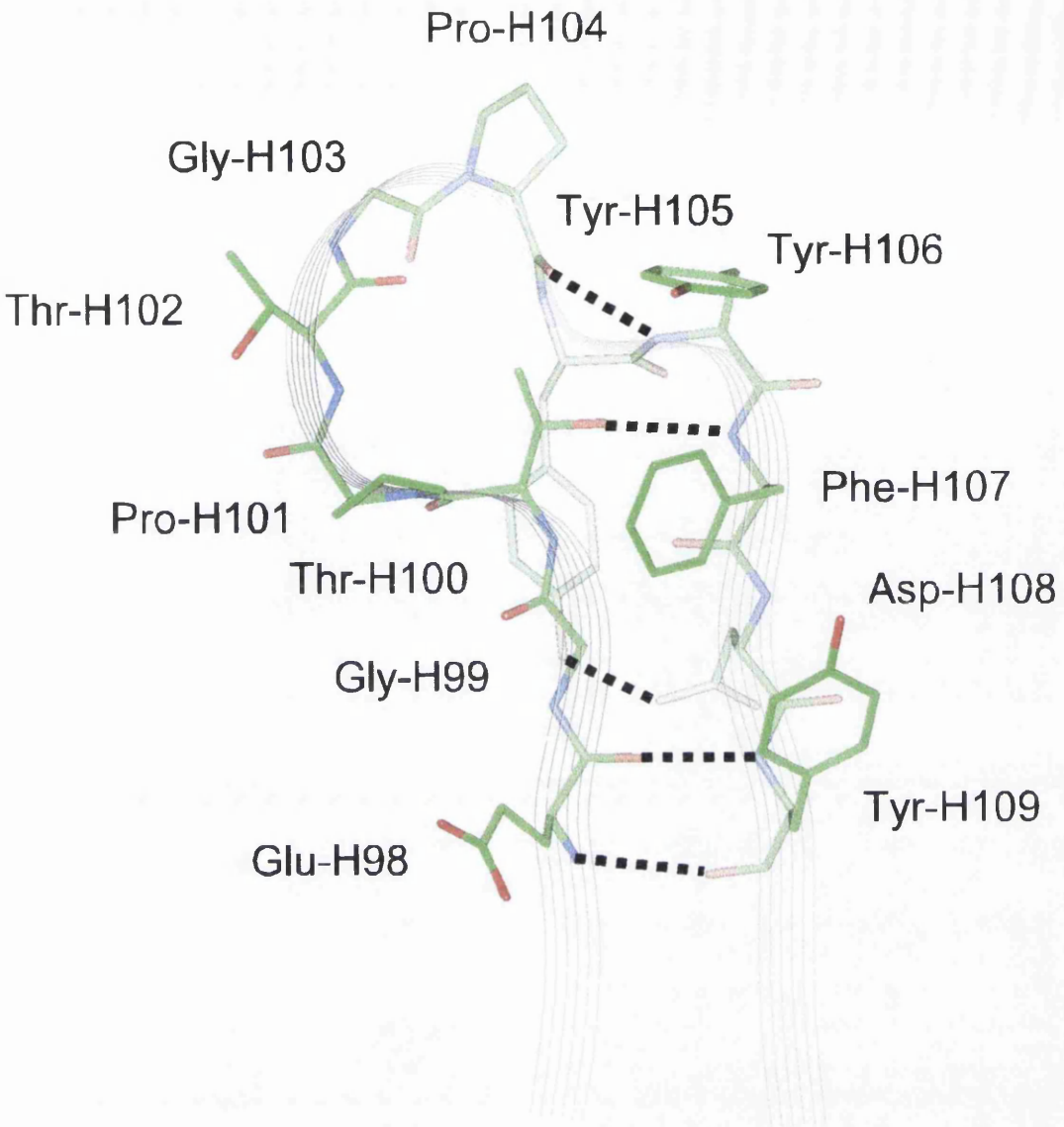


Figure 6.11. The structure of CDR-*H3* of MFE-23. The structure for residues H98 to H109 is shown. The mainchain of CDR-*H3* is shown as a black ribbon representation. This ribbon representation is extended either side of the loop to show the directions of β-strands F and G. Atoms are coloured according to type; green is carbon, blue is nitrogen and red is oxygen. Hydrogen bonds (broken lines) were assigned if the angle between the acceptor and donor is greater than or equal to 120° and their separation within 2.5 Å for H-N or H-O bonds or 3.0 Å for N-O or O-O bonds.

sidechain occupies a cavity formed by the looped mainchain conformation between residues H100 and H107 and the aromatic sidechains of Tyr-H105, Tyr-H106 and Phe-H107. Its OH group forms a hydrogen bond with the mainchain NH of H107 and its mainchain NH group forms one with the CO of Phe-H107. The sidechains of Thr-H102, Pro-H104 and Tyr-H106 have high solvent accessibilities (Table 6.5). Residues H98 and H109 form part of the β-ladder between β-strands F and G. As found for many other *H3* loops of similar size (Al-Lazikani *et al.*, 1997), the ϕ and ψ angles generally correspond to β-strand conformations except for H101, H103 and H106 whose conformations induce bends that characterize the overall shape of *H3*. In this context, Pro-H101 resembles a right-handed helix, Gly-H103 adopts uncommon ϕ and ψ values of 83° and -165° respectively, and Tyr-H106 has a right-handed helical conformation in which its amide group forms a hydrogen bond with the mainchain carbonyl group of H104.

The remaining *H1*, *H2*, *L1*, *L2* and *L3* loops in MFE-23 (Figure 6.5) correspond to known canonical structures that were predictable from their sequences (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Al-Lazikani *et al.*, 1997). These are summarised as follows:

The structural *H1* loop contains residues H26 to H32 which correspond to the single known canonical class of structures for this loop. Structurally important residues are Gly-H26 which produces a sharp turn in the loop, Phe-H27 which is partially buried, and Ile-H29 which is completely buried (Chothia *et al.*, 1989). The structurally important Arg residue normally found at H98 is replaced by the buried residue Glu-H98, whose carboxyl group forms a hydrogen bond with the carbonyl oxygen of Ile-H29. The hypervariable *H1* loop contains residues H31 to H35, in which the buried residue Met-H34 is also important for its structure, and Tyr-H33 and His-H35 occur at SDR positions.

The structural *H2* loop is residues H52 to H57 and belongs to canonical class 2 (type A). The conformations of Gly-H56 and Pro-H53 are important for its structure. The hypervariable *H2* loop is H50 to H66, and residues Trp-H50, Asp-H52, Glu-H54, Asp-H57 and Glu-H59 occur at SDR positions. This corresponds to a striking arrangement of acidic residues.

The structural *L1* loop is residues L26 to L31 and belongs to canonical class 1 for κ V_L domains. The buried residue Val-L29 is important for its structure. The hypervariable *L1* loop is residues L24 to L33, and the buried residues Ala-L25 and Met-L32 adjacent to *L1* and Tyr-L71 on β-strand E are important for the *L1* structure. His-L33 occupies an SDR position.

The structural *L2* loop is a three-residue turn consisting of residues L49 to L51 for which only one canonical class is known, and Ile-L47 and Gly-L63 are structurally important. Thr-L50 was the only outlier on a Ramachandran plot (Figure 6.9). The hypervariable *L2* loop is residues L49 to L55, and residues Ser-L49, Asn-L52 and Ala-L54 occur at SDR positions.

The structural *L3* loop is residues L90 to L95 and belongs to canonical class 1. Pro-L94 is a cis-peptide which is an important determinant of the extended conformation of residues L92 to L95, and Gln-L89 provides the hydrogen bonds that stabilise this loop. The hypervariable *L3* loop is L88 to L96. Gln-L88, Arg-L90, Ser-L91, Ser-L92 and Tyr-L93 occur at SDR positions.

6.3.4. Appearance of the CEA-binding site of MFE-23

Despite the 50% carbohydrate content of CEA, MFE-23 binds to the protein surface of CEA. Two lines of evidence show this. First, the N-terminal domain pair of the seven in CEA was expressed in an *E. coli* bacterial system, to which carbohydrate is not attached, and MFE-23 binds to this expression product with high affinity at a level comparable to its binding to CEA (J. D. Thornton, P. A. Keep, K. A. Chester and R. H. J. Begent, unpublished results). Second, the solution scattering modelling of the intact CEA structure showed that carbohydrate was fully extended away from CEA potentially to leave large protein surfaces accessible for antibody attachment (Chapter 4).

When viewed sideways, the six antigen-binding loops in MFE-23 creates a generally convex surface, in which the *H3* loop protrudes slightly to form an apex that may interact with an indentation on the surface on CEA (Figure 6.10). When the loops are viewed face-on (Figure 6.12), the mainchain backbone of the six loops forms a large cavity into which sidechains from all six loops protrude. Loops *H3* and *L3* are sandwiched by *H1* and *H2* on one side and *L2* and *L1* on the other (Figure 6.12a). The

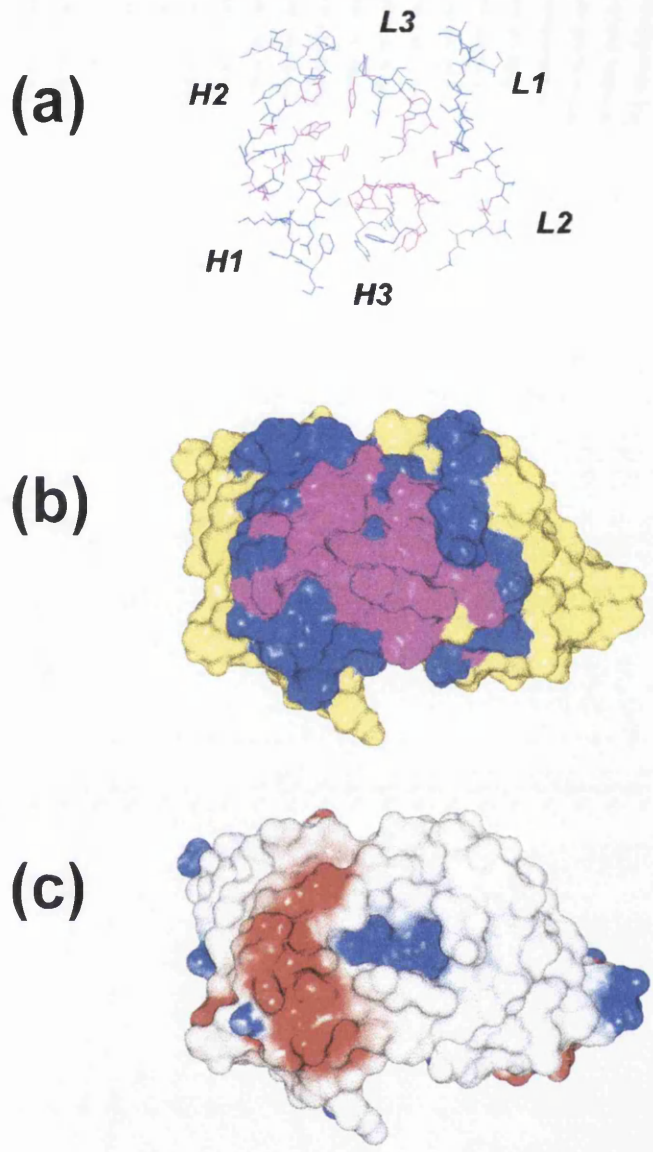


Figure 6.12. Three representations of the antigen-binding site of MFE-23. All three are viewed from the same perspective looking into the site.

(a) Parts of the six structural and hypervariable loops of MFE-23 (loop *H1*: residues H26 to H35; *H2*:H50 to H56; *H3*:H99 to H108; *L1*:L24 to L33; *L2*: L49 to L56 and *L3*:L88 to L96). The specificity-determining residues (SDRs) are coloured magenta (H33, H35, H50, H52, H54, H57, H59, H99, H100, H101, H102, H103, H104, H105, H106, L33, L49, L52, L54, L88, L90, L91, L93; Padlan *et al.*, 1995) and the remaining loop residues are coloured blue.

(b) The antigen-binding site is shown as a Connolly surface that is accessible to a probe of 1.4 Å radius. The loops are coloured magenta and blue as above, and the remainder of the MFE-23 structure is coloured yellow.

(c) Electrostatic view of the MFE-23 antigen binding site. Red represents a potential of less than -4kT (acidic), blue a potential of more than +4kT (basic) and white as 0kT (neutral). Linear interpolation of the colours represents potentials between -4kT and +4kT.

SDR positions are clustered at the centre of the binding region (magenta in Figures 6.12a,b), and occupy ridges surrounding the central cavity (Padlan *et al.*, 1995). The use of the structural and hypervariable loop definitions showed that all the antigen-binding residues were clustered together at the centre of the CEA-binding site (blue and magenta in Figure 6.12b). The 64 structural and hypervariable residues in the six loops have a total sidechain surface area of 2340 $Å^2$, of which 750 $Å^2$ (32%) is solvent exposed. Examination of the antigen-binding region of MFE-23 using electrostatic calculations (Figure 6.12c) showed that a small positive charge existed at its centre, primarily due to Arg-L90, the sidechain of which is 17% solvent exposed. This is flanked by a larger band of negative charge on one side that corresponded to Asp-H31 (70% of sidechain is solvent exposed), Asp-H52 (16%), Glu-H54 (64%), Asp-H57 (65%) and Glu-H59 (36%) on the *H1* and *H2* loops.

The murine monoclonal antibody A5B7 also binds to human CEA. A comparison of the MFE-23 crystal structure and sequence with that of A5B7 (Banfield *et al.*, 1996) showed that the increased CEA binding affinity of MFE-23 compared to A5B7 can be assigned to residues in the *H2* and *H3* loops. This is shown by the remarkable similarities of the *L1*, *L2* and *L3* loops in both antibodies, and the *H1* loop also showed a high degree of similarity. Each pair of these four MFE-23 and A5B7 loops contained the same number of residues and showed sequence identities of 50-100% (Figure 6.10). These four loops also showed similar appearances in Figure 6.10, and belong to the same canonical groups. In distinction to these, the *H2* and *H3* loops differed in that they have different lengths of 1-2 residues and reduced sequence similarity of 40-50%, the *H2* loop of MFE-23 has four acidic groups compared to one in A5B7, and the *H3* loop of MFE-23 is more hydrophobic than that of A5B7. Their appearances in Figure 6.10 are different where *H2* is classified in canonical group 4 instead of 2A. *H3* is larger in MFE-23, while *H2* is larger in A5B7. An electrostatic map of the antigen-binding site in A5B7 (not shown) did not exhibit the prominent distribution of acidic groups associated with the *H2* loop in MFE-23. Crystal structures for three antibody-antigen complexes have been reported for a common antigen, hen egg white lysozyme. The antibodies bind to essentially nonoverlapping epitopes, and their combining sites are quite different (Padlan, 1990). It is therefore possible that the

similar shape of the MFE-23 and A5B7 antigen-binding site may correspond to their binding to a similar region in CEA.

6.3.5. Identification of MFE-23 residues important for CEA binding

One approach to predict the MFE-23 residues important for CEA binding is to analyse the 64 antigen loop residues. This analysis was based on their sidechain solvent accessibilities, the frequency of contacts made with antigen in known structures, and their occurrence at SDR positions (Figure 6.13) (Padlan *et al.*, 1995). Figure 6.13 showed that a total of 10 residues (Tyr-H33, Glu-H54, Asp-H57, Glu-H59, Thr-H102, Pro-H104, Tyr-H106, Arg-L90, Ser-L91 and Tyr-L93) both presented solvent-exposed areas of over 10 $Å^2$, made contact with antigen at a frequency of at least 30% in known structures, and occurred at SDR positions. These were predicted to be strong candidates for mediating the binding of MFE-23 to CEA. Studies of the amino acid distributions in antigen-binding sites, the observed contact residues in antibody-antigen crystal structures, and the chemical and physical properties best suited to antigen-binding show that Trp and Tyr residues play important roles in antigen contacts, and that Ser, Asn and His residues are also typically involved (Padlan, 1990; Mian *et al.*, 1991; Kabat *et al.*, 1977; Lea & Stuart, 1995; Janin & Chothia, 1990). This supports the predicted involvement of 4 of the above 10 residues to bind to CEA (Tyr-H33, Tyr-H106, Ser-L91 and Tyr-L93). The presence of a further 5 loop residues (His-H35, Trp-H50, Tyr-L31, His-L33, and Asn-L55) is consistent with the observed frequencies of antigen-binding residues, and Figure 6.13 shows that these also had a high probability for antigen contacts or have a large sidechain exposure.

A potentially more direct approach to predict the MFE-23 residues important for CEA binding involved the examination of the crystal lattice packing contacts (Table 6.6). Since the comparison between the MFE-23 and A5B7 antigen binding loops showed that the largest differences were present in the *H2* and *H3* loops, it was interesting that the lattice contacts occurred primarily between the *H1*, *H2* and *H3* loops of MFE-23 and the EBA β-sheet of an adjacent V_L domain (reduction of 150 $Å^2$ per domain in the accessible surface area of 2560 $Å^2$ per domain). Less lattice contact was made between *L1* and *L2* and the A' β-strand of an adjacent V_H domain from a different

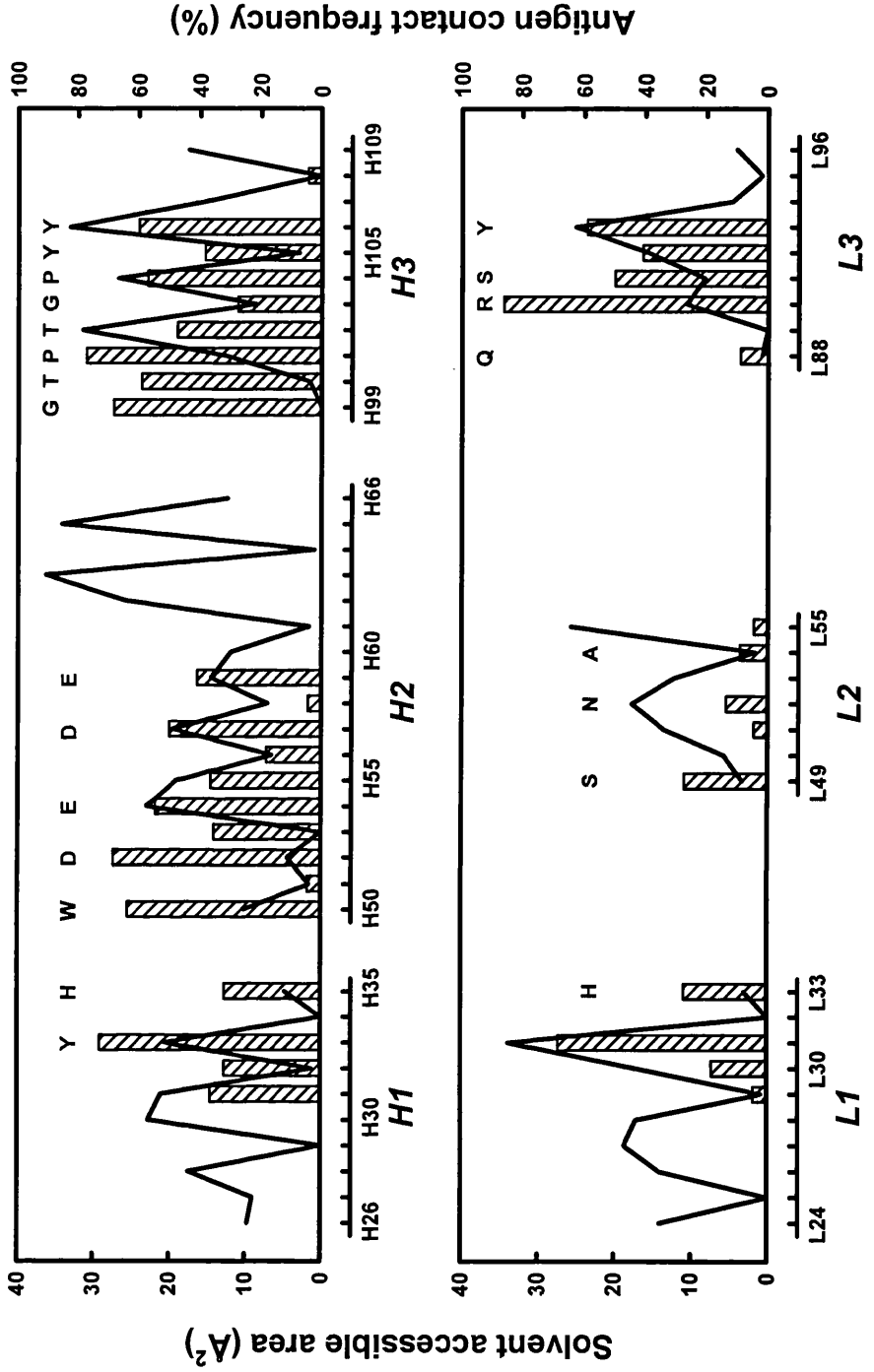


Figure 6.13. Probability of antigen-binding function at residues in the structural and hypervariable loops of MFE-23. The total sidechain surface accessibility (Å^2) determined using a probe of 1.4 Å is plotted for each hypervariable region residue (continuous line). The frequency of each loop residue making contact with antigen in antibody-antigen complexes (bars) was determined from Table 1 in Padlan *et al.* (1995). The 23 SDR positions highlighted in Figures 6.12(a) and 6.12(b) are denoted by their single-letter amino acid codes.

Table 6.6. Contact residues between the MFE-23 antigen-binding loops and the adjacent MFE-23 molecule in the crystal lattice

Loop	MFE-23 antigen binding site residues	Contact residues on adjacent molecule
H1	Lys-H30 C=O	Lys-L18 NZ (2.8 Å; hydrogen bond)
	Asp-H31 CO_2^-	Ser-L20 OG (5.1 Å; hydrogen bond via H_2O)
	Tyr-H33* OH	Arg-L76 NH2 (3.6 Å; hydrogen bond)
H2	Asp-H52 CO_2^-	Lys-L18 NZ (2.6 Å, 3.0 Å; salt bridge)
	Glu-H54 CO_2^-	Arg-L76 NH1 (2.9 Å; 3.6 Å; salt bridge)
H3	Thr-H102 OH	Lys-L18 C=O (2.6 Å; hydrogen bond)
	Pro-H104	Pro-L8 (4.3 Å; hydrophobic)
	Pro-H104	Leu-H11 (4.1 Å; hydrophobic)
	Tyr-H106 OH	Thr-L20 C=O (2.7 Å; hydrogen bond)
L1	Tyr-L31 OH	Glu-H10 CO_2^- (3.4 Å; hydrogen bond)
L2	Ser-L51 OH	Arg-H13 mainchain NH (3.4 Å; hydrogen bond)

* Adjacent to H1 structural loop

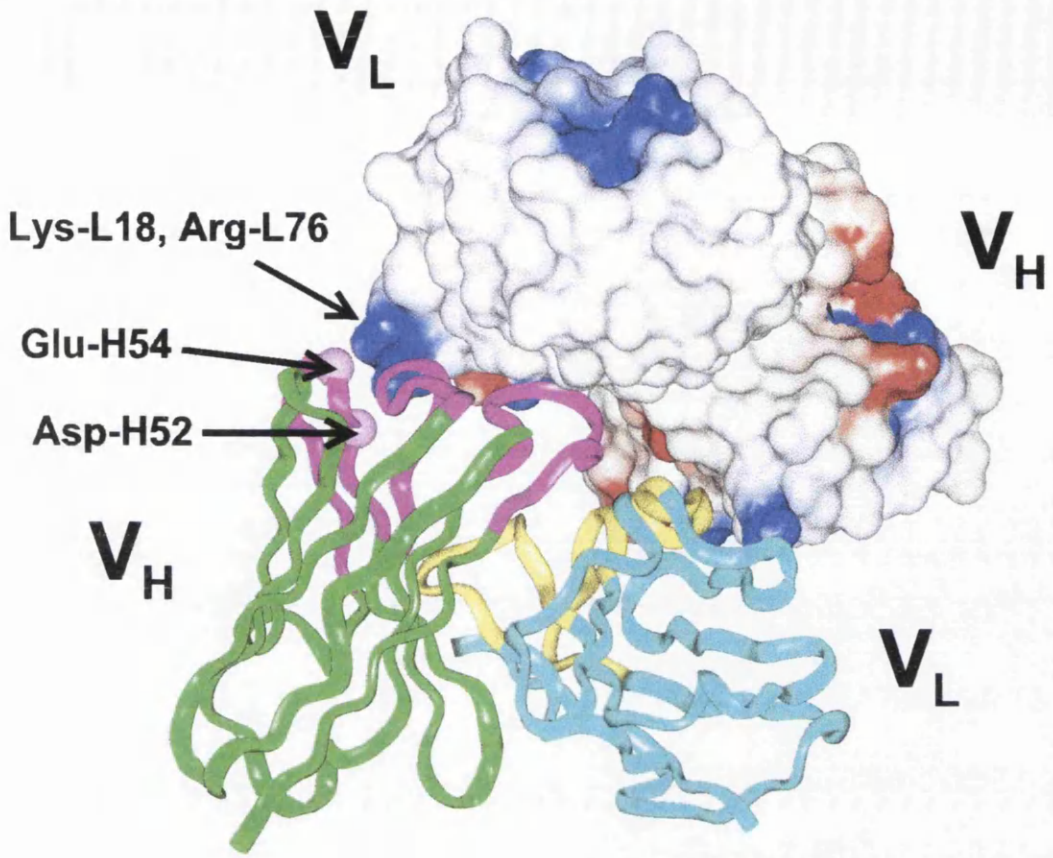


Figure 6.14. Interaction between the MFE-23 antigen-binding loops with two other different MFE-23 molecules in the crystal lattice packing of MFE-23. The ribbon view of the MFE-23 molecule is coloured to follow Figure 6.10. The positions of Asp-H52 and Glu-H54 (arrowed) are shown as spheres. These form contacts with Lys-L18 and Arg-L76 in the V_L domain (arrowed).

molecule (reduction of 80 Å^2 in the surface accessible area), while *L3* made no contact. This is consistent with the observed frequency of participation of *L3*, *H2* and *H3* loops in antibody-antigen interactions (Wilson & Stanfield, 1994). Although the buried solvent accessible surface is much lower than the values of 632-916 Å^2 that have been observed in structures of complexes between antibodies and protein antigens (Braden & Poljak, 1995). The view of Figure 6.14 showed good surface complementarity in that *H3* was inserted at an indentation formed between the two domains, while *H2* was positioned on the other side of the V_L domain. This caused the *H2* and *H3* loops to be wrapped around the V_L domain. The changes in surface accessibilities were used to construct a list of 10 contact residues (Table 6.6). This showed that Asp-H52 and Glu-H54 on the *H2* loop made salt bridge contacts with Lys-L18 and Arg-L76 on the adjacent molecule. This proximity relationship is highlighted in the electrostatic maps of Figures 6.12(c) and 6.14. *H3* formed hydrogen bond and hydrophobic contacts with both the V_L and V_H domains. Six of these 10 contact residues (Tyr-H33, Glu-H54, Thr-H102, Pro-H104, Tyr-H106 and Tyr-L31) correspond to the 15 predicted antigen-binding residues identified above. The agreement between the *H1*, *H2* and *H3* lattice contacts and predicted contacts made it likely that they are important for MFE-23 binding to CEA. It should however by noted that the two neighbouring V_L and V_H domains in the lattice are orientated antiparallel to each other, while the domains in CEA are parallel. This means that only the *H1*, *H2* and *H3* interactions are likely to be biologically significant.

The interaction of MFE-23 with CEA has been examined by site-specific mutagenesis (Read *et al.*, 1995). A Tyr-H106-Pro mutation abolished MFE-23 binding to CEA, in agreement with the lattice and prediction analyses above and the general involvement of Tyr residues in combining sites (Padlan, 1990). Note that the structural consequence of the Tyr-H106-Pro mutation should leave the mainchain structure of the *H3* loop unaltered as the ϕ and ψ angles of Tyr-H106 were similar to those expected for Pro residues. A Tyr-H105-Ala mutation caused a minor perturbation in MFE-23 binding, in agreement with the absence of this residue from the lattice and prediction analyses. This is attributed to the burial of Tyr-H105 within *H3*, so its mutation would not have affected MFE-23 binding. However a Thr-H102-Ala mutation improved the

MFE-23 binding to CEA, which confirmed the lattice and prediction analyses which suggested that this was important. The effect of these three V_H mutations suggest that the lattice contacts can act as a good model for the complex between MFE-23 and CEA.

6.3.6. Comparison of the MFE-23 crystal structure with two homology models

Prior to this study, two homology models of MFE-23 had been constructed using the program AbM (Martin *et al.*, 1989; Pedersen *et al.*, 1992; Pedersen & Rees, 1993; Read *et al.*, 1995). In the first model, the *L2, L3, H1* and *H2* loops were constructed using canonical structures, and the *L1* and *H3* loops were constructed using a database search. In the second model, the full Fv structure was modelled using database methods. Both models were energy minimised, in which the framework backbone was fixed and the sidechains and antigen-binding loops were allowed to move. These models permitted a blind test of the prediction of a complete antigen-binding site (Bajorath & Sheriff, 1996). For each comparison, the α-carbon atoms of each peptide being compared were superimposed using INSIGHT II. Comparison of the models with the crystal structure showed r.m.s. deviations of 1.49 Å and 1.24 Å at 223 α-carbon positions (Table 6.5). This is within the expected r.m.s. deviation of 1.73 Å for satisfactory homology modelling (Sutcliffe *et al.*, 1987). The framework and the individual *L2, L3, H1* and *H2* loops also showed low r.m.s. deviations of 0.09-0.95 Å, which are comparable to those reported for the anti-tumour antibody BR96 (Bajorath & Sheriff, 1996), while that for the *H3* loop was noticeably higher at 2.36-2.73 Å in both models. This showed that the canonical structures gave good *L2, L3, H1* and *H2* loop models, and that the database method gave only marginally worsened agreements. In accord with general experience (Wilson & Stanfield, 1994; Banfield *et al.*, 1996; Bajorath & Sheriff, 1996), the *H3* loop was poorly predicted. Four of the nine residues in the modelled *H3* loops had solvent accessibilities that were significantly different from those in the crystallographic *H3* loop (Table 6.5), while only 5 of the other 28 modelled loop residues showed less than 25% differences in accessibility with those in the crystal structure. This difference is attributable to the lack of recognition of the significance of Thr-H100 and Gly-H99 in stabilising the loop structure.

These results may be explained in terms of hydrogen bond formation in the

crystal structure and the two models. Using INSIGHT II, the crystal structure was determined to possess 156 standard mainchain hydrogen bonds, while the two models contained 90-91 hydrogen bonds. In particular, the models did not identify hydrogen bonds formed between the buried sidechain of Asp-H108 and those of Asn-H97, Tyr-H105 and Trp-H110. Likewise they did not identify that formed between the buried sidechain of Thr-H100 and the mainchain NH group of Phe-H107. Better loop prediction algorithms may result from the more accurate recognition of potential hydrogen bonds.

6.3.7. The structural classes of the MFE-23 V_H and V_L domains

The framework regions correspond to residues not involved in the antigen-binding loops. In the murine and human V_H domains, these exist as one of three structural classes 1, 2 or 3 depending on the residues at H9 and H67 at opposite edges of the β-sheet sandwich (Kabat numbering; Saul & Poljak, 1993). During crystallographic refinement, it became apparent that MFE-23 belonged to a different V_H class from that of the initial J539 model because of differences at residues H8 to H10. Residue H9 occurs at the kink between β-strands A and A' which are hydrogen bonded to β-strands B and G respectively. In MFE-23, Ala-H9 is within the A'GFCC'C" β-sheet, while this is not so in J539. Using Kabat numbering, MFE-23 contained Ala-H9 and Ala-H67 and belonged to class 2, whereas J539 contains Gly-H9 and Phe-H67 and belongs to class 3 (Kabat numbering). Had Pro-H9 and a non-aromatic residue at H67 been present, this would have identified class 1 (Figure 6.5).

The "humanisation" of MFE-23 will be affected by perturbations of β-strands A, A', B and G in the three V_H classes. "Humanisation" can be achieved either by grafting the six MFE-23 loops onto a human framework region or by converting the MFE-23 framework region into the human equivalent (Saul & Poljak, 1993). To identify the most compatible human V_H framework for murine MFE-23, the MFE-23 V_H domain was superimposed onto murine crystal structures that correspond to the three V_H classes as well as all six human Fab crystal structures known in early 1998 (Table 6.7). The superimposition was based on the 172 mainchain atoms of 43 structurally-conserved framework residues and excluded the conformationally-flexible edge β-

Table 6.7. Superposition of the V_H and V_L domains of MFE-23 with those from known Fab structures

Fab structure	PDB code	Source	Chain subgroup[1]	V_H framework class[2]	Atoms	R.m.s deviation (Å)	Chain subgroup[1]	Atoms	R.m.s deviation (Å)
			V_H domains				V_L domains		
MFE-23	n/a	Murine	IIC	2	172	-	κIV or κVI	140/132	-
D44.1	1mlb	Murine	IIA	2	172	0.45	κV	140	0.56
D1.3	1vfa	Murine	IB	1	172	0.52	κV	140	0.36
J539	2fbj	Murine	IIIB	3	172	0.55	κVI	140	0.40
TR1.9	1vge	Human	n.d.	2	172	0.47	κI	140	0.46
New	7fab	Human	II	1	172	0.71	λI	132	0.78
HuIgM	1igm	Human	III	3	172	0.68	κI	140	0.53
Hil	8fab	Human	III	3	172	0.69	n.d.	140	0.79
Kol	2fb4	Human	III	3	172	0.76	λI	140	0.80
3D6	1dfb	Human	III	3	172	0.79	κI	140	0.48

[1] The V_H and V_L chain subgroups are defined in Kabat *et al.* (1991).

[2] The V_H framework classes are defined in Saul & Poljak (1993).

N.d., not defined.

strands A and G. The superimposition onto the murine V_H domains gave lower r.m.s. deviations than the corresponding human ones (Table 6.7). Both the murine and human class 2 V_H domains gave the lowest r.m.s. deviation of 0.45-0.47 Å, and showed that a human class 2 V_H structure offered the best starting point for "humanisation".

Light chains occur as κ and λ classes. The κ chain of the MFE-23 V_L structure was superimposed onto other murine and human Fab V_L domains using 140 mainchain atoms from 35 framework residues. The murine V_L domains were all κ chains, whereas the human ones were both κ and λ domains. The κ domains gave r.m.s. deviations between 0.36-0.56 Å, while the λ domains gave higher r.m.s. deviations between 0.78-0.82 Å. This is attributed to a major difference between the κ and λ domains at residues L7 to L10 which occupy a kink in β-strand A (Kabat numbering). The κ domains contain a Pro residue at position L8, whereas λ domains contain Pro residues at both L7 and L8 and a deletion at L10. The superimpositions showed again that the best agreement with the MFE-23 structure was determined by the chain type.

6.3.8. Two "humanisation" prediction strategies for MFE-23

The Fv fragment from human Fab TR1.9 offered the best framework model for the "humanisation" of MFE-23 because both its V_H and V_L domains belong to the same classes as those of MFE-23. It is necessary that the framework does not change during "humanisation" (Bajorath *et al.*, 1995). The framework region also defined the origin of the six loops. Comparisons between MFE-23 and the nine structures of Table 6.7 showed that the 12 α-carbon atoms just before and after the six loops were maintained to within 0.6 ± 0.3 Å in position with the exception of H99 (see above), which had to be replaced by H98 for this comparison. Both considerations showed that a "humanisation" strategy based on a transfer of the MFE-23 loops to the TR1.9 framework region was feasible. "Humanisation" strategies have been analysed using crystal structures (Holmes & Foote, 1997; Eigenbrot *et al.*, 1993, 1994; Banfield *et al.*, 1997).

In the first approach for the generation of a humanised MFE-23 model based on the human TR1.9 framework (Padlan, 1994), the TR1.9 sequence was retained as much

as possible after the replacement of its loops with those from MFE-23. The loops were identified using both the structural and hypervariable definitions. Framework residues that may influence the loop conformations were identified using solvent accessibility calculations, in which the loops were deleted from the model, then the accessibility was calculated in order to identify newly-exposed sidechains. In the V_H domain, for the *H1* region, it was deduced that all 10 MFE-23 loop and 3 MFE-23 framework residues were required in TR1.9. For the *H2* region, only 10 of the 17 MFE-23 loop residues were required. For the *H3* region, all 11 MFE-23 loop and 1 MFE-23 framework residue were required. In the V_L domain, a higher sequence and structural similarity between MFE-23 and TR1.9 simplified the extent of residue replacements. In the *L1* region, 10 loop and 3 framework residues were required in TR1.9. In the *L2* region, 7 loop residues and 2 framework residues were required in TR1.9. In the *L3* region, 9 loop residues and 1 framework residues were required in TR1.9. These 67 residue changes were incorporated into a homology model based on the TR1.9 framework with MFE-23 loops. This model was subjected to energy refinement to show that very little sidechain movement resulted and that the model was plausible.

The second approach to "humanisation" (Padlan, 1994) was based on the comparison of the solvent-accessible surfaces of MFE-23 and TR1.9 in order to predict the residue changes in the MFE-23 sequence required to form a humanised framework while leaving unaltered its affinity for CEA. Thus residues in MFE-23 and TR-1.9 that were over 30% accessible were eliminated if they were either identical in the two proteins, or located within one of the six structural or hypervariable antigen-binding loops. Of the total of 226 residues in MFE-23, 116 were over 30% accessible in MFE-23, of which only 19 residues were chemically different between the two proteins, and 10 more were chemically similar (identified in Figure 6.5). A total of 16 of these 29 residues were located in the loops at the opposite end of the MFE-23 structure to that of the antigen-binding site, while 10 were on the DEBA and A'G β-sheets. The construction of a homology model based on these 29 residue changes showed that no detectable effect was seen on the MFE-23 binding loops after energy refinement. The second approach required fewer residue changes and is preferable for this reason. Nonetheless both "humanisation" predictions showed that, in the absence of further

information to reduce the number of required residue changes, the "humanisation" of murine MFE-23 involves considerable effort.

6.4. Conclusions

There are now over 230 antibody crystal structures in the PDB (summer 1998). The overwhelming majority of these are Fab fragments, while others include intact IgG structures, Fc and Fv fragments, and Fab-hapten, Fab-protein and other Fab complexes. The present MFE-23 structure determination is the sixth one involving a scFv molecule, and is the first scFv structure that demonstrates binding specificity for an immunoglobulin fold as found in CEA. Previous ones have been specific for oligosaccharides, membrane glycoproteins, steroids or viruses. Of particular interest is that the small size of the scFv molecule restricts the way in which the lattice packing can be formed. Antibody-antigen contacts reflect the shape and electrostatic complementarity between the antigen-binding loops and their antigen (Braden *et al.*, 1998). Thus the MFE-23 lattice packing may well provide clues on how the antigen-binding loops interact with their target, given that both MFE-23 and CEA belong to the immunoglobulin superfamily. This view was given credibility by the observed interactions between the MFE-23 *H1*, *H2* and *H3* loops and the EBA β-sheet of a neighbouring MFE-23 V_H domain. Evidence that these contacts are biologically significant was obtained by comparison of the observed MFE-23 loop interactions with those predicted for MFE-23, the outcome of site-specific mutagenesis work with MFE-23, and the similarities seen between the antigen-binding loops of MFE-23 and A5B7 that highlighted the importance of the *H2* and *H3* loops. It was concluded from comparisons with other antibody-antigen crystal structures and the observed lattice contacts that Tyr-H33, Glu-H54, Thr-H102, Pro-H104, Tyr-H106 and Tyr-L31 may be important for binding to CEA. In addition a set of five acidic residues, Asp-H31, Asp-H52, Glu-H54, Asp-H57 and Glu-H59, in the MFE-23 binding site might be important for binding to CEA.

Less information was obtained on the MFE-23 binding site on CEA. The immunoglobulin fold does not demonstrate any noticeable features on its surface, apart from indentations at the junctions between successive domains and the presence of

extended carbohydrate chains on several EBA faces in CEA (Chapter 4). No sequence similarity was detected between the MFE-23 V_H domain and any potential CEA epitopes resembling the residues summarised in Table 6.6. It was concluded that MFE-23 may interact directly with a β-sheet of CEA or at the indentation between two CEA domains to follow Figure 6.14, and further structural studies are required. (see Chapter 7)

Other features in the MFE-23 crystal structure correspond to those seen in other scFv crystal structures. Thus, in common with these other studies, MFE-23 was observed as a dimeric four-domain structure in its crystal lattice and its linker peptide was invisible in the electron density map. While the β-strands and the canonical and non-antigen binding loops possessed structures as expected, the *H3* antigen-binding loop (Figures 6.7 and 6.8) once again corresponded to a structure that could not be predicted by homology modelling. The crystal structure provided insight on possible "humanisation" strategies for murine MFE-23 on the basis of similarities with human TR1.9 and experience from other crystallographic studies.

Chapter 7

The Solution Structure of MFE-23 by Neutron Scattering and an Outline Model of the Complex Formed Between MFE-23 and CEA

7.1. Introduction

Two major divisions of the immunoglobulin superfamily (IgSF) are the antibodies (immunoglobulins) and cell surface proteins (Williams & Barclay, 1988; Chapter 1). While immunoglobulin G and A (IgG and IgA) exist a 12-domain proteins constructed from four V-set and eight C1-set immunoglobulin folds, the carcinoembryonic antigen (CEA) is constructed from one V-set, three I-set and three C2-set immunoglobulin folds and has a particularly high carbohydrate content of 50% by weight (Harpaz & Chothia, 1994; Chapter 4). Both crystallography and solution scattering work show that the domains in IgSF proteins are frequently arranged as extended chains of multiple domains which either exist as single chains or in association with other chains by means of interactions between domain pairs (Harris *et al.*, 1998b; Perkins *et al.*, 1991; Mayans *et al.*, 1995; Chapters 4, 5 and 6). Members of the IgSF are frequently involved in protein-protein interactions, where adhesive interactions may be formed with antigens or with cell-surface proteins on opposing cells. From a clinical perspective, the interactions between antibodies and the cell-surface protein CEA are important because they are used for developing tumour targeting strategies. Recently, crystal structures have been determined for the MFE-23 and A5B7 antibody fragments, which are both anti-CEA tumour targeting agents (Chapter 6; Banfield *et al.*, 1996). Knowledge of how such antibody fragments behave in solution and how they interact with their target antigen is important for the development of successful anti-tumour strategies. In the case of MFE-23, two sets of lattice interactions were observed which have significant implications for its solution structure and its binding to CEA (Chapter 6).

A single MFE-23 molecule contains a V-set heavy chain (V_H) immunoglobulin domain which is joined to a V-set light chain (V_L) domain by a $(Gly_4Ser)_3$ linker. Due to the first set of crystal lattice contacts, MFE-23 was seen to form a four-domain structure (Figure 7.1). However no linker was visible in the electron density map. Consequently it was not possible to distinguish between monomeric or dimeric forms of MFE-23 in its crystal structure. Table 7.1 summarises five earlier crystal structures that correspond to the scFv molecules Se155-4, NC10, L5MK16, CC49/212 and C219 (Zdanov *et al.*, 1994; Kortt *et al.*, 1994; Perisic *et al.*, 1994; Raag & Whitlow, 1995;

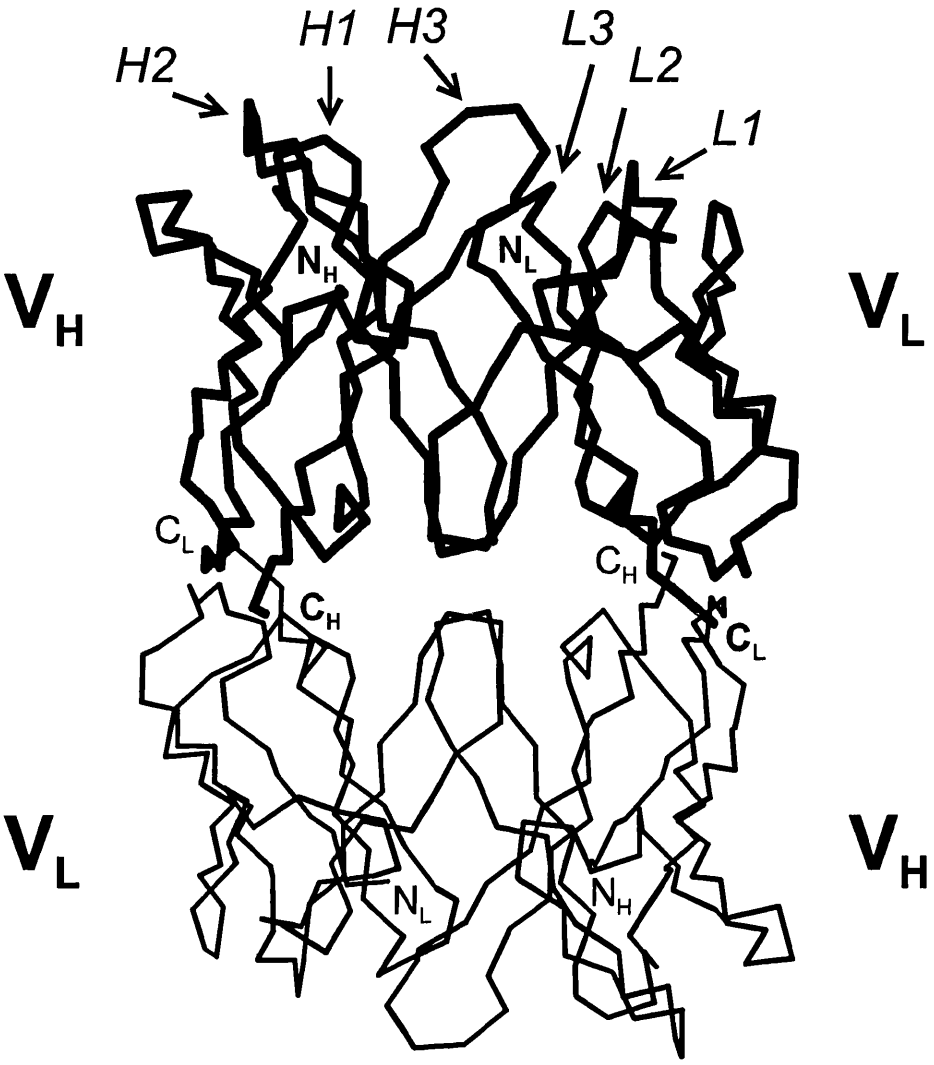


Figure 7.1. Dimeric association of Fv structures in MFE-23 crystals. A single Fv structure as observed in the asymmetric unit of MFE-23 crystals is shown as an α-carbon trace (bold line), and the positions of the N- and C-terminal residues of its V_H and V_L domains are indicated. The Fv structure produced by two-fold (Y, X, -Z) crystalline symmetry (thin line) is also shown to illustrate the close association of MFE-23 Fv structures in the crystalline state. The positions of the six antigen-binding loops *H1*, *H2*, *H3*, *L1*, *L2* and *L3* show that they are fully exposed.

Table 7.1. Comparison of MFE-23 with five other single chain Fv crystal structures

Name	Specificity	Resolution (nm)	Domain orientation	Linker residues	Oligomer	PDB code	Reference
MFE-23	Carcinoembryonic antigen	0.24	V_H-V_L	$(G_4S)_3$	Monomer	n.a.	Present study
Se155-4	Salmonella liposaccharide	0.17	V_L-V_H	19-residue	Monomer	1mfa	Zdanov *et al.*, 1994
L5MK16	Phosphatidylinositol-specific phospholipase $C_{\delta 1}$	0.26	V_H-V_L	G_4S	Dimer	1lmk	Perisic *et al.*, 1994
NC10	Influenza virus silidase	0.30	V_H-V_L	$(G_4S)_3$	Dimer	n.a.	Kortt *et al.*, 1994
CC49/212	Carcinoma-specific trisaccharide	0.25	V_H-V_L	12-residue	Dimer	n.a.	Raag & Whitlow, 1995
C219	P-glycoprotein	0.24	V_L-V_H	17-residue (proteolytically removed)	Not known	1ap2	Hoedemaeker *et al.*, 1997

n.a., not available

Hoedemaeker *et al.*, 1997). All five were seen to be dimers in their crystal structures, although crystal packing considerations showed that Se155-4 was monomeric. This raised the question of whether MFE-23 is monomeric or dimeric in solution. This is important because two-domain Fv molecules possess improved functionality compared to four-domain Fab fragments for tumour targeting purposes (Yokota *et al.*, 1992). Neutron scattering and constrained curve fit modelling was used to resolve this question (Perkins *et al.*, 1998a). The solution structure of MFE-23 is presented, in which MFE-23 is monomeric and its V_H and V_L domains associate through their GFC faces.

The molecular basis of the interaction of MFE-23 with CEA is important for the improvement of tumour targeting strategies. As CEA is 50% carbohydrate by weight, it is unlikely that this can be crystallised intact, and alternative structural strategies are required. Previously, an outline solution structure for CEA was determined on the basis of the similarity of its domain structure with the V-set and C2-set domains in the cell surface proteins CD2 and CD4 (Bates *et al.*, 1992; Chapter 4). A reevaluation of the immunoglobulin superfamily resulted in the identification of an I-set domain which is intermediate in structure between those of the V-set and C1-set domains (Harpaz & Chothia, 1994; Chothia & Jones, 1997). The structure of an I-set domain is represented by crystal structures for the cell-surface protein, vascular cell adhesion molecule-1 (VCAM-1) (Jones *et al.*, 1995; Wang *et al.*, 1995, 1996), and intercellular cell adhesion molecule-1 (ICAM-2) (Casasnovas *et al.*, 1997). These developments mean that a more realistic full homology model can now be created for CEA, the domain arrangement of which can be validated using the X-ray and neutron scattering data (Chapter 4). The new CEA homology model together with the crystal structure for MFE-23 and the solution structures of CEA and MFE-23 leads to a structural model of the complex formed between CEA and MFE-23. The second set of lattice packing contacts seen in the MFE-23 crystal structure provides the basis of this model, as these show how the *H1*, *H2* and *H3* antigen binding loops of MFE-23 may interact with an immunoglobulin fold (Chapter 6). Using the two-domain crystal structures for CD2, CD4, ICAM-2 and VCAM-1 as structural templates, and homology models for the CEA domains, a prediction is made of an MFE-23 binding site on CEA. This strategy is a novel approach to the prediction of outline structures of antibody-antigen complexes, and the

putative binding-site could usefully form the basis of experiments for manipulating the interaction between CEA and MFE-23.

7.2. Materials and methods

7.2.1. Purification of MFE-23 and neutron scattering data

MFE-23 was purified as described in Chapter 6. For scattering work, MFE-23 was concentrated by ultrafiltration (PM10 membrane, Amicon), subjected to gel filtration on Sephacryl-S100 (Pharmacia), and concentrated (YM10 membrane, Amicon). For neutron scattering, samples were dialyzed into 0% or 100% ^2H_2O Dulbecco's phosphate-buffered saline with four buffer changes over 36 h at 6°C (137 mM NaCl, 0.5 mM $MgCl_2$, 2.7 mM KCl, 8.1 mM Na_2HPO_4, 1.5 mM KH_2PO_4, pH 7.4). The concentration of MFE-23 was determined from an absorption coefficient of 20.0 calculated from its sequence (1%, 280 nm, 1 cm path; Perkins, 1986), which is 40% higher than the value of 14.3 used by Verhaar *et al.* (1995).

Neutron scattering data using Instrument D22 at the neutron reactor at the Institut Laue-Langevin, Grenoble, France, which is analogous to Instrument D11 (Lindner *et al.*, 1992), were obtained at 15 °C using sample-detector and collimation distances of 5.6 m, a wavelength λ of 1.00 nm and a rectangular beam aperture of 7 × 10 mm. This gave a Q range of 0.07 to 0.8 nm^{-1} ($Q = 4 \pi \sin \theta / \lambda$; scattering angle = 2θ; wavelength = λ). Data acquisition times were typically 6 min in ^2H_2O buffers for MFE-23 samples at 1.1 to 3.7 mg/ml in rectangular quartz Hellma cells of path length 2 mm. Neutron data using Instrument LOQ at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, U.K. (Heenan & King, 1993) were also obtained at 15°C using a proton beam current of 190 mA to generate neutrons. Acquisitions were for 8×10^6 to 16×10^6 monitor counts lasting 70 to 140 min for protein concentrations of 8.0 to 1.5 mg/ml, also using 2 mm Hellma cells. Other details are given in Chapters 2 and 4 and Ashton *et al.* (1997).

Guinier analyses at low Q give the radius of gyration R_G and the forward scattering at zero angle $I(0)$ (Glatter & Kratky, 1982):

$$\ln I(Q) = \ln I(0) - R_G^2 Q^2/3.$$

This expression is valid in a $Q.R_G$ range up to 1.5. The R_G is a measure of structural elongation. The relative $I(0)/c$ values (c = sample concentration) for samples measured in the same buffer during a data session gives the relative molecular weights M_r of the proteins when referenced against a suitable standard (Wignall & Bates, 1987).

7.2.2. Homology modelling of CEA

Homology models for CEA-1 to CEA-7 (commonly designated as N, IA, IB, IIA, IIB, IIIA and IIIB in that order in the sequence) were constructed using the sequence alignment of Figure 7.2 and INSIGHT II, BIOPOLYMER, HOMOLOGY and DISCOVERY software (Biosym/MSI, San Diego, U.S.A.) on INDY Workstations (Silicon Graphics, Reading, U.K.). Loops were built using the pdb_select.1995-jun-01 database derived from 349 crystal structures at 0.2 nm resolution or better (Hobohm *et al.,* 1992; Hobohm & Sander 1994). Energy refinements were performed at the loop splice junctions, then on the sidechain atoms of all residues in the structurally conserved regions, and all the atoms of both types of loop residues. Energy refinements were based on the consistent valence force field. Iterations were made using the steepest descent algorithm to improve the connectivity of the model and minimize bad contacts or stereochemistry. The secondary structure backbone was retained by fixing the mainchain atoms in the conserved regions, and tethering these in the loop regions. Models were verified using PROCHECK (Laskowski *et al.*, 1993), and solvent accessibilities were calculated using a probe of 0.14 nm in COMPARER (Lee & Richards, 1971; Šali & Blundell, 1990). Secondary structures were identified using DSSP (Kabsch & Sander, 1983).

Residues 1-108 of CEA-1 were modelled using residues 4-104 in human CD2 (1hnf; Bodian *et al.*, 1994), residues 1-99 in human CD4 (3cd4; Wang *et al.*, 1990), and residues 2-114 in human CD8 (1cd8; Leahy *et al.*, 1992). The three reference structures were superimposed using structurally conserved β-sheet residues (Figure 7.2). Using the rigid body fragment assembly method, coordinates were assigned to eight structurally-conserved β-strand regions (37 residues) and four designated conserved loops (39 residues) using the reference structure with the most similar sequence (underlined in Figure 7.2a). Coordinates for five searched loops that correspond to

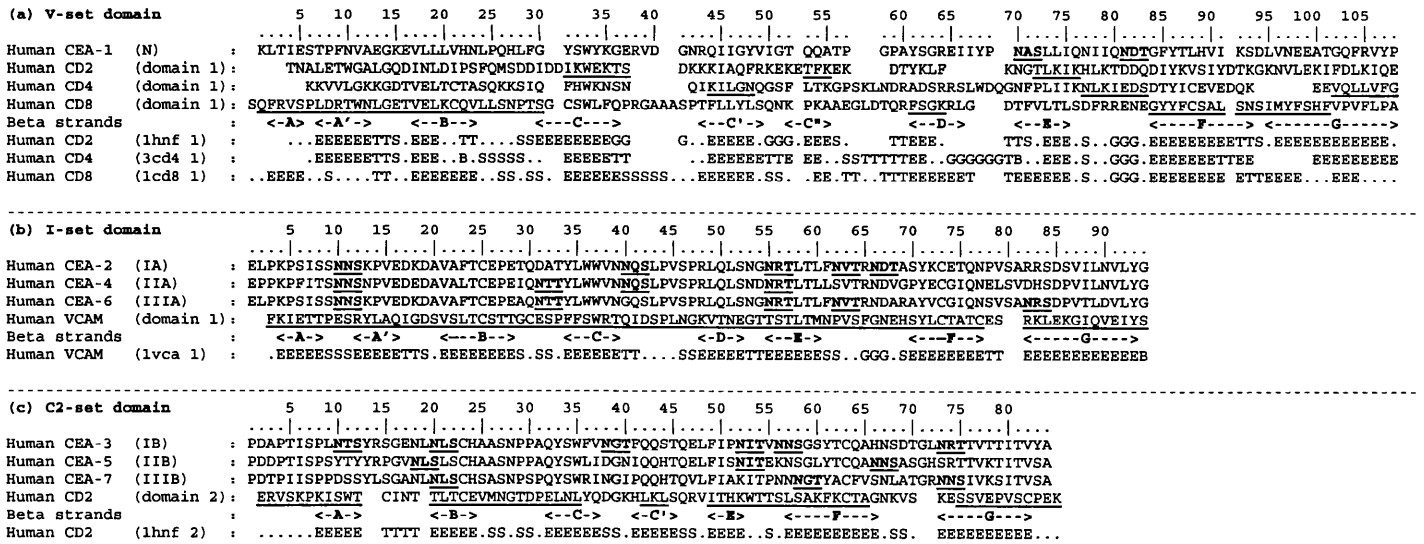


Figure 7.2. Sequence alignment used for the modelling of CEA based on homologous crystal structures. The 28 oligosaccharide sites in CEA are denoted by bold underlining. Sequences are labelled with the appropriate Brookhaven Protein Database code, and underlined regions indicate the structures used to construct the CEA homology models. The β-strands (E) identified by the DSSP program are labelled A to G. (*a*) The V-type domain of CEA is compared with the first V-type domains in CD2, CD4 and CD8. (*b*) The three I-type domains of CEA are compared with that found in the first domain of the vascular cell adhesion molecule-1 (VCAM-1). (*c*) The three C2-type domains of CEA are compared with that found in the second domain of the cell-surface protein CD2.

insertions or deletions (32 residues) were assigned from loop database searches (not underlined in Figure 7.2a).

Residues 3-94 of the CEA-2, CEA-4 and CEA-6 domains were modelled using residues 1-90 in human vascular cell adhesion molecule-1 (1vca; A-chain; Jones *et al.*, 1995), using 7 structurally conserved β-strand regions (38 residues), 7 designated loops (50 residues) and 1 searched loop (4 residues) that corresponded to an insertion at residues 78-81 in these CEA sequences (Figure 7.2b). Residues 2-84 of the CEA-3, CEA-5 and CEA-7 domains were modelled using residues 104-181 of human CD2 (1cd8; Leahy *et al.*, 1992), using 7 structurally conserved β-strand regions (29 residues), 6 designated loops (38 residues) and 2 searched loops that corresponded to short insertions or deletions (16 residues) in these sequences (Figure 7.2c).

The seven homology models were superimposed using INSIGHT II on the best-fit and CD2-derived CEA models derived by X-ray and neutron scattering curve fits (Chapter 4). The missing inter-domain linker residues (9 residues) were modelled as searched loops. Triantennary $Man_3GlcNAc_6Gal_3Fuc_3NeuNAc_1$ carbohydrate structures was added to each of the 28 putative N-linked sites on the CEA model as before (Chapter 4). Electrostatic surfaces were calculated using INSIGHT II and DELPHI software (Biosym/MSI, San Diego, U.S.A.). Red represents a potential of less than -4kT (acidic), blue a potential of more than +4kT (basic) and white as 0kT (neutral). Linear interpolation of the colours represents potentials between -4kT and +4kT.

7.2.3. Scattering curve modelling fits for MFE-23 and CEA

The crystal structure of MFE-23 was adapted for the presence of a 15-residue C-terminal peptide including the myc-tag by adding these as an extended β-strand using BIOPOLYMER. To make the Debye sphere models used to calculate scattering curves, the atomic coordinates of MFE-23 were placed in a three-dimensional grid of cubes of side 0.45 nm. A sphere of volume equal to a single cube was placed at the centre of each cube if a specified number of atoms were present in the cube. The cutoff was based on the requirement that the total volume of spheres was that of the dry protein volume of 34.8 nm^3 for all 256 residues of MFE-23 (Perkins, 1986). This meant that the volume

of the unknown coordinates for the $(Gly_4Ser)_3$ linker were included in the sphere model, e.g. the sphere model for the compact monomer contained 381 spheres. For models with a compact C-terminal myc-tag, the C-terminal peptide was removed, and the cube size was increased to 0.487 nm to maintain the correct volume, e.g. the sphere model for the compact monomer contained 302 spheres. Hydration is not detectable by neutron scattering and was not considered (Perkins *et al.*, 1998a). Scattering curves for comparison with experiment were calculated by the Debye equation assuming a uniform sphere scattering density in the program SCT (Perkins & Weiss, 1983; Chapter 3). This procedure has been tested with crystal structures in a molecular weight range of 23,000-127,000 (Smith *et al.*, 1990; Perkins *et al.*, 1993; Ashton *et al.*, 1997). A full-width-half-height wavelength spread of 10% for λ of 1.0 nm and a beam divergence of 0.024 radians were used to correct the calculated curve before fits to the D22 and LOQ data (Ashton *et al.*, 1997). The agreement between the modelled and experimental curves was determined using the R_G value derived from the calculated curve in the same Q range used for experimental Guinier fits and the R-factor for the Q range extending to 2 nm^{-1} (Smith *et al.*, 1990).

The two CEA homology models were converted into Debye sphere models using a cube side of 0.571 nm. This unhydrated sphere model consisted of 970 spheres and was used to calculate the neutron scattering curve. To account for the hydration shell observed in X-ray scattering experiments, the hydrated sphere model was prepared by uniformly expanding the dry model to correspond to the addition of a volume equal to a hydration of 0.3g H_20/g of glycoprotein and an electrostricted volume of 0.0245 nm^3 per bound water molecule (Perkins, 1986).

7.3. Results and discussion

7.3.1. Crystallographic dimer in the MFE-23 structure and five other scFv structures

All six scFv crystal structures solved to date (Table 7.1) showed dimeric molecules within their crystal lattice. Only the NC10 scFv structure corresponded to the same domain construction as that of MFE-23, with an N-terminal V_H domain joined to a C-terminal V_L domain by a $(Gly_4Ser)_3$ linker. The other four corresponded to both

types of domain orientations, in which the linkers were of lengths between 5-19 residues, and in one structure the linker was proteolytically removed. In antibody Fab fragment structures, the V_H and V_L domains interact through their GFC β-sheets in order to form the antigen binding site (Chothia *et al.*, 1985). In all six scFv structures, this association of V_H and V_L domains is observed. Except for the 5-residue linker in the L5MK16 scFv structure, electron density was not visible for these linkers for reason of disorder or flexibility. This agrees with NMR spectroscopy which showed that the $(Gly_4Ser)_3$ linker is more flexible than the remainder of the molecule (Freund *et al.*, 1993, 1994; Takahashi *et al.*, 1994). The lack of density for the linkers meant that the unambiguous identification of the multimeric state of the scFv molecules could not be made.

The crystallographically-observed dimer structure for MFE-23 is shown in Figure 7.1. The well-defined interactions responsible for dimer formation are reported in Chapter 6. Dimers could be formed in one of two ways. The V_H and V_L domains within the same scFv molecule can make face-to-face contact through their GFC faces with each other, so the dimer is formed by non-covalent interactions between two scFv molecules (upper or lower pairs in Figure 7.1). Alternatively, the two domains in scFv are dissociated from each other so that the V_H domain makes an extended link with its V_L domain, and consequently the dimer is formed from two extended monomers held together by two inter-scFv V_H-V_L domain interactions and two linkers (right or left pairs in Figure 7.1). The latter are known as diabodies. Such a dimer has been directly observed in the electron density map of L5MK16 which has a 5-residue linker that forced this arrangement. A crystal packing analysis of Se155-4 was carried out based on the length of its linker, and this showed that this had crystallised as a monomer which then formed noncovalent dimers in the crystal (Zdanov *et al.*, 1994). However solution data for NC10 by gel filtration and ultracentrifugation and for CC49/212 by chromatography showed that these could form dimers (Kortt *et al.*, 1994; Raag & Whitlow, 1995). This identification is not known for C219. For MFE-23, an identification is important because the six antigen binding loops are fully exposed at both ends of the dimer. If dimers existed, MFE-23 can form bivalent contacts with CEA (Figure 7.1). The rational planning of anti-tumour strategies is affected by whether

MFE-23 is monomeric or dimeric.

7.3.2. Monomeric MFE-23 structure by neutron scattering

Neutron scattering is advantageous in that it gives both molecular weight and structural information. It will identify both the oligomeric state of MFE-23 and its domain arrangement in solution. As radiation damage effects are frequently encountered for antibodies studied by synchrotron X-ray scattering (Beavil *et al.*, 1995), MFE-23 was studied by neutron scattering on Instruments LOQ and D22 to avoid this (Methods). Using a MFE-23 concentration range of 1.1 to 8.0 mg/ml, linear Guinier plots were obtained in an appropriate $Q.R_G$ range from 0.6 to 1.1 (Figure 7.3a). The R_G and $I(0)/c$ values derived from these showed a linear concentration dependence (Figures 7.3b and 7.3c). On extrapolation to zero concentration, MFE-23 had an R_G value of 1.88 ± 0.03 nm and an $I(0)/c$ value of 0.0294 ± 0.0007 relative to a polymer standard. By reference to a linear calibration graph of 12 pairs of $I(0)/c$ and M_r values for Instrument LOQ (Figure 2.10; Chapter 2), the $I(0)/c$ value was determined to correspond to a M_r of 27,300 ± 1,200. This is in close agreement with the M_r of 27,200 calculated from its sequence. Figure 7.3(c) showed that MFE-23 was monomeric below 1 mg/ml, and became oligomeric at higher concentrations.

Curve modelling fits constrained by the MFE-23 crystal structure identified the MFE-23 domain arrangement in solution. As the hydration shell surrounding proteins is invisible by neutron scattering, this permitted the direct comparison of the neutron data with the crystal structure without the need to add water molecules to it (Perkins, 1986; Svergun *et al.*, 1998; Perkins *et al.*, 1998a). Three possible alternative domain arrangements for MFE-23 exist based on the crystal packing. Two monomers were created from a compact and a dissociated, extended V_H-V_L domain arrangement (upper pair and left pair of domains in Figure 7.1), in which the GFC faces either interact or are freely exposed respectively. The latter would be significant as it would not have a six-loop combining site and would have reduced antigen-binding affinity. The dimer was created from all four domains. Since the myc-tag is not visible in the electron density map, the solution structure of the myc-tag is assumed to be flexible and adopt a range of conformations. The 15-residue C-terminal peptide was therefore represented by

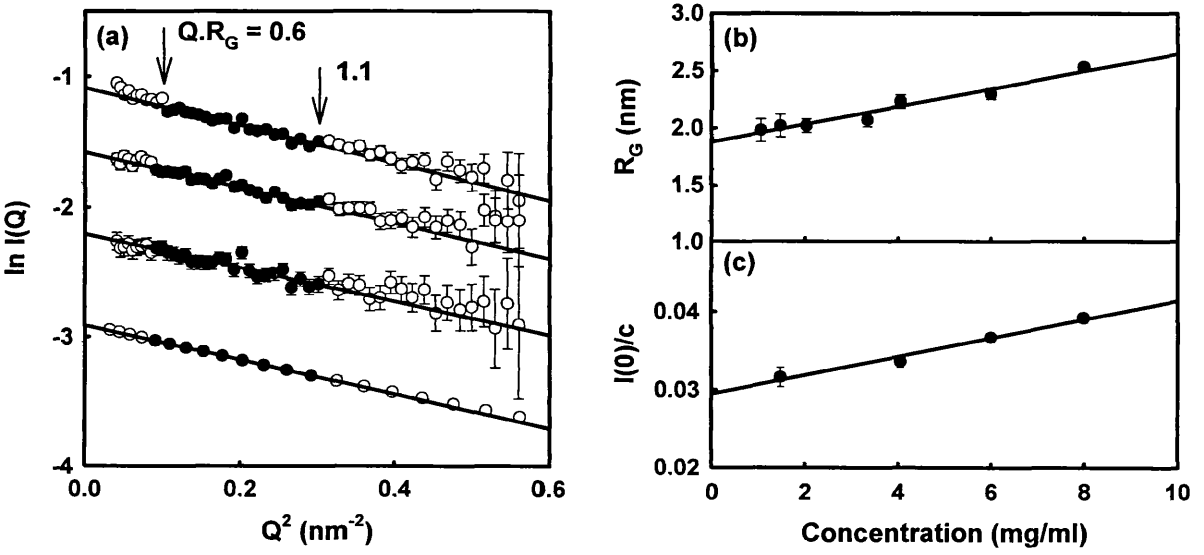


Figure 7.3. Neutron Guinier analyses for MFE-23. (*a*) Guinier R_G plots for MFE-23 concentrations of 3.4, 2.0 and 1.1 mg/ml measured on Instrument D22, and the compact Fv model in Figure 3a. The filled circles show the data used to determine R_G values in a Q range of 0.30 to 0.55 nm^{-1}. (*b*) Concentration dependence of the experimental R_G values for MFE-23 between concentrations of 1.1 and 8.0 mg/ml. Linear regression gave an R_G value of 1.88± 0.03 nm for MFE-23 at zero concentration. (*c*) Concentration dependence of the $I(0)/c$ values for MFE-23 concentrations between 1.5 and 8.0 mg/ml. Linear regression gave an $I(0)/c$ of 0.0294 ± 0.0007 at zero concentration.

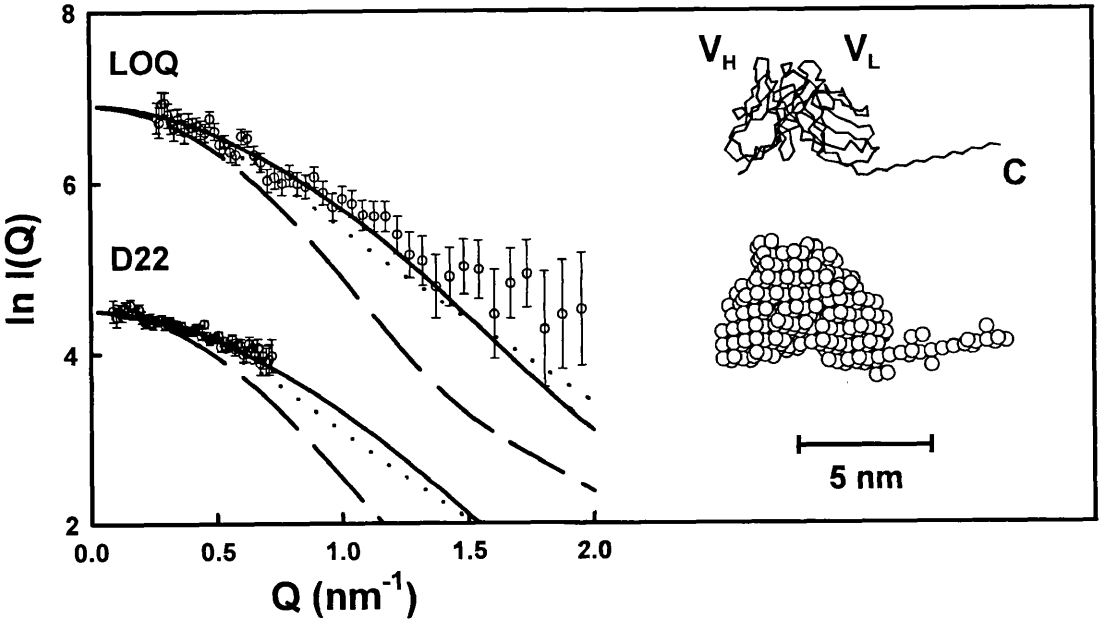


Figure 7.4. Scattering curve fits from molecular models for MFE-23. The experimental MFE-23 neutron data was obtained at a concentration of 1.5 mg/ml using LOQ (○) and 1.1 mg/ml using D22 (○). The calculated scattering curves for MFE-23 correspond to the compact Fv monomer (continuous line: two upper or lower domains in Figure 7.1), the Fv dimer (dashed line: all four domains in Figure 7.1), and the elongated monomer (dotted line: two left or right domains in Figure 7.1). Compact Fv monomer: R_G = 2.00 nm; *R*-factor = 9.5%, dimer: R_G = 2.53 nm; *R*-factor = 25.0% and elongated Fv monomer: R_G = 2.21 nm; *R*-factor = 12.1%. The compact Fv model is shown as an α-carbon trace and as a sphere model, both with an extended C-terminal myc-tag.

maximally extended or compact conformations to give the maximum range of structures for each arrangement of domains. The extended myc-tag was modelled as an extended β-strand at the V_L C-terminus (Figure 7.4; Methods). The compact myc-tag was modelled from the MFE-23 coordinates without a myc-tag, and adjusting the sphere conversion so that additional spheres to correspond to the volume of the myc-tag were positioned close to the surface of MFE-23 (Methods). Guinier fits of the three scattering curves calculated for the extended myc-tag models gave R_G values of 2.00 nm and 2.20 nm for the compact and extended monomers and 2.52 nm for the dimer. Those for the compact myc-tags R_G values of 1.80 nm, 2.12 nm and 2.39 nm in that order. The observed R_G value of 1.88 nm agreed best with the R_G range of 1.80-2.00 nm for the compact monomer in both myc-tag conformations. This agreement showed that this was the solution structure of MFE-23 at 1 mg/ml. Interestingly, MFE-23 was crystallised at this concentration (Chapter 6).

To confirm the Guinier analysis, the calculated scattering curves from the three models were compared with the experimental scattering curve in the Q range up to 0.2 nm^{-1} (Figure 7.4). The calculated curves were visibly different from each other. That from the compact monomer model gave the best agreement for LOQ and D22 data, despite the weak counting statistics caused by the need to work below 1 mg/ml concentration. The compact monomer model gave the best R-factor. The R-factor goodness of agreement was 9.5%, 12.1% and 25.0% for the compact and extended monomers and the dimer curve fits respectively, all with extended myc-tag models, using the LOQ data at 1.5 mg/ml MFE-23.

Further curve-fit calculations for MFE-23 at higher concentrations were difficult to interpret. For example, the modelled R_G values of dimeric MFE-23 were 2.39-2.52 nm (above). According to Figure 7.3(c), the molecular weight had only increased by 25% at these R_G values. This showed that oligomer formation of MFE-23 involved extended associations of MFE-23 molecules. This may affect the use in patients of MFE-23 at high concentrations.

7.3.3. Homology modelling of CEA

CEA had previously been modelled as seven V-set and C2-set Ig folds in the order V-C2-C2-C2-C2-C2-C2 (Williams & Barclay, 1988; Bates *et al.*, 1992; Chapter 4). This has recently been reclassified as a V-I-C2-I-C2-I-C2 arrangement (Harpaz & Chothia, 1994) and I-set crystal structures have now appeared (Jones *et al.*, 1995; Wang *et al.*, 1995, 1996; Casasnovas *et al.*, 1997). The I-set fold is intermediate between the V-set and the C1-set Ig folds. Its β-sheet structure contains DEBA and A'GFCC' β-sheets, in distinction to V-set folds with DEBA and A'GFCC'C" β-sheets, C1-set folds with DEBA and GFCC' β-sheets, and C2-set folds with EBA and GFCC' β-sheets (Chothia & Jones, 1997; Bork *et al.*, 1994). The CEA-2, CEA-4 and CEA-6 domain sequences were readily aligned with that of vascular cell adhesion molecule-1 (VCAM-1), whose I-set structure lacks β-strand C' (Figure 7.2). Unlike the previous C2-set alignments (Bates *et al.*, 1992; Chapter 4), there were almost no gaps or insertions. The sequences identities were low at 16-20%, but this increased to 34-37% when residue similarity was considered (defined in the legend to Figure 4.6; Chapter 4). The CEA-3, CEA-5 and CEA-7 domain sequences showed sequence identities of 12-15% with CD2, which increased to 28-35% when residue similarity was considered. These improvements lead us to construct full homology models for the seven CEA domains (Methods).

Previously, the V-set domain from the crystal structure of human CD2 had been used to represent the V-set CEA-1 domain (Chapter 4). The homology modelling of this domain was complicated by insertions or deletions in most of the loop regions (Figure 7.2a). Accordingly it was decided to use the V-set domain structures from human CD2, CD4 and CD8 as templates. Their conserved β-strand residues were superimposed using the conserved Trp residue in β-strand C as an initial reference point (Figure 7.2a; Methods). This revealed structural differences in the lengths and sequences of corresponding β-strands and loops. First, the CEA-1 sequence was aligned with those in the three template structures, and structurally conserved regions were defined using the conserved β-strands. Next, for each of the ten β-strands and for each conserved designated loop, the template structure with the highest sequence similarity to CEA-1 was used in the CEA-1 homology model (Figure 7.2a). Five loops were reconstructed

from database searches, all of which were located at the exposed N-terminal end of the CEA-1 domain (i.e. the antigen-binding site in antibody V-set domains) or the exposed edge C" β-strand in the A'GFCC'C" β-sheet. This was performed because all five loops displayed structural variability between the three templates. This variability may reflect the characteristic functional properties of V-set domains in different cell-surface proteins.

Previously, the C2-set domain from the crystal structure of human CD2 had been used to represent the CEA-2 to CEA-7 domains (Chapter 4). While this CD2 domain of length 78 residues is similar in size to the CEA-3, -5 and -7 domains of length 84 residues each, it is significantly shorter than the CEA-2, -4 and -6 domains of length 95 residues each. The replacement of C2-set domains by I-set domains from VCAM-1 in homology models for domains CEA-2, -4 and -6 required only one insertion in the loop between β-strands F and G. This was modelled using a database search (Figure 7.2b). The effect of moving to an I-set fold is to position residues now assigned to β-strands A' and C' on the opposite side of the β-sandwich, which affects the position of oligosaccharides at Asn118, Asn297 and Asn 474. Homology models of CEA-3, -5 and -7 were constructed using the C2-set domain of CD2. Two insertions were required in the loops between β-strands A and B and β-strands F and G, and these were modelled using database searches (Figure 7.2c).

7.3.4. Scattering curve fits for the CEA homology model

CEA had been determined to be monomeric in solution (Chapter 4). In that study, the domain arrangement of CEA was derived from the combination of X-ray and neutron scattering curve fits with the CD2 coordinates for one V-set and six C2-set domains and coordinates for 28 oligosaccharides. An automated search for domain arrangements that were consistent with the CEA scattering curves had been based on an optimisation of the average interdomain rotation between two neighbouring domains. Only elongated zig-zag domain arrangements fitted the data. The best-fit CEA model from that analysis gave an averaged interdomain orientation that resembled the orientation between the two domains in the CD2 crystal structure (Jones *et al.*, 1992; Bodian *et al.*, 1994; Table 4.2; Chapter 4). This showed that two new full CEA

homology models could be created by superimposing the seven new homology models for the CEA domains on the CD2-domains of both the best-fit CEA and the CD2-derived CEA model described in Chapter 4. The peptide connections between the domains in each model were created using database searches of loop structures. Oligosaccharide chains were added in the extended conformations that were identified by the scattering analyses.

The domain arrangement in the two new CEA homology models were tested using scattering curve fits using the X-ray and neutron data from Chapter 4. The fit procedure was improved by the use of an optimised set of instrumental corrections for neutron fits (Ashton *et al.*, 1997). Previously, in reflection of the 50% carbohydrate content in CEA, two-density sphere modelling had been used to fit the X-ray and neutron data. Here, the use of new homology models gave improved X-ray and neutron curve fits using a single density sphere model for CEA. In effect, this meant that the presence and absence of a hydration shell was sufficient to account for the differences between the two scattering curves. The outcomes for both new CEA homology models are now summarised:

(a) For the best-fit model (Figure 7.5), the modelled X-ray R_G value was 7.9 nm, in good agreement with the experimental X-ray R_G value of 8.0 nm. The modelled X-ray R_{XS} value was 1.9 nm, in good agreement with the experimental X-ray R_{XS} value of 2.1 nm. The X-ray and neutron *R*-factors were reduced to 4.3% and 8.2% respectively in place of the values of 4.7% and 8.7% reported previously for the two-density model.

(b) For the CD2-derived model (Figure 7.6), the modelled X-ray R_G value was also 7.9 nm. The modelled X-ray R_{XS} value was 2.0 nm, in closer agreement with the experimental X-ray R_{XS} value of 2.1 nm than the best-fit model. The X-ray and neutron *R*-factors were 4.4% and 8.1% and also represented an improvement compared to earlier work.

Even though scattering does not result in unique structure determinations, but only those that are consistent with the scattering curve, both sets of agreements confirmed the previously-determined extended zig-zag domain arrangement of CEA. Other possible structure arrangements had previously been ruled out (Chapter 4).

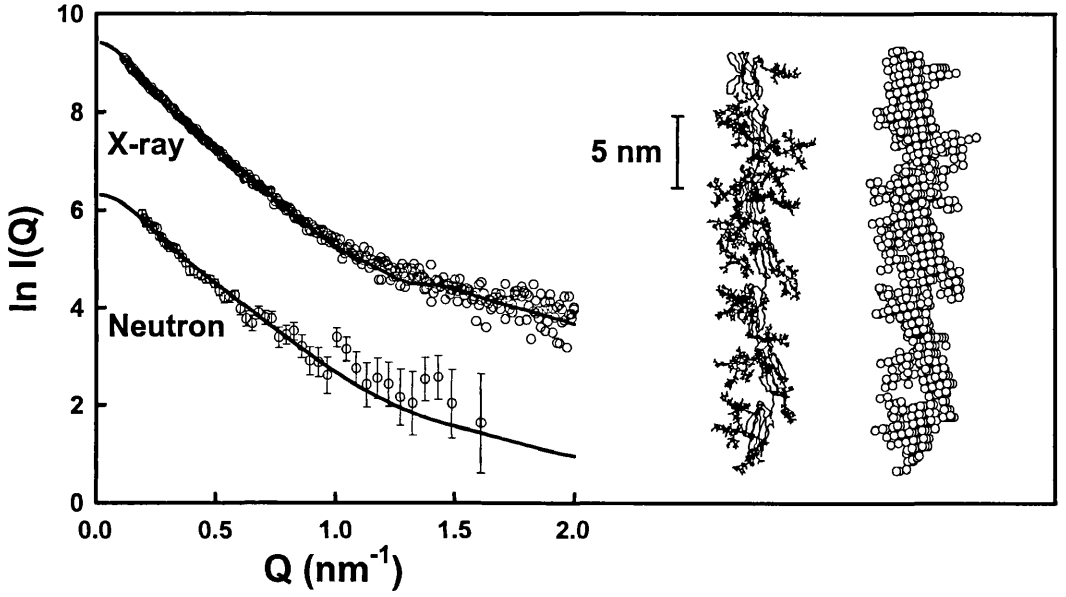


Figure 7.5. Comparison of the simulated X-ray and neutron scattering curves for a homology model of CEA based on the best-fit model of Chapter 4 with experimental X-ray and neutron data. The seven CEA domains in each model are shown as α-carbon traces, whereas the carbohydrate chains are represented in full. The corresponding sphere model (sphere diameter, 0.571 nm) is also shown. The calculated neutron curve was corrected for wavelength resolution and beam divergence. For the experimental X-ray curve (○), the R_G value is 7.9 nm and the $R_{2.0}$ value for the X-ray data is 4.3%. For the experimental neutron curve (○), the $R_{1.7}$ value is 8.2%.

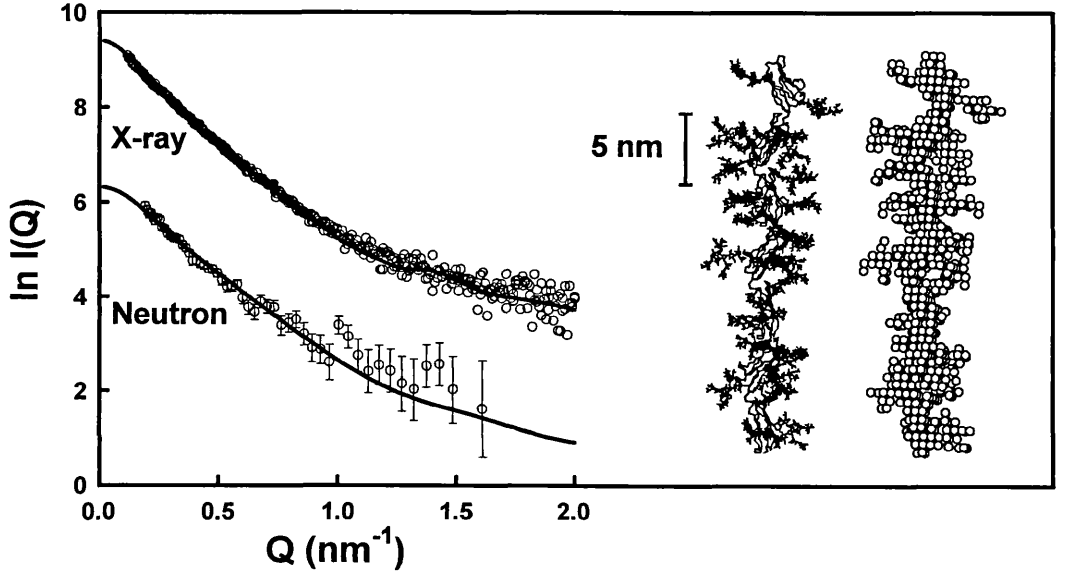


Figure 7.6. Comparison of the simulated X-ray and neutron scattering curves for the homology model of CEA based on the CD2-derived model of Chapter 4 with experimental X-ray and neutron data. The seven CEA domains in each model are shown as α-carbon traces, whereas the carbohydrate chains are represented in full. The corresponding sphere model (sphere diameter, 0.571 nm) is also shown. The calculated neutron curve was corrected for wavelength resolution and beam divergence. For the experimental X-ray curve (○), the R_G value is 7.9 nm and the $R_{2.0}$ value for the X-ray data is 4.4%. For the experimental neutron curve (○), the $R_{1.7}$ value is 8.1%.

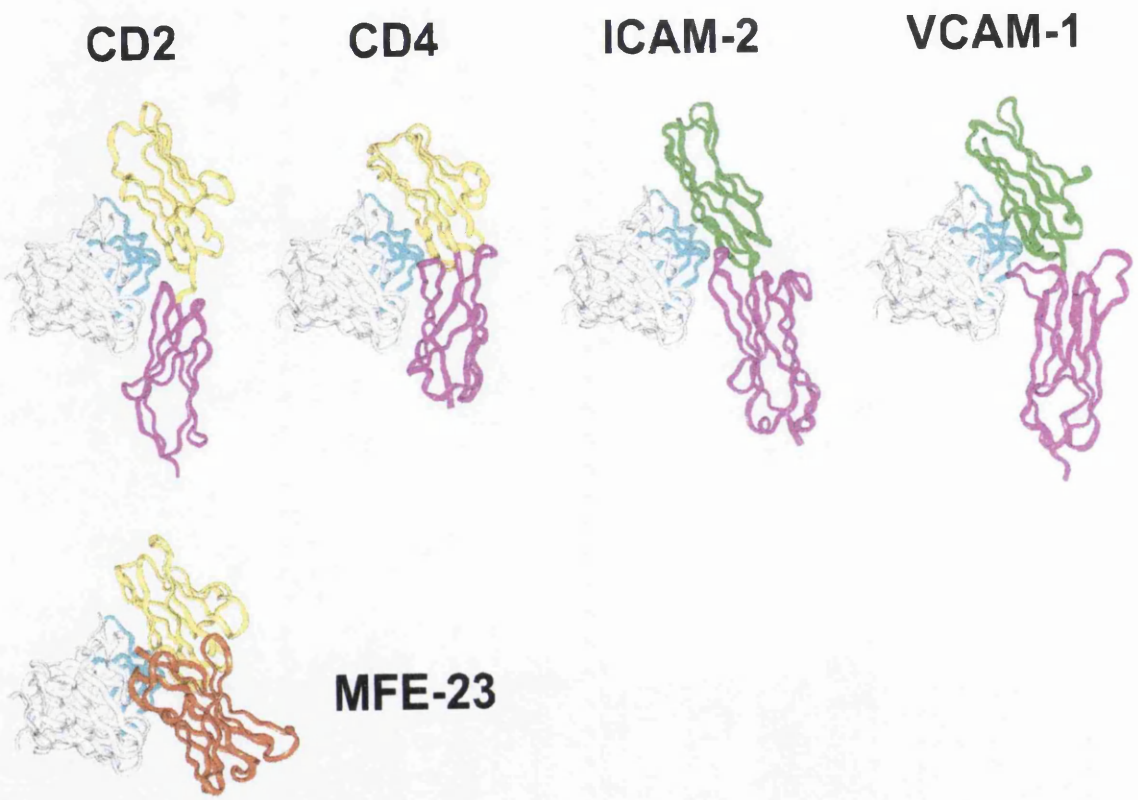


Figure 7.7. Hypothetical complexes formed between MFE-23 and the two-domain cell-surface proteins CD2, CD4, ICAM-2 and VCAM-1 (PDB codes 1hnf, 3cd4, 1zxq, 1vca-A respectively). The observed lattice contacts between the antigen-binding loops *H1*, *H2* and *H3* (cyan) of the V_H domain of MFE-23 with the DEBA face of the adjacent V_L domain in the crystal lattice was used to position MFE-23 in relation to the N-terminal V-type domains of these four cell-surface proteins. MFE-23 V_H and V_L domains white; V-type domains: yellow; I-type domains: green; C2-type domains: purple. For comparison, the packing of MFE-23 in its crystal lattice against the V_L (yellow) and V_H (orange) domains of two different adjacent MFE-23 molecules is illustrated to show how the V_H domain is antiparallel and deviates from the packing seen for the cell-surface proteins.

7.3.5. Modelling of the interaction between MFE-23 and CEA

Given the present experimental determinations of monomeric solution structures for MFE-23 and CEA above, a model could now be constructed for the complex formed between MFE-23 and CEA. The crystal structure of MFE-23 showed a second type of lattice packing which is characterised by the interaction of the antigen-binding loops *H1*, *H2* and *H3* in the V_H domain of MFE-23 with the V_L domain of a neighbouring MFE-23 molecule. This was predicted to provide a good model for the interaction of MFE-23 with the immunoglobulin folds in CEA (Chapter 6). This is exemplified in Figure 7.7 where an MFE-23 molecule and the V_L and V_H domains of two different adjacent MFE-23 molecules are shown. It is clear that the adjacent V_H domain (orange) does not interact with the MFE-23 antigen-binding loops. This is consistent with the back-to-back antiparallel pairing of the adjacent V_L and V_H domains in the crystal lattice (Figure 7.1), since the seven domains in CEA form a parallel arrangement.

The interaction between MFE-23 and CEA was investigated by superimpositions of the adjacent V_L domain in the MFE-23 lattice onto the N-terminal domain of the related two-domain crystal structures of CD2, CD4, ICAM-2 and VCAM-1 (Bodian *et al.*, 1994; Wang *et al.*, 1990; Casasnovas *et al.*, 1997; Jones *et al.*, 1995). This presumed that the crystallographic N-terminal domains each represented a CEA domain. The superimpositions were performed using residues belonging to the invariant immunoglobulin β-strands B, C, E and F. Figure 7.7 showed that, in all four cases, MFE-23 was positioned within an angled surface formed by the two domains. While the C-terminal domain of CD2, CD4, ICAM-2 and VCAM-1 was positioned close to the *L1*, *L2* and *L3* antigen-binding loops of MFE-23, that of CD2 presented the most favourable contacts with MFE-23, followed by CD4. Figure 7.7 also showed that the *H1*, *H2* and *H3* loops of MFE-23 were inserted into the angle formed between the lower part of the DEBA face of the N-terminal domain and the upper part of the GFC face of the C-terminal domain in all four crystal structures. The complementary nature of the MFE-23 and CD2 surfaces showed that this interaction could indeed be modelled by reference to the observed lattice contacts with the V_L domain in the MFE-23 crystal structure.

This proposed interaction between MFE-23 and CD2 was adapted to that in CEA by replacing the two CD2 domains by the homology models for CEA-1 and CEA-2. The replacement used the CD2 crystal structure in preference to the CD4, ICAM-2 and VCAM-1 structures on the grounds that (a) this formed the most favourable interaction with MFE-23 (Figure 7.7); (b) the linker peptide between the CD4 domains was not long enough when compared with those joining CEA-1 and CEA-2; and (c) the domain pair in ICAM-2 and VCAM-1 corresponded to an I-C2 arrangement in which the first domain is not a V-set one and the second one is unusual in that it is about 20% larger than a typical C2-set fold (Figure 7.7). The accuracy of each domain homology model is expected to be within an r.m.s. deviation of 0.173 nm in the α-carbon positions for satisfactory homology modelling (Sutcliffe *et al.*, 1987). This will be higher at β-strand positions in the protein core, and reduced in the rebuilt loop regions. The position of the superimposed CEA-1 domain will be more accurately known than that of the adjacent attached CEA-2 domain. In addition there may be conformational changes of the order of 0.1-0.15 nm that involve the *H3* loop (Stanfield *et al.*, 1990; Wilson & Stanfield, 1994). Nonetheless it is expected that the modelling is sufficiently accurate to assess an outline structure of the complex between MFE-23 and CEA.

Evidence to support the model of this complex was obtained from the calculation of electrostatic surfaces for CEA and MFE-23 and the examination of Asp, Glu, Lys and Arg residues in both proteins using stereo glasses. The MFE-23 crystal structure showed that the *H1* and *H2* loops in the antigen binding site of MFE-23 contain negatively-charged residues at Asp-H31, Asp-H52, Glu-H54, Asp-H57 and Glu-H59, four of which are highly exposed. Together with Glu-L1 in the MFE-23 framework, these form a continuous stripe of acidic residues across the surface of MFE-23 (Figure 7.8b). The CEA homology model showed that there is a corresponding stripe of basic residues (arrowed in Figure 7.8a) across the CEA-1 and CEA-2 domains. This is formed by Lys15 and Arg64 on the DEBA face of CEA-1 and by Lys112, Lys180, Arg190 and Arg191 on the GFC face of CEA-2. The α-carbon positions of these six basic residues in the CEA homology model were well placed to form potential ionic interactions with the above six acidic residues in the MFE-23 crystal structure. The proposed participation of Lys15 and Arg64 in CEA-1 in forming ionic interactions with the *H2*

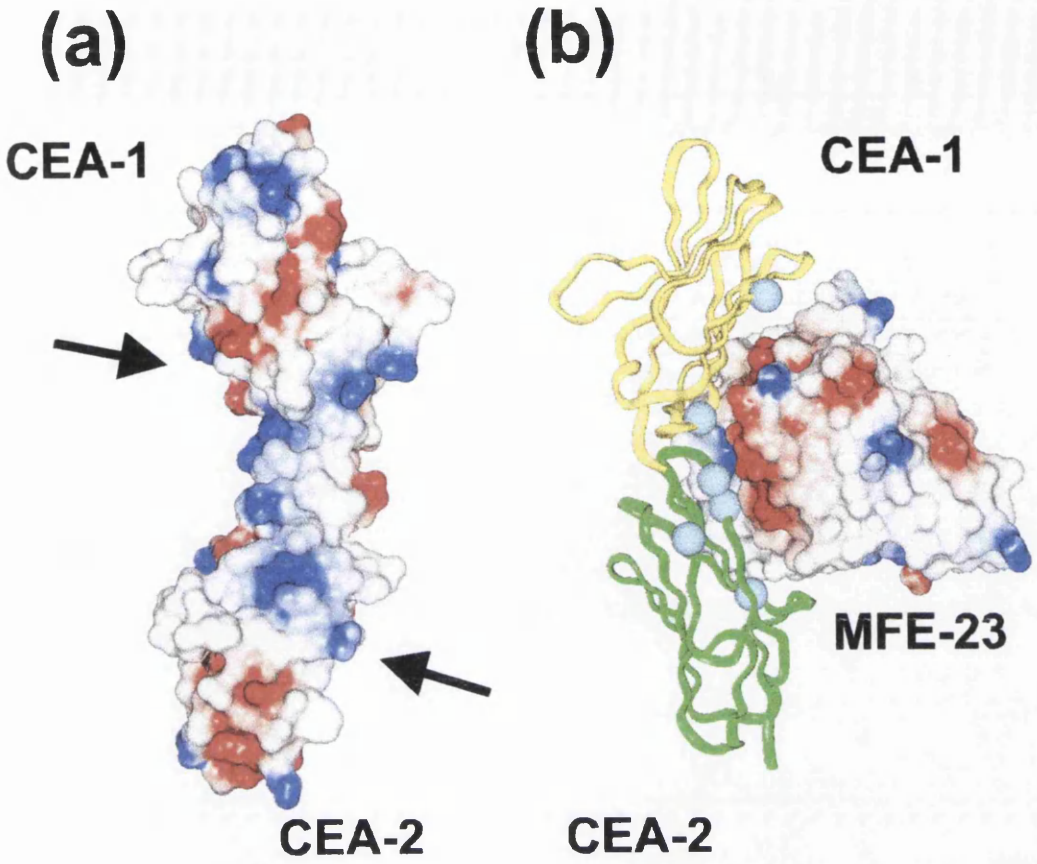


Figure 7.8. Electrostatic maps of CEA-1 and CEA-2 and the antigen binding site of MFE-23.

(*a*) The CEA-1 and CEA-2 domains are built in the same orientation as the two domains of human CD2 in Figure 6. A stripe of basic residues at the interface between the CEA-1 and CEA-2 domains is arrowed. This involves Lys15, Arg64, Lys112, Lys180, Arg190 and Arg191.

(*b*) The α-carbons of these six basic residues are shown as blue spheres in the ribbon view of the CEA homology model (V, yellow; I, green) which is rotated by 180° about the vertical axis relative to the electrostatic view of the CEA-1 and CEA-2 domains. These basic residues are complementary to a stripe of acidic residues seen on the MFE-23 electrostatic surface. These acidic residues are located on the *H1* and *H2* loops of MFE-23 and at the N-terminus of the V_L domain, and involve Asp-H31, Asp-H52, Glu-H54, Asp-H57, Glu-H59 and Glu-L1.

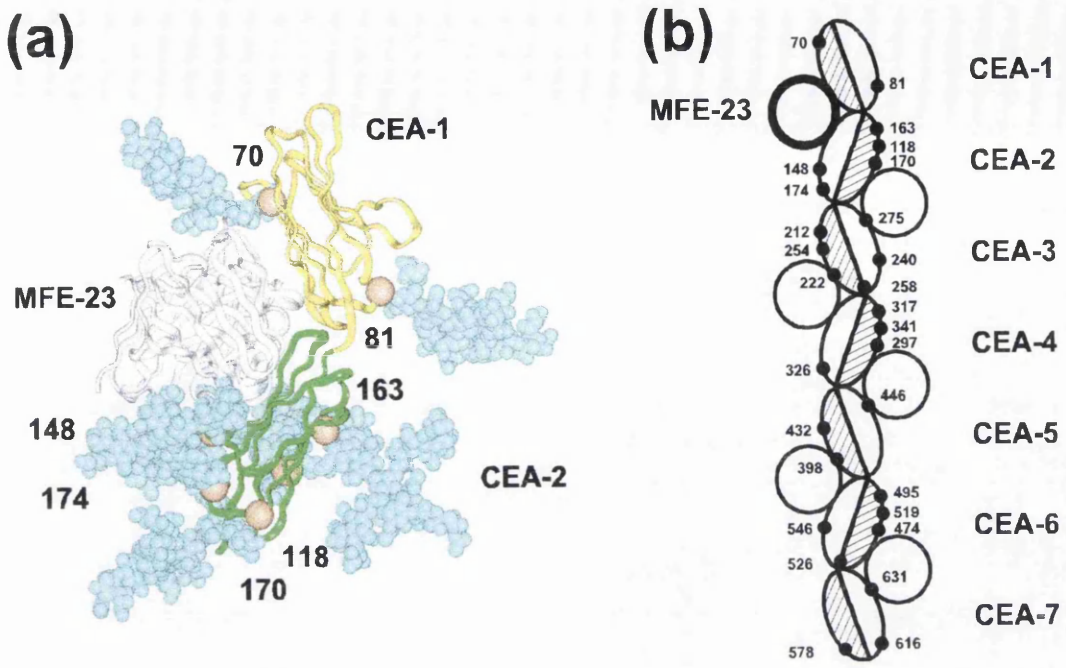


Figure 7.9. Comparison of the oligosaccharide arrangement in CEA relative to the modelled MFE-23 binding site on CEA.

(*a*) A ribbon diagram of MFE-23 (white) is shown attached to the CEA-1 and CEA-2 domains (yellow and green respectively), in which the CEA-1 and CEA-2 domains are built in the same orientation as the two domains in CD2 (Figure 6). Seven oligosaccharide chains at Asn70 and Asn81 in CEA-1 and at Asn118, Asn148, Asn163, Asn170 and Asn174 in CEA-2 are shown in blue solid representations. These do not prevent MFE-23 binding at the interface between the CEA-1 and CEA-2 domains.

(*b*) A schematic view of the seven CEA domains together with MFE bound at all six possible interfaces between adjacent domains. The CEA domain orientation follows that in Figure 5. The preferred MFE-23 binding site is shown in bold. The DEBA face of each CEA domain is shown hatched, and the GFC face is shown open. The position of the 28 carbohydrate sites is shown by filled symbols (●). Each site is identified by its Asn residue number. The number is positioned outside an MFE-23 circle if the Asn residue offers no steric hindrance to MFE-23 binding, and inside the circle if there is steric overlap. For the CEA model in Figure 5, only the preferred MFE-23 molecule is able to bind to CEA without steric overlap with carbohydrate.

loop of MFE-23 is analogous to the role of Lys-L18 and Arg-L76 in forming lattice contacts with Asp-H52 and Glu-H54 in the MFE-23 crystal structure (Figure 6.14; Chapter 6). Similar salt bridges have been observed to be formed between two V_H glutamates and antigen Arg residues to stabilize the antibody-antigen complexes in two crystal structures (Braden & Poljak, 1995; Davies & Cohen, 1996)

Further support for the model of the complex was obtained from the positions of the 28 carbohydrate chains in CEA. The scattering modelling showed that these possess structures that are on average extended freely into solution (Chapter 4). Complex formation between MFE-23 and CEA required the absence of steric conflict with carbohydrate chains at the protein surface, where in particular glycosylated Asn residues have to be absent from the MFE-23 binding site on CEA. In relation to two such sites on CEA-1 and five more on CEA-2, Figure 7.9(a) showed that the model of the MFE-23 complex with CEA-1 and CEA-2 fully satisfied this requirement. One site in CEA-1 is located at Asn70 at the tip of the domain from where the carbohydrate can project freely into solution, while the other at Asn81 is located at the base of the GFC face on the reverse side of CEA-1 to where MFE-23 binds in the complex. The five sites in CEA-2 at Asn118, Asn148, Asn163, Asn170 and Asn174 are all located around the base of the I-set domain, from which a ruff of carbohydrate chains project into solution. The seven chains do not conflict with the modelled MFE-23 binding site between CEA-1 and CEA-2, although interactions between the oligosaccharide chains at Asn148 and Asn174 and MFE-23 cannot be ruled out. Figure 7.9(a) also indicated the complementary nature of the MFE-23 and CEA surfaces, where MFE-23 is inserted into the junction between the CEA-1 and CEA-2 homology models in the same manner as that observed for CD2 and CD4 (Figure 7.7).

Analysis of the remaining carbohydrate sites on CEA did not favour the binding of MFE-23 at five other possible locations that are formed at the angle between the lower part of a DEBA face and the upper part of a GFC face. Of the 28 sites, 16 could be located to the DEBA faces, and 10 to the GFC faces (Figure 7.9b). Using the full CEA homology model of Figure 7.6, the MFE-23 structure was duplicated five times as indicated in Figure 7.9(b). The relationship of each MFE-23 structure to nearby

carbohydrate chains was checked using stereo glasses. This showed that a second potential MFE-23 location between CEA-2 and CEA-3 would be blocked by Asn275, a third location would be blocked by Asn222, a fourth location would be blocked by Asn446, a fifth location would be blocked by Asn398, and a sixth location would be blocked by Asn631. In fact, the orientation of these 5 carbohydrate chains in CEA at the angled interface between five pairs of domains suggested that these may help to maintain CEA as an extended zig-zag structure. Two further carbohydrate chains at Asn578 and Asn616 at the base of CEA may act to buttress the elevation of CEA vertically away from the cell surface. The overall effect of the location of carbohydrate in CEA is to leave the protein surfaces of CEA-1 and the upper half of CEA-2 largely exposed. Since the carbohydrate appears to buttress the domain arrangement of CEA, it would appear that the role of the seven domain structure of CEA is to present highly exposed biologically active surfaces at the cell surface. The immunoglobulin superfamily forms the dominant component of cell surface proteins, and many of these are the one-, two- or four-domain high CD8, CD2 and CD4 cell surface proteins. In comparison, CEA projects well beyond these (Figure 7.10).

7.4. Conclusions

7.4.1. Solution structures of MFE-23 and CEA

An understanding of the complex between MFE-23 and CEA is dependent on knowledge of the structures of the free proteins in solution. The crystal structure of MFE-23 showed that this was a dimeric four-domain structure, however neutron scattering showed that it is a monomeric two-domain protein in solution under the conditions used therapeutically, and that its domain arrangement corresponds to the classic interaction between two GFC faces that is seen in two-domain Fv fragments. Neutron scattering is advantageous in that gel filtration methods for determining oligomerisation are qualitative at best. Another key benefit of the neutron work included the demonstration that the correct molecular weight of MFE-23 is obtained if the absorption coefficient for MFE-23 at 280 nm (1%, 1 cm) is 20.0 as predicted from its sequence and not that of 14.3 used previously (Verhaar *et al.*, 1995). These results illustrate the importance of solution scattering methods to confirm and extend new results obtained from crystallography (Perkins *et al.*, 1998a).

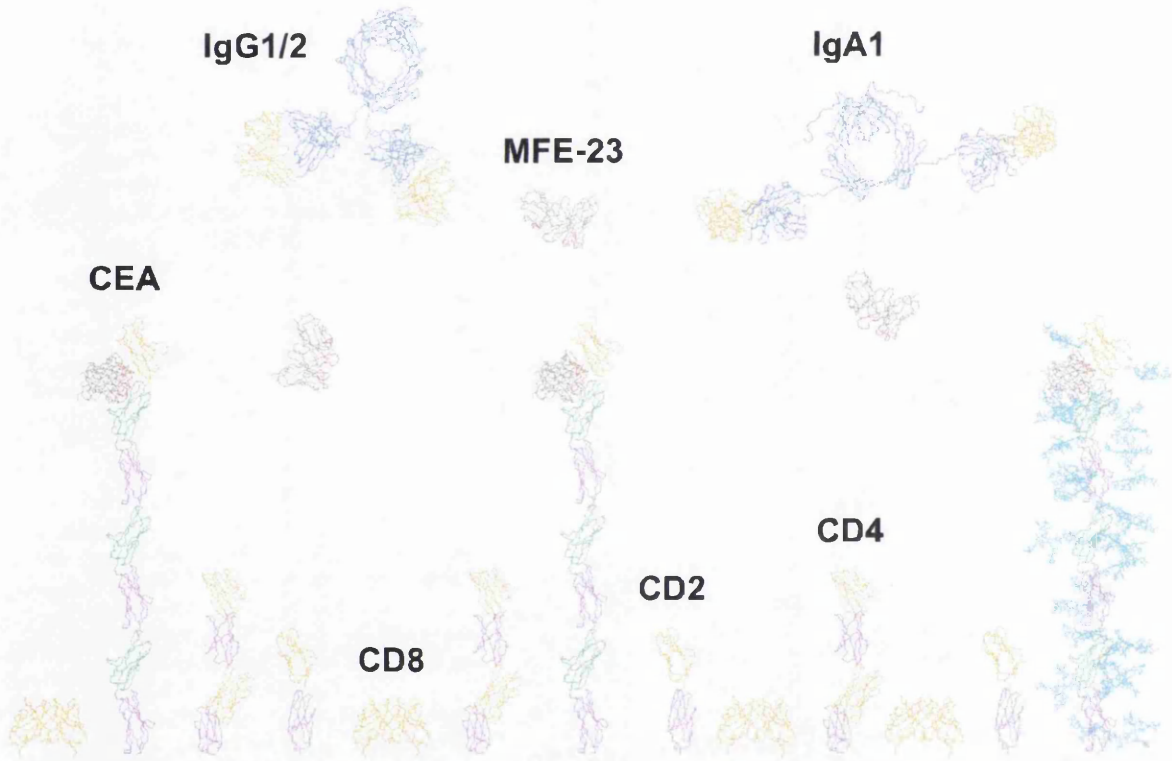


Figure 7.10. Association of MFE-23 with CEA on a cell surface. The sizes of several CD8, CD2 and CD4 molecules (PDB codes 1cd8, 1hnf, 1cid and 3cd4) are shown as α-carbon views relative to three molecules of CEA with and without its 28 carbohydrate chains (cyan). MFE-23 molecules (black, with red antigen sites) are shown attached to its modelled binding site on CEA. Representations of the averaged solution structures of the 12-domain antibodies IgG1/2 and IgA1 are shown (Mayans *et al.*, 1995; Chapter 5). The domain types are indicated as follows: V, yellow; I, green; C1, blue; C2, purple.

The solution structure of CEA is an extended zig-zag arrangement of seven domains (Chapter 4). This was reconfirmed in this study by the incorporation of full homology models for the V-, I- and C2-set domains into the previous scattering models to follow Harpaz & Chothia (1994), whereupon improved curve fits were obtained. The scattering modelling is precise enough to identify a zig-zag domain arrangement in CEA and its extended carbohydrate structures, but not a precise interdomain orientation. Nonetheless the sequence and structural similarities between CEA and CD2 make it likely that CEA is well represented by the interdomain orientation seen in the two-domain crystal structure of CD2 (Jones *et al.*, 1992; Bodian *et al.*, 1994). This was used for the full homology model of CEA (Figures 7.6, 7.7, 7.8, 7.9 and 7.10).

7.4.2. Structural model for the complex between MFE-23 and CEA

The crystal structure of MFE-23 and the homology model for CEA derived from the crystal structure of CD2 provided a wealth of information that could be used to deduce a structural model for the complex formed between them. The observed lattice contacts between the *H1*, *H2* and *H3* loops of MFE-23 with an adjacent MFE-23 molecule (Chapter 6) provided the basis for the positioning of the V-set domain in the CD2 crystal structure relative to the MFE-23 antigen-binding site. It was readily deduced from this that the homology model for CEA-1 and CEA-2 based on the interdomain arrangement seen in the CD2 crystal structure offered excellent surface and electrostatic complementarity with the antigen binding site of MFE-23, while at the same time showing that there is no steric conflict with any of the carbohydrate sites in CEA. The success of this approach is attributable to the high affinity of MFE-23 for CEA (Kd = 2.5 nM) which implies that the unbound MFE-23 antigen combining site already has good shape and electrostatic complementarity for CEA and can in principle by matched to CEA. It also implies that the domain arrangement at CEA-1 and CEA-2 does in fact resemble that observed in CD2 and CD4. These results illustrate the continuing power of the joint use of constrained scattering modelling and known crystal structures to develop realistic and useful models for understanding the function of multidomain proteins (Perkins *et al.*, 1998a). The determination of a structural model for the complex now makes possible the rational design of MFE-23 mutations to enhance its binding affinity to CEA to improve its tumour targeting properties, as well

as performing the quantitative assessment of these mutations and the definition of predictions to test the structural model in detail.

Chapter 8

Summary and Conclusions

8.1. Carcinoembryonic antigen (CEA)

CEA is a seven-domain, highly glycosylated cell-surface protein, and is an important marker for colon cancer. In Chapter 4, the solution arrangement at low resolution of the seven domains in CEA cleaved from its membrane anchor was determined for the first time by X-ray and neutron scattering. Guinier analyses showed that the X-ray radius of gyration R_G of CEA was 8.0 nm. The length of CEA was 27-33 nm, and is consistent with an extended arrangement of seven domains. The X-ray cross-sectional radius of gyration R_{XS} was 2.1 nm, and is consistent with extended carbohydrate structures in CEA. The neutron data gave a CEA molecular weight of 150,000, in agreement with a value of 152,500 from composition data, and validated the X-ray analyses. The CEA scattering curves were analysed using a automated computer modelling procedure based on the crystal structure of CD2. The V-set and C2-set domains in CD2 were separated, and the C2-set domain was duplicated five times to create a linear seven-domain starting model for CEA. A total of 28 complex-type oligosaccharide chains in extended conformations were added to this model. Assuming that all six interdomain orientations in CEA were the same, three-parameter searches of the rotational orientions between the seven domains were performed to give 4,056 possible CEA models which explored all conformations. The best curve fits corresponded to models with a limited family of extended zig-zag conformations with length 27 nm and width 8 nm. Interestingly, the best-fit model was similar to a CEA model derived from the CD2 crystal structure by successive direct superimpositions of adjacent domains. Both low resolution models showed that the protein face of the GFCC' β-sheet in neighbouring CEA domains lie on alternate sides of the structure. Such a model has implications for the adhesion interactions between CEA molecules on adjacent cells or for the antibody targeting of CEA. In Chapter 7, this work was continued in which full homology models for the seven domains of CEA were created using the V-set, I-set and C2-set immunoglobulin folds found in crystal structures for CD2, CD4, CD8 and VCAM-1. Two complete CEA homology models were built based upon the best-fit and CD2-derived models from Chapter 4, both of which gave improved scattering curve fits.

8.2. Human immunoglobulin A (IgA)

Human IgA is an abundant antibody that mediates immune protection at mucosal surfaces as well as in plasma. The IgA1 isotype contains two four-domain Fab fragments and a four-domain Fc fragment analogous to that in immunoglobulin G (IgG), and linked by a glycosylated hinge region made up of 23 amino acids from each of the heavy chains. IgA1 also has two 18-residue tailpieces at the C-terminus of each heavy chain in the Fc fragment. In Chapter 5, X-ray scattering using H_2O buffers and neutron scattering using 100% ^2H_2O buffers is described for monomeric IgA1 and a recombinant IgA1 that lacks the tailpiece (PTerm455). The radii of gyration R_G from Guinier analyses were similar at 6.11-6.20 nm for IgA1 and 5.84-6.16 nm for PTerm455, and their cross-sectional radii of gyration R_{XS} were also similar. The similarity of the R_G and R_{XS} values shows that the tailpiece of IgA1 is not extended outwards into solution. The IgA1 R_G values are higher than those for IgG, and the distance distribution function $P(r)$ showed two distinct peaks whereas a single peak was observed for IgG. Both results show that the hinge of IgA1 results in an extended Fab and Fc arrangement that is different from that in IgG. Automated curve fit searches constrained by homology models for the Fab and Fc fragments were used to model the experimental IgA1 scattering curves. A limited family of IgA1 structures that gave good curve fits to the experimental data were identified. These contained extended hinges of length about 7 nm that positioned the Fab-to-Fab centre-to-centre separation 17 nm apart while keeping the corresponding Fab-to-Fc separation at 9 nm. The resulting extended T-shaped IgA1 structures are distinct from IgG structures previously determined by scattering and crystallography which have Fab-to-Fab and Fab-to-Fc separations of 7-9 nm and 6-8 nm respectively. It was concluded that the IgA1 hinge is structurally distinct from that in IgG, and this results in a markedly different antibody structure that may account for a unique immune role of monomeric IgA1 in plasma and mucosa.

8.3. The anti-CEA single-chain Fv MFE-23

MFE-23 is a murine single-chain Fv antibody molecule used for the monitoring and targeting of colon cancer through its specificity for CEA. In Chapter 6, its crystal structure was determined at 2.4 Å resolution by molecular replacement, giving an *R*-factor of 19.0%. Five of the six antigen-binding loops (*L1, L2, L3, H1, H2*) agreed with

the canonical structure classes. The sixth loop *H3* demonstrated clear electron density, and was fitted to a well-ordered structure defined by a buried Thr-H100 residue that had not been predicted previously. The antigen binding site displayed a basic centre flanked by a large acidic region located at *H1* and *H2*. The antigen-binding loops *H1*, *H2* and *H3* of MFE-23 formed extensive contacts with the DEBA β-sheet of an adjacent domain within the crystal lattice. These involved salt bridges and hydrogen bonds, and demonstrated good surface complementarity in which the hydrophilic *H2* and hydrophobic *H3* loops were wrapped around the adjacent domain surface. The crystal structure of another anti-CEA antibody A5B7 also demonstrated a similar loop topology. The lattice contacts agreed with residues in the *H1*, *H2* and *H3* loops that were predicted to be important for antigen contacts. A previously described mutation that significantly perturbed MFE-23 binding to CEA corresponded to an *H3* residue that participates in these lattice contacts. The crystal structure was also used to design two "humanised" forms of MFE-23. These were based on the human TR1.9 Fab fragment structure which belongs to the same V_H structural class 2 as MFE-23, and also like MFE-23 it has a κ light chain. It is concluded that crystal structures for small single-chain Fv molecules that are specific for immunoglobulin folds may provide a good model for their antibody-antigen interactions, and in addition provide insight on "humanisation" strategies.

In the MFE-23 crystal structure, a dimeric four-domain structure was observed. In Chapter 7, its solution structure was determined by neutron scattering and curve modelling fits to show that MFE-23 formed compact V_H-V_L-linked monomers at concentrations below 1 mg/ml, at which MFE-23 is used in therapeutic applications, and that oligomers exist at higher concentrations.

8.4. Predicted interaction between MFE-23 and CEA

The structure of the complex formed between MFE-23 and CEA was modelled using lattice contacts in the MFE-23 crystal structure which showed how the antigen-binding loops may interact with an adjacent immunoglobulin fold (Chapter 7). By superimposition based on the crystal structure of CD2 and homology models for the CEA domains, it was found that a monomer of MFE-23 could only be inserted into the

junction formed between the first two domains of CEA. In this model for the complex, the six antigen-binding loops formed contacts with both domains, and the antigen binding site protruded into the indentation between these two domains. Good surface complementarity existed, appropriate electrostatic contacts between MFE-23 and CEA were formed, and no steric conflicts with the 28 carbohydrate chains occurred. This structural model of the complex between MFE-23 and CEA will permit for the first time the rational development of improved strategies for tumour targeting.

8.5. Final thoughts

The only experimental methods to determine atomic coordinate models of proteins are X-ray crystallography and NMR methods. However, the general structural features of immunoglobulin superfamily proteins, which include large sizes, glycosylation and inter-domain flexibility, can prohibit the application of such methods. If this is the case, useful structural information can be obtained from low resolution methods such as homology modelling, small angle solution scattering and constrained scattering curve fits. In this thesis three such low resolution models have been determined by the joint application of all three methods, namely those for IgA, CEA and for the interaction between CEA and MFE-23. Although these models require experimental verification, they and the crystal structure of MFE-23 should be most useful for the ongoing design of experiments to elucidate the functions of these protein systems.

References

Aleshin, A., Golubev, A., Firsov, L. M. & Honzatko, R. B. (1992). Crystal structure of glucoamylase from *Aspergillus awamori* var. X100 to 2.2 A resolution. *J. Biol. Chem.* **267**, 19291-19298.

Aleshin, A. E., Hoffman, C., Firsov, L. M. & Honzatko, R. B. (1994). Refined crystal structures of glucoamylase from *Aspergillus awamori* var. X100. *J. Mol. Biol.* **238**, 575-591.

Al-Lazikani, B., Lesk, A. M. & Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* **273,** 927-948.

Alters, S. E., Gadea, J. R., Sorich, M., O'Donoghue, G., Talib, S. & Phillip, R. (1998). Dendritic cells pulsed with CEA peptide induce CEA-specific CTL with restricted TCR repertoire. *J. Immunotherapy,* **21,** 17-26.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Alignment Search Tool. *J. Mol. Biol.* **215**, 403-410.

Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986). Three-dimensional structure of an antigen-antibody complex at 2.8 Å resolution. *Science,* **233,** 747-753.

Apodaca, G., Bomsel, M., Arden, J., Breitfeld, P. P., Tang, K. & Mostov, K. E. (1991). The polymeric immunoglobulin receptor: a model protein to study transcytosis. *J. Clin. Invest.* **87,** 1877-1882.

Arakawa, F., Kuroki, M., Misumi, Y., Matsuo, Y. & Matsuoka, Y. (1990). The nucleotide and deduced amino acid sequences of a cDNA encoding a new species of pregnancy-specific beta 1-glycoprotein (PS beta G). *Biochim. Biophys. Acta* **1048,** 303-305.

Arulanandam, A. R. N., Moingeon, P., Concino, M. F., Recny, M. A., Kato, K., Yagita, H., Koyasu, S. & Reinherz, E. L. (1993). A soluble multimeric recombinant CD2 protein identifies CD48 as a low affinity ligand for human CD2: divergence of CD2 ligands during the evolution of humans and mice. *J. Exp. Med.* **177,** 1439-1450.

Ashton, A. W., Boehm, M. K., Gallimore, J. R., Pepys, M. B. & Perkins, S. J. (1997). Pentameric and decameric structures in solution of the serum amyloid P component by X-ray and neutron scattering and molecular modelling analyses. *J. Mol. Biol.* **272,** 408-422.

Atkin, J. D., Pleass, R. J., Owens, R. J. & Woof, J. M. (1996). Mutagenesis of the human IgA1 heavy chain tailpiece that prevents dimer assembly. *J. Immunol.* **157,** 156-159.

Baezinger, J. & Kornfeld, S. (1974). Structure of the carbohydrate units of IgA1 immunoglobulin. II. Structure of the O-glycosidically linked oligosaccharde units. *J. Biol. Chem.* **249,** 7260-7269.

Bagshawe, K. D. (1989). The 1st Bagshawe lecture - towards generating cyto-toxic agents at cancer sites. *Br. J. Cancer,* **60,** 275-281.

Bagshawe, K. D., Sharma, S. K., Springer, C. S. & Antoniw, P. (1995). Antibody directed enzyme prodrug therapy: pilot scale clinical trial. *Tumour Target.* **1,** 17-29.

Bakos, M., Kurowsky, A., Woodward, C. S., Denney, R. M. & Goldblum, R. M. (1991a). Probing the topography of free and polymeric Ig-bound human secretory component with monoclonal antibodies. *J. Immunol.* **146,** 162-168.

Bakos, M., Kurowsky, A. & Goldblum, R. M. (1991b). Characterization of a critical binding site for human polymeric Ig on secretory component. *J. Immunol.* **147,** 3419-3426.

Bairoch, A. (1991). PROSITE - a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **19,** 2241-2245.

Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence databank and its supplement TrEMBL. *Nucleic Acids Res.* **25,** 31-36.

Bajorath, J. & Sheriff, S. (1996). Comparison of an antibody model with an X-ray structure: The variable fragment of BR96. *Proteins: Struct. Funct. Genet.* **24,** 152-157.

Bajorath, J., Harris, L. & Novotny, J. (1995). Conformational similarity and systematic displacement of complementarity determining region loops in high resolution antibody X-ray structures. *J. Biol. Chem.* **270**, 22081-22084.

Banfield, M. J., King, D. J., Mountain, A. & Brady, R. L. (1996). Structure of the Fab fragment of a monoclonal antibody specific for carcinoembryonic antigen. *Acta Cryst.* **D52**, 1107-1113.

Banfield, M. J., King, D. J., Mountain, A. & Brady, R. L. (1997). V_L:V_H domain rotations in engineered antibodies: Crystal structures of the Fab fragments from two murine antitumor antibodies and their engineered human constructs. *Proteins: Struct. Funct. Genet.* **29**, 161-171.

Barclay, A. N., Birkeland, M. L., Brown, M. H., Beyers, A. D., Davis, S. J., Somoza, C. & Williams, A. F. (1993). *The leukocyte antigen factsbook.* Academic press, New York.

Barnett, T., Goebel, S. J., Nothdurft, M. A. & Elting, J. J. (1988). Carcinoembryonic antigen family: characterization of cDNAs coding for NCA and CEA and suggestion of nonrandom sequence variation in their conserved loop-domains. *Genomics,* **3,** 59-66.

Barnett, T. R., Drake, L. & Pickle II, W. (1993). Human biliary glycoprotein gene: characterization of a family of novel alternatively spliced RNAs and their expressed proteins. *Mol. Cell. Biol.* **13,** 1273-1282.

Barton, G. F. (1996). Protein sequence alignment and database scanning. In: *Protein structure prediction.* Editor, M. J. E. Sternberg. Oxford University Press, Oxford. pp31-63.

Bastian, A., Kratzin, H., Eckart, K. & Hilschmann, N. (1992). Intra- and interchain disulfide bridges of the human J chain in secretory immunoglobulin A. *Biol. Chem. Hoppe-Seyler,* **373,** 1255-1263.

Bateman, A., Eddy, S. R. & Chothia, C. (1996). Members of the immunoglobulin superfamily in bacteria. *Protein Sci.* **5,** 1936-1942.

Bates, P. A., Luo, J. & Sternberg, M. J. E. (1992). A predicted three-dimensional structure for the carcinoembryonic antigen (CEA). *FEBS Lett.* **301**, 207-214.

Beavil, A. J., Beavil, R. L., Chan, C. M. W., Cook, J. P. D., Gould, H. J., Henry, A. J., Owens, R. J., Shi, J., Sutton, B. J. & Young, R. J. (1993). Structural basis of the IgE-FcεRI interaction. *Biochem. Soc.Transact.* **21**, 968-972.

Beavil, A. J., Young, R. J., Sutton, B. J. & Perkins, S. J. (1995). Bent domain structure of recombinant human IgE-Fc in solution by X-ray and neutron scattering in conjunction with an automated curve fitting procedure. *Biochemistry*, **34**, 14449-14461.

Becker, J. W. & Reeke Jr., G. N. (1985). Three-dimensional structure of β-2-

microglobulin. *Proc. Natl. Acad. Sci. USA,* **82,** 4225-4229.

Becker, J. W., Erickson, H. P., Hoffman, S., Cunningham, B. A. & Edelman, G. M. (1989).Topology of cell adhesion molecules. *Proc. Natl. Acad. Sci. USA,* **86,** 1088-1092.

Begent, R. H. J., Verhaar, M. J., Chester, K. A., Casey, J. L., Green, A. J., Napier, M. P., Hope-Stone, L. D., Cushen, N., Keep, P. A., Johnson, C. J., Hawkins, R. E., Hilson, A. J. W. & Robson, L. (1996). Clinical evidence of efficient tumor targeting based on single-chain Fv antibody selected from a combinatorial library. *Nature Med.* **2,** 979-984.

Benchimol, S., Fuks, A., Jothy, S., Beauchemin, N., Shirota, K. and Stanners, C. P. (1989). Carcinoembryonic antigen, a human tumour marker, functions as an intercellular adhesion molecule. *Cell,* **57,** 327-334.

Bentley, G. A., Boulot, G., Karjalainen, K. & Mariuzza, R. A. (1995). Crystal structure of the β chain of a Tcell antigen receptor. *Science,* **267,** 1984-1987.

Berek, C. & Milstein, C. (1987). Mutation drift and repertoire shift in the maturation of the immune-response. *Immunol. Rev.* **96,** 23-41.

Berendt, A. R., McDowall, A., Craig, A. G., Bates, P. A., Sternberg, M. J. E., Marsh, K., Newbold, C. I. & Hogg, N. (1992). The binding site on ICAM-1 for *Plasmodium falciparum*-infected erythrocytes overlaps but is distinct from the LFA-1 binding site. *Cell,* **68,** 71-81.

Berling, B., Kolbinger, F., Grunert, F., Thompson, J. A., Brombacher, F., Buchegger, F., von Kleist, S. & Zimmermann, W. (1990). Cloning of a carcinoembryonic antigen gene family member expressed in leukocytes of chronic myeloid leukemia patients and bone marrow. *Cancer Res.* **50,** 6534-6539.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535-542.

Bhat, T. N., Bentley, G. A., Fischmann, T. O., Boulet, G. & Poljak, R. J. (1990). Small rearrangements in structures of Fv and Fab fragments of antibody D1.3 on antigen binding. *Nature,* **347,** 483-485.

Bhat, T. N., Bentley, G. A., Boulet, G., Green, M. I, Tello, D., Dall'Acqua, W., Souchon, H., Schwarz, F. P., Mariuzza, R. A. & Poljak, R. J. (1991). Bound water molecules and conformational stabilization help mediate an antibody-antigen association. *Proc. Natl. Acad. Sci. USA,* **91,** 1089-1093.

Biewenga, J. & van Run, P. E. M. (1992). Effects of limited reduction on disulphide bonds in human IgA1 and IgA1 fragments. *Molec. Immunol.* **29,** 327-334.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. & Wiley, D. C. (1987). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature,* **329,** 506-512.

Bjorkman, P. J. & Parham, P. (1990). Structure, function and diversity of class I major histocompatibility complex molecules. *Ann. Rev. Biochem.* **90,** 253-288.

Bode, W., Meyer, E.,, Jr. & Powers, J. C. (1989). Human leukocyte and porcine pancreatic elastase: X-ray crystal structures, mechanism, substrate specificity and mechanism-based inhibitors. *Biochemistry,* **28,** 1951-1963.

Bodian, D. L., Jones, E. Y., Harlos, K., Stuart, D. I. & Davis, S. J. (1994). Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2.5 Å resolution. *Structure,* **2,** 755-766.

Boehm, M. K., Mayans, M. O., Thornton, J. D., Begent, R. H. J., Keep, P. A. & Perkins, S. J. (1996). Extended glycoprotein structure of the seven domains in human carcinoembryonic antigen by X-ray and neutron solution scattering and an automated curve fitting procedure: implications for cellular adhesion. *J. Mol. Biol.* **259,** 718-736.

Bogers, W. M. J. M., Stad, R. K., van Es, L. A. & Daha, M. R. (1991). Immunoglobulin A: interaction with complement, phagocytic cells and endothelial cells. *Complement Inflamm.* **8,** 347-358.

Bork, P., Holm, L. & Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* **242,** 309-320.

Bos, M. P., Grunert, F. & Belland, R. J. (1997). Differential recognition of members of the carcinoembryonic antigen family by Opa variants of *Neisseria gonorrhoeae. Infect. Immun.* **65,** 2353-2361.

Boulin, C., Kempf, R., Koch, M. H. J. & McLaughlin, S. M. (1986). Data appraisal, evaluation and display for synchrotron radiation experiments: hardware and software. *Nucl. Instrum. Meth.* **A249**, 399-407.

Bourne, Y., Mazurier, J., Legrand, D., Rouge, P., Montreuil, J., Spik, G. & Cambillau, C. (1994). Structures of a legume lectin complexed with the human lactotransferrin N2 fragment, and with an isolated biantennary glycopeptide - role of the fucose moiety. *Structure,* **2,** 209-219.

Bowie, J. U. & Eisenberg, D. (1993). Inverted protein structure prediction. *Curr. Opin. Struct. Biol.* **3**, 437-444.

Braden, B. C. & Poljak, R. J. (1995). Structural features of the reactions between antibodies and protein antigens. *FASEB J.* **9,** 9-16.

Braden, B. C., Souchon, H., Eisele, J.-L., Bentley, G. A., Bhat, T. N., Navaza, J. & Poljak, R. J. (1994). Three-dimensional structures of the free and the antigen-complexed Fab from monclonal anti-lysozyme antibody D44.1. *J. Mol. Biol.* **243,** 767-781.

Braden, B. C., Goldman, E. R., Mariuzza, R. A. & Poljak, R. J. (1998). Anatomy of an antibody molecule: structure, kinetics, thermodynamics and mutational studies of the antilysozyme antibody D1.3. *Immunol. Rev.* **163,** 45-57.

Brady, R. L., Dodson, E. J., Dodson, G. G., Lange, G., Davis, S. J., Williams, A. F. & Barclay, A. N. (1993). Crystal structure of domains 3 and 4 of rat CD4: relation to the NH2-terminal domains. *Science,* **260,** 979-983.

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L. & Wiley, D. C. (1993). 3-dimensional structure of the human class-II histocompatibility antigen HLA-DR1. *Nature,* **364,** 33-39.

Brümmendorf, T. & Rathjen, F. G. (1995). Cell adhesion molecules 1: immunoglobulin superfamily. *Protein Profiles,* **2.**

Brummer, J., Neumaier, M., Gopfert, C. & Wagener, C. (1995). Association of pp60 (c-src) with biliary glycoprotein (CD66a), an adhesion molecule of the carcinoembryonic antigen family down-regulated in colorectal carcinomas. *Oncogene,* **11,** 1649-1655.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). Crystallographic R-factor refinement by molecular dynamics. *Science,* **235,** 458-460.

Brünger, A. T. (1988). Crystallographic refinement by simulated annealing applications to a 2.8 Å resolution structure of aspartate aminotransferase. *J. Mol. Biol.* **203,** 803-816.

Brünger, A. T. (1992a). *X-PLOR version 3.1.* Yale University Press, New Haven, USA.

Brünger, A. T. (1992b). The free R value: a novel statistical quantity for assessing the accuracy crystal structures. *Nature,* **355,** 472-474.

Buckley, C. D., Simmons, D. L. (1997). Cell adhesion: a new target for therapy. *Molec. Med. Today,* **3,** 449-456.

Burmeister, W. P., Gastinal, L. N., Simister, N. E., Blum, M. L. & Bjorkman, P. J. (1994a). Crystal structure at 2.2 Å resolution of the MHC-related neonatal Fc receptor. *Nature,* **372,** 336-343.

Burmeister, W. P., Huber, A. H. & Bjorkman, P. J. (1994b). Crystal structure of the complex of rat neonatal Fc receptor with Fc. *Nature,* **372,** 379-383.

Burrows, P. D., Schroeder Jr, H. W. & Cooper, M. D. (1995). B-cell differentiation in humans. In: *Immunoglobulin Genes (Second Edition).* Editors, T. Honjo & F. W. Alt. Academic Press, London. pp3-31.

Burton, D. R. (1990). Antibody: the flexible adaptor molecule. *Trends Biochem. Sci.* **15,** 64-69.

Burton, D. R. & Woof, J. M. (1992). Human antibody effector functions. *Adv. Immunol.* **51,** 1-84.

Campbell, I. D. & Downing, K. (1994). Building protein structure and function from modular units. *Trends Biotech.* **12,** 168-172.

Campbell, I. D. & Dwek, R. A. (1984). *Biological Spectroscopy.* Benjamin/Cummings Publishing Company, Menlo Park, California.

Campbell, I. D. & Spitzfaden, C. (1994). Building proteins with fibronectin type III modules. *Structure,* **2,** 333-337.

Canfield, S. M. & Morrison, S. L. (1991). The binding affinity of human IgG for its high affinity Fc receptor is determined by multiple amino acids in the CH2 domain and is modulated by the hinge region. *J. Exp. Med.* **173,** 1483-1491.

Carayannopoulos, L., Max, E. E. & Capra, J. D. (1994). Recombinant human IgA expressed in insect cells. *Proc. Natl. Acad. Sci. USA,* **91,** 8348-8352.

Carayannopoulos, L., Hexham, J. M. & Capra, J. D. (1996). Localization of the binding site for monocyte immunoglobulin (Ig) A-Fc receptor (CD89) to the domain boundary between Ca2 and Ca3 in human IgA1. *J. Exp. Med.* **183,** 1579-1586.

Casasnovas, J. M., Springer, T. A., Liu, J.-H., Harrison, S. C. & Wang, J.-H. (1997). Crystal structure of ICAM-2 reveals a distinctive integrin recognition surface. *Nature,* **387,** 312-315.

Casey, J. L., Keep, P. A., Chester, K. A., Robson, L., Hawkins, R. E. & Begent, R. H. J. (1995). Purification of bacterially expressed single chain Fv antibodies for clinical applications using metal chelate chromatography. *J. Immunol. Methods,* **179,** 105-116.

Castellano, E. E., Oliva, G. & Navaza, J. (1992). Fast rigid-body refinement for molecular-replacement techniques. *J. Appl. Cryst.* **25,** 281-284.

CCP4 (1994). The CCP4 suite: programs for protein crystallography. *Acta Cryst.* **D50,** 760-763.

Chacko, S., Padlan, E. A., Portolano, S., McLachlan, S. M. & Rapoport, B. (1996). Structural studies of human autoantibodies - crystal-structure of a thyroid peroxidase autoantibody Fab. *J. Biol. Chem.* **271,** 12191-12198.

Chamberlain, D., Keeley, A., Aslam, M., Arenas-Licea, J., Brown, T., Tsaneva, I. R. & Perkins, S. J. (1998). A synthetic Holliday junction is sandwiched between two

tetrameric *Myobacterium leprae* RuvA structures in solution: new insights from neutron scattering contrast variation and modelling. *J. Mol. Biol.*, in press.

Chandrasekaran, E. V., Davila, M., Nixon, D. W., Goldfarb, M. & Mendicino, J. (1983). Isolation and structures of the oligosaccharide units of carcinoembryonic antigen. *J. Biol. Chem.* **258**, 7213-7222.

Chappel, M. S., Isenman, D. E.., Everett, M., Xu, Y.-Y., Dorrington, K. J. & Klein, M. H. (1991). Identification of the Fcγ receptor class I binding site in human IgG through the use of recombinant IgG1/IgG2 hybrid and point-mutated antibodies. *Proc. Natl. Acad. Sci. USA,* **88,** 9036-9040.

Chen, D. S., Asanaka, M., Chen, F. S., Shively, J. E. & Lai, M. M. C. (1997a). Human carcinoembryonic antigen and biliary glycoprotein can serve as mouse hepatitis virus recpetors. *J. Virol.* **71,** 1688-1691.

Chen, T., Grunert, F., MedinaMarino, A & Gotschlich, E. C. (1997b). Several carcinoembryonic antigens (CD66) serve as receptors for gonococcal opacity proteins. *J. Exp. Med.* **185,** 1557-1564.

Chester, K. A., Begent, R. H. J., Robson, L., Keep, P. A., Pedley, R. B., Boden, J. A., Boxer, G., Green, A., Winter, G., Cochet, O. & Hawkins, R. E. (1994). Phage libraries for generation of clinically useful antibodies. *Lancet,* **343,** 455-456.

Chester, K. A. & Hawkins, R. E. (1995). Clinical issues in antibody design. *Trends Biotechnol.* **13,** 294-300.

Chintalacharuvu, K. R., Tavill, A. S., Loui, L. N., Vaerman, J. P., Lamm, M. E. & Kaetzel, C. S. (1994). Disulphide formation between dimeric IgA and the polymeric immunoglobulin receptor during hepatic transcytosis. *Hepatology,* **19,** 162-173.

Chintalacharuvu, K. R., Raines, M. & Morrison, S. L. (1994). Divergence of human α-chain constant region gene sequences. A novel recombinant α2 gene. *J. Immunol.* **152,** 5299-5304.

Chintalacharuvu, K. R. & Morrison, S. L. (1996). Residues critical for H-L disulphide bond formation in human IgA1 and IgA2. *J. Immunol.* **157,** 3443-3449.

Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254,** 304-308.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature,* **357,** 543-544.

Chothia, C. & Finkelstein, A. V. (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59,** 1007-1039.

Chothia, C. & Janin, J. (1981). Relative orientation of close-packed β-pleated sheets in proteins. *Proc. Natl. Acad. Sci. USA,* **21,** 4146-4150.

Chothia, C. & Jones, E. Y. (1997). The molecular structure of cell adhesion molecules. *Annu. Rev. Biochem.* **66,** 823-862.

Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196,** 901-917.

Chothia, C., Novotny, J., Bruccoleri, R. & Karplus, M. (1985). Domain association in immunoglobulin molecules: the packing of variable domains. *J. Mol. Biol.* **186,** 651-663.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, **342,** 877-883.

Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins

from their amino acid sequence. *Advan. Enzymol. Relat. Areas Mol. Biol.* **47,** 45-148.

Chuang, P. D. & Morrison, S. L. (1997). Elimination of N-linked glycosylation sites from the human IgA1 constant region. *J. Immunol.* **158,** 724-732.

Clements, J. M., Newham, P., Shepherd, M., Gilbert, R., Dudgeon, T. J., Needham, L. A., Edwards, R. M., Berry, L., Brass, A. & Humphries, M. J. (1994). Identification of a key integrin-binding sequence in VCAM-1 homologous to the LDV active-site in fibronectin. *J. Cell Science* **107,** 2127-2135.

Cohn, E. J. & Edsall, J. T. (1943). Density and apparent specific volume of proteins. In: *Proteins, amino acids and peptides.* Reinhold Publ. Corp. New York. pp370-381.

Collaborative Computational Project Number 4. (1994). The CCP4 suite: programs for crystallography. *Acta Cryst.* **D50,** 760-763.

Colman, P. M. (1988). Structure of antibody-antigen complexes: implications for immune recognition. *Adv. Immunol.* **43,** 99-132.

Corper, A. L., Sohi, M. K., Bonagura V. R., Steinitz M., Jefferis R., Feinstein A., Beale D., Taussig M.J., Sutton B.J. (1997). Structure of human IgM rheumatoid factor Fab bound to its autoantigen IgG Fc reveals a novel topology of antibody-antigen interaction. *Nat. Struct. Biol.* **4,** 374-381.

Correll, C. C., Ludwig, M. L., Bruns, C. M., and Karplus, P. A. (1993). Structural prototypes for an extended family of flavoprotein reductases - comparison of phthalate dioxygenase reductase with ferredoxin reductase and ferredoxin. *Prot. Sci.* **2,** 2112-2133.

Coyne, R. S., Siebrecht, M., Pietsch, M. C. & Casanove, J. E. (1994). Mutational analysis of polymeric immunoglobulin receptor/ligand interactions. *J. Biol. Chem.* **269,** 31620-31625.

Crowther, R. A. (1972). In: *The Molecular Replacement Method.* Editor, M. G. Rossmann. Gordon & Breach, New York. pp173-178.

Crowther, R. A. & Blow, D. M. (1967). *Acta Cryst.* **23,** 544-548.

Cupp-Vickery, J. R. & Poulos, T. L. (1995). Structure of cytochrome p450eryf involved in erythromycin biosynthesis. *Nat. Struct. Biol.* **2,** 144-153.

Davies, D. R. & Cohen, G. H. (1996). Interactions of protein antigens with antibodies. *Proc. Natl. Acad. Sci. USA,* **93,** 7-12.

Davies, D. R. & Padlan, E. A. (1992). Twisting into shape. *Curr. Biol.* **2,** 254-256.

Davis, S. J., Brady, R. L., Barclay, A. N., Harlos, K., Dodson, G. G. & Williams, A. F. (1990). Crystallization of a soluble form of the rat T-cell surface glycoprotein CD4 complexed with Fab from the W3/25 monoclonal antibody. *J. Mol. Biol.* **213,** 7-10.

Dayhoff, M. O., Park, C. M. & McLaughlin, P. J. (1972). *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington DC.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington DC.

Deisenhofer, J. (1981). Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from Staphylococcus aureus at 2.9 and 2.8 Å resolution. *Biochemistry,* **20,** 2361-2370.

de Vos, A. M., Ultsch, M. & Kossiakoff, A. (1992). Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science,* **255,** 306-312.

Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? *Science,* **214,** 149-159.

Doolittle. R. F. (1994). Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19,** 15-18.

Doolittle. R. F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64,** 287-314.

Doolittle, R. F. & Bork, P. (1993). Evolutionary mobile modules in proteins. *Sci. Amer.* **269,** 50-56.

Drenth, J. (1994). *Principles of protein X-ray crystallography.* Springer advanced texts in chemistry, New York.

Driscoll, P. C., Cyster, J. G., Campbell, I. D. & Williams, A. F. (1991). Structure of domain 1 of rat T lymphocyte CD2 antigen. *Nature,* **353,** 762-765.

Dveksler, G. S., Dieffenbach, C. W., Cardellichio, C. B., McCuaig, K., Pensiero, M. N., Jiang, G. S., Beauchemin, N. & Holmes, K. V. (1993). Several members of the mouse carcinoembryonic antigen-related glycoprotein family are functional receptors for the coronavirus mouse hepatitis virus-A59. *J. Virol.* **67,** 1-8.

Edelman, G. M. (1970). The covalent structure of a human IgG. XI. Functional implications. *Biochemistry,* **9,** 3197-3205.

Edwards, Y. J. K. & Perkins, S. J. P. (1995). The protein fold of the von Willebrand factor type A domain is predicted to be similar to the open twisted β-sheet flanked by α-helices found in human ras-p21. *FEBS Lett.* **358,** 283-286.

Edwards, Y. J. K. & Perkins, S. J. (1996). Assessment of protein fold predictions from sequence information: the predicted α/β doubly wound fold of the von Willebrand Factor Type A domain is similar to its crystal structure. *J. Mol. Biol.* **260,** 277-285.

Eigenbrot, C., Randal, M., Presta, L., Carter, P. & Kossiakoff, A. A. (1993). X-ray structures of the antigen-binding domains from three variants of humanized antip185HER2 antibody 4D5 and comparison with molecular modelling. *J. Mol. Biol.* **229,** 969-995.

Eigenbrot, C., Gonzalez, T., Mayeda, J., Carter, P., Werther, W., Hotaling, T., Fox, J. & Kessler, J. (1994). X-ray structures of fragments from binding and nonbinding versions of a humanized anti-CD18 antibody: structural indications of the key role of VH residues 59-65. *Proteins: Struct. Funct. Genet.* **18,** 49-62.

Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein-structure refinement. *Acta Cryst.* **A47,** 392-400.

Evans, P. R. (1993). Data reduction. In: *Data collection and processing. Proceedings of the CCP4 study weekend 29-30 January 1993.* SERC Daresbury Laboratory.

Fallgreen-Gebauer, E., Gebauer, W., Bastian, A., Kratzin, H. D., Eiffert, H., Zimmermann, B., Karas, M. & Hilschman, N. (1993). The covalent linkage of secretory component to IgA. *Biol. Chem. Hoppe-Seyler,* **374,** 1023-1028.

Field, M. C., Dwek, R. A., Edge, C. J. & Rademacher, T. W. (1989). O-linked oligosaccharides from human serum immunoglobulin A1. *Biochem. Soc. Trans.* **17,** 1034-1035.

Field, M. C., Amatayakul-Chantler, S., Rademacher, T. W., Rudd, P. M. & Dwek, R. A. (1994). Structural analysis of the N-glycans from human immunoglobulin A1: comparison of normal human serum immunoglobulin A1 with that isolated from patients with rheumatoid arthritis. *Biochem. J.* **299,** 261-275.

Flores, T. P., Orengo, C. A., Moss, D. & Thornton, J. M. (1993). Comparisons of

conformational characteristics in structurally similar protein pairs. *Prot. Sci.* **2,** 1811-1826.

Fong, S., Hamill, S. J., Proctor, M., Freund, S. M. V., Benian, G. M., Chothia, C., Bycroft, M. & Clarke, J. (1996). Structure and stability of an immunoglobulin superfamily domain from twitchin, a muscle protein of the nematode *Caenorhabditis elegans*. *J. Mol. Biol.* **264,** 624-639.

Fremont, D. H., Hendrickson, W. A., Marrack, P. & Kappler, J. (1996). Structures of an MHC class II molecule with covalently bound single peptides. *Science,* **272,** 1001-1004.

Freund, C., Ross, A., Guth, B., Plückthun, A. & Holak, T. A. (1993). Characterization of the linker of the single-chain Fv fragment of an antibody by NMR spectroscopy. *FEBS Lett.* **320,** 97-100.

Freund, C., Ross, A., Plückthun, A. & Holak, T. A. (1994). Structural and dynamic properties of the Fv fragment and the single-chain Fv fragment of an antibody in solution investigated by heteronuclear three-dimensional NMR spectroscopy. *Biochem*istry, **33,** 3296-3303.

Frutiger, S., Hughes, G. J., Hanly, W. C., Kingzette, M. & Jaton, J. C. (1986). The amino terminal domain of rabbit secretory component is responsible for noncovalent binding to immunoglobulin A dimer. *J. Biol. Chem.* **261,** 16673-16681.

Gamulin, V., Rinkevich, B., Schacke, H., Kruse, M., Muller, I. M. & Muller, W. E. G. (1994). Cell-adhesion receptors and nuclear receptors are highly conserved from the lowest metazoa (marine sponges) to vertebrates. *Biol. Chem. Hoppe-Seyler,* **375,** 583-588.

Gangopadhyay, A. & Thomas, P. (1996). Processing of carcinoembryonic antigen by Kupffer cells - recognition of a penta-peptide sequence. *Arch. Biochem. Biophys.* **334,** 151-157.

Gangopadhyay, A., Lazure, D. A., Kelly, T. M. & Thomas, P. (1996a). Purification and analysis of an 80-kDa carcinoembryonic antigen-binding protein from Kupffer cells. *Arch. Biochem. Biophys.* **328,** 151-157.

Gangopadhyay, A., Bajenova, O., Kelly, T. M. & Thomas, P. (1996b). Carcinoembryonic antigen induces cytokine expression in Kuppfer cells - implications for hepatic metastasis from colorectal-cancer. *Cancer Res.* **56,** 4805-4810.

Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E. & Wiley, D. C. (1996). Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature,* **384,** 134-141.

Garcia, K. C., Degano, M., Stanfield, R. L., Brunmark, A., Jackson, M. R., Peterson, P. A., Teyton, L. & Wilson, I. A. (1996). An alpha beta T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science,* **274,** 209-219.

Garcia de la Torre, J. & Bloomfield, V. A. (1977a). Hydrodynamic properties of macromolecular complexes. I. Translation. *Biopolymers*, **16,** 1747-1761.

Garcia de la Torre, J. & Bloomfield, V. A. (1977b). Hydrodynamics of macromolecular complexes. III. Bacterial viruses. *Biopolymers*, **16,** 1779-1793.

Garcia de la Torre, J. (1989). Hydrodynamic properties of macromolecular assemblies. In: *Dynamic Properties of Biomolecular Assemble.* Editors, S. E. Harding & A. J. Rowe. Royal Society of Chemistry, Cambridge. pp3-31

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and

implications of simple methods for predicting the secondary structures of globular proteins. *J. Mol. Biol.* **120**, 97-120.

Garrett, T. P. J., Wang, J., Yan, Y., Liu, J. & Harrison, S. C. (1993). Refinement and analysis of the structure of the 1st 2 domains of human CD4. *J. Mol. Biol.* **234,** 763-778.

Gerken, T. A., Butehhof, K. J. & Shogren, R. (1989). Effects of glycosylation on the conformation and dynamics of O-linked glycoproteins - C-13 NMR-studies of ovine submaxillary mucin. *Biochemistry,* **28,** 5536-5543.

Ghosh, R. E. (1989) *A Computing Guide for Small Angle Scattering Experiments.* Institut Laue Langevin Internal Publication 89GH02T.

Gibrat, J. F., Garnier, J. & Robson B. (1987). Further developments of protein secondary structure predictions using information theory - new parameters and consideration of residue pairs. *J. Mol. Biol.* **198**, 425-443.

Glatter, O. & Kratky, O. (1982). Editors of *Small-angle X-ray scattering.* Academic Press, New York.

Glusker, J. P. & Trueblood, K. N. (1985). *Crystal structure analysis. A primer.* Oxford University Press, New York.

Gold, P. and Freedman, S. O. (1965). Demonstration of tumour specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. *J. Exp. Med.* **121,** 439-462.

Gordon, A. H. (1975). In: *Electrophoresis of Proteins in Polyacrylamide and Starch Gels.* Editors, T. S. Work & E. Work. North Holland Publ. Co., Amsterdam. pp153s-164s

Gray-Owen, S. D., Dehio, C., Haude, A., Grunert, F. & Meyer, T. F. (1997). CD66 carcinoembryonic antigens mediate interactions between Opa-expressing *Neisseria gonorrhoeae* and human polymorphonuclear phagocytes. *EMBO J.* **16,** 3435-3445.

Griffiss, J. M. & Goroff, D. K. (1983). IgA blocks IgM and IgG-initiated immune lysis by separate molecular mechanisms. *J. Immunol.* **130,** 2882-2885.

Grimm, T. & Johnson, J. P. (1995). Ectopic expression of carcinoembryonic antigen by a melanoma cell leads to changes in the transcription and expression of 2 additional cell-adhesion molecules. *Cancer Res.* **55,** 3254-3257.

Gumbiner, B. M. (1996). Cell-adhesion - the molecular-basis of tissue architecture and morphogenesis. *Cell,* **84,** 345-357.

Hahn, T. (1996). Editor of *International Tables for Crystallography, Vol. A.* Kluwer Academic Publishers, Dordrecht, Holland.

Hammarström, S., Engvall, E., Johansson, B. G., Svensson, S., Sundblad, G. & Goldstein, I. J. (1975). Nature of the tumor-associated determinant(s) of carcinoembryonic antigen. *Proc. Nat. Acad. Sci., USA,* **72**, 1528-1532.

Hammarström, S., Shively, J. E., Paxton, R. J., Beatty, B. G., Larsson, A., Ghosh, R., Bormer, O., Buchegger, F., Mach, J.-P., Burtin, P., Seguin, P., Darbouret, B., Degorce, F., Sertour, J., Jolu, J. P., Fuks, A., Kalthoff, H., Schmiegel, W., Arndt, R., Kloppel, G., von Kleist, S., Grunert, F., Schwarz, K., Matsuoka, Y., Kuroki, M., Wagener, C., Weber, T., Yachi, A., Imai, K., Hishikawa, N. & Tsujisaki, M. (1989). Antigenic sites in carcinoembryonic antigen. *Cancer Res.* **49**, 4852-4858.

Harlos, K., Martin, D. M. A, O'Brien, D. P., Jones, E. Y., Stuart, D. I., Polikarpov, I, Miller, A., Tuddenham, E. G. D. & Boys, C. W. G. (1994). Crystal structure of

the extracellular region of human tissue factor. *Nature*, **370**, 662-666.

Haris, P. I. & Chapman, D. (1994). Analysis of polypeptide and protein structures using Fourier transform infrared spectroscopy. *Methods Mol. Biol.* **22,** 183-202.

Harpaz, Y. & Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface-receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.* **127,** 1211-1219.

Harriman, W., Vőlk, H., Defranoux, N. & Wabl, M. (1993). Immunoglobulin class switch recombination. *Annu. Rev. Immunol.* **11,** 361-384.

Harris, L. J., Larson, S. B., Hasel, K. W., Day, J., Greenwood, A. & McPherson, A. (1992). The three-dimensional structure of an intact monoclonal antibody for canine lymphoma. *Nature,* **360,** 369-372.

Harris, L. J., Larson, S. B., Hasel, K. W. & McPherson, A. (1997). Refined structure of an intact IgG2a monoclonal antibody. *Biochemistry,* **36,** 1581-1597.

Harris, L. J., Skaletsky, E. & McPherson, A. (1998a). Crystallographic structure of an intact IgG1 monoclonal antibody. *J. Mol. Biol.* **275,** 861-872.

Harris, L. J., Larson, S. B., Skaletsky, E. & McPherson, A. (1998b). Comparison of the conformations of two intact monoclonal antibodies with hinges. *Immunol. Rev.* **163,** 35-43.

Hashino, J., Fukuda, Y., Iwao, K., Krop-Watorek, A., Oikawa, S., Nakazato, H. & Nakanishi, T. (1993). Production and characterization of monoclonal antibodies to N-domain and domain III of carcinoembryonic antigen. *Biochem. Biophys. Res. Comm.* **197**, 886-893.

Hauck, C. R., Meyer, T. F. Lang, F. & Gulbins, E. (1998). CD66-mediated phagocytosis of Opa_{52} *Neisseria gonorrhoeae* requires a Src-like tyrosine kinase- and Rac1-dependent signalling pathway. *EMBO J.* **17,** 443-454.

Heenan, R. K. & King, S. M. (1993). Development of the small-angle diffractometer LOQ at the ISIS pulsed neutron source. In *Proceedings of an International Seminar on Structural Investigations at Pulsed Neutron Sources, Dubna, 1st-4th September 1992. Report E3-93-65,* Joint Institute for Nuclear Research, Dubna.

Heenan, R. K., King, S. M., Osborn, R. & Stanley, H. B. (1989). *COLETTE Users Guide.* Internal publication RAL-89-128, Rutherford Appleton Laboratory, Didcot, U. K.

Hefta, S. A., Hefta, L. J. F., Lee, T. D., Paxton, R. J. & Shively, J. E. (1988). Carcinoembryonic antigen is anchored to membranes by covalent attachment to a glycosylphosphadidylinositol moiety: identification of the ethanolamine linkage site. *Proc. Natl. Acad. Sci., USA,* **85**, 4648-4652.

Hefta, L. J. F., Chen, F. S., Ronk, M., Sauter, S. L., Sarin, V., Oikawa, S., Nakazato, H., Hefta, S. & Shively, J. E. (1992). Expression of carcinoembryonic antigen and its predicted immunoglobulin-like domains in HeLa cells for epitope analysis. *Cancer Res.* **52,** 5647-5655.

Henikoff, S. & Henikoff, J. G. (1992). Amino-acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA,* **89,** 10915-10919.

Hiemstra, P. S., Gorter, A., Stuurman, M. E., van Es, L. A. & Daha, M. R. (1987). Activation of the alternative pathway of complement by human serum IgA. *Eur. J. Immunol.* **17,** 321-326.

Hiemstra, P. S., Biewenga, J., Gorter, A., Stuurman, M. E., Faber, A., van Es, L. A. & Dahi, M. R. (1988). Activation of complement by human serum IgA, secretory

IgA and IgA1 fragments. *Mol. Immunol.* **25,** 527-533.

Hilbert, M., Bohm, G. & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins,* **17,** 138-151.

Hill, R. L., Delaney, R., Fellows, R. E. & Lebovitz, H. E. (1966). The evolutionary origins of the immunoglobulins. *Proc. Natl. Acad. Sci. USA,* **56,** 1762-1769.

Hinoda, Y., Neumaier, M., Hefta, S. A., Drzeniek, Z., Wagener, C., Shively, L., Hefta, L. J., Shively, J. E. & Paxton, R. J. (1988). Molecular cloning of a cDNA coding biliary glycoprotein I: primary structure of a glycoprotein immunologically crossreactive with carcinoembryonic antigen. *Proc. Natl. Acad. Sci. USA,* **85,** 6959-6963.

Hjelm, R. P. (1985).The small-angle approximation of X-ray and neutron scatter from rigid rods of non-uniform cross section and finite length. *J. Appl. Cryst.* **18,** 452-460.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science,* **3,** 522-524.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Science,* **1,** 409-417.

Hoedemaeker, P. J., Signorelli, T., Johns, K., Kuntz, D. A. & Rose, D. R. (1997). A single chain Fv fragment of P-glycoprotein-specific monoclonal antibody C219: Design, expression and crystal structure at 2.4 Å resolution. *J. Biol. Chem.* **272,** 29784-29789.

Holden, H. M., Ito, M., Hartshorne, D. J. & Rayment, I. (1992). X-ray structure determination of Telokin, the C-terminal domain of myosin light chain kinase, at 2.8 Å resolution. *J. Mol. Biol.* **227,** 840-851.

Holm, L. & Sander, C. (1998). Touring protein fold space with DALI/FSSP. *Nucl. Acids Res.* **26,** 316-319.

Holmes, M. A. & Foote, J. (1997). Structural consequences of humanizing an antibody. *J. Immunol.* **158,** 2192-2201.

Holmgren, A. & Bränden, C. -I. (1989). Crystal structure of chaperone protein PapD reveals an immunoglobulin fold. *Nature,* **342,** 248-251.

Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science,* **268,** 1144-1149.

Hubbard, T. J. P. & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1,** 159-171.

Huston, J. S., Levinson, D., Mudgett-Hunter, M., Tai, M.-S., Novotný, J., Margolies, M. N., Ridge, R. J., Bruccoleri, R. E., Haber, E., Crea, R. & Oppermann, H. (1988). Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in *Eschericia coli. Proc. Natl. Acad. Sci. USA,* **85,** 5879-5883.

Ikeda, S., Kuroki, M., Haruno, M., Oikawa, S., Nakazato, H., Kosaki, G. & Matsuoka, Y. (1992). Epitope mapping of the carcinoembryonic antigen with various related recombinant proteins expressed in Chinese hamster ovary cells and 25 distinct monoclonal antibodies. *Molec. Immunol.* **29,** 229-240.

Ilantzis, C., Jothy, S., Apert, L. C., Draber, P. & Stanners, C. P. (1997). Cell-surface levels of human carcinoembryonic antigen are inversely correlated with colonocyte differentiation in colon carcinogenesis. *Lab. Invest.* **76,** 703-716.

Improta, S., Politon, A. S. & Pastore, A. (1996). Immunoglobulin-like modules from titin I-band extensible components of muscle elasticity. *Structure,* **4,** 323-337.

Islam, S. A., Luo, J. & Sternberg, M. J. E. (1995). Identification and analysis of domains in proteins. *Protein Eng.* **8,** 513-525.

Jacrot, B. & Zaccai, G. (1981). Determination of molecular-weight by neutron-scattering. *Biopolymers,* **20,** 2413-2426.

Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265,** 16027-16030.

Jardetzky, T. S., Brown, J. H., Gorga, J. C., Stern, L. J., Urban, R. G., Chi, Y. I., Stauffacher, C., Stominger, J. L. & Wiley, D. C. (1994). Three-dimensional structure of a human class II histocompatibility molecule complexed with superantigen. *Nature,* **368,** 711-718.

Jarvis, G. A. & Griffiss, J. M. (1991). Human IgA1 blockade of IgG-initiated lysis of *Neisseria meningitidis* is a function of antigen-binding fragment binding to polysaccharide capsule. *J. Immunol.* **147,** 1962-1967.

Jean, F., Malapert, P., Rougon, G. & Barbet, J. (1988). Cell membrane, but not circulating, carcinoembryonic antigen is linked to a phosphatidylinositol-containing hydrophobic domain. *Biochem. Biophys. Res. Comm.* **155,** 794-800.

Johnson, G., Kabat, E. A. & Wu, T. T. (1996). Kabat database of sequences of proteins of immunological interest. In: *Weir's handbook of experimental immunology I. Immunochemistry and molecular immunology. 5th edition.* Editors, W. M. Weir, L. A. Herzenberg, & C. Blackwell. Blackwell Science Inc., Cambridge, MA. pp6.1-6.21.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992b). A new approach to protein fold recognition. *Nature,* **358,** 86-89.

Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. & Stuart, D. I. (1992). Crystal structure at 2.8 Å resolution of a soluble form ot the cell adhesion molecule CD2. *Nature,* **360,** 232-239.

Jones, E. Y., Harlos, K., Bottomley, M. J., Robinson, R. C., Driscoll, P. C., Edwards, R. M., Clements, J. M., Dudgeon, T. J. & Stuart, D. I. (1995). Crystal structure of an integrin-binding fragment of vascular cell adhesion molecule-1 at 1.8 Å resolution. *Nature,* **373,** 539-544.

Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. & Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature,* **321,** 522-525.

Jones, T. A. & Liljas, L. (1984). Crystallographic refinement of macromolecules having non-crystallographic symmetry. *Acta Cryst.* **A40,** 50-57.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Cryst.* **A47,** 110-119.

Kabat, E. A., Wu, T. T. & Bilofsky, H. (1977). Unusual distributions of amino acids in complementary-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. *J. Biol. Chem.* **252,** 6609-6616.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). Sequences of proteins of immunological interest. 5th edition. US Department of Health and Human Services, Public Health Service, National Institutes of Health (NIH Publication No. 91-3242).

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Kaetzel, C. S., Robinson, J. K., Chintalacharuvu, K. R., Vaerman, J. P. & Lamm, M. E. (1991). The polymeric immunoglobulin receptor (secretory component) mediates transport of immune complexes across epithelial cells: a local defense function for IgA. *Proc. Natl. Acad. Sci. USA,* **88,** 8796-8800.

Kaetzel, C. S., Robinson, J. K. & Lamm, M. E. (1994). Epithelial transcytosis of monomeric IgA and IgG cross linked through antigen to polymeric IgA. *J. Immunol.* **152,** 72-78.

Keck, U., Nedellec, P., Beauchemin, N., Thompson, J. & Zimmermann, W. (1995). The CEA10 gene encodes a secreted member of the murine carcinoembryonic antigen family and is expressed in the placenta, gastrointestinal-tract and bone-marrow *Eur. J. Biochem.* **229,** 455-464.

Keep, P. A., Leake, B. A. & Rogers, G. T. (1978). Extraction of CEA from tumour tissue, foetal colon and patients' sera, and the effect of perchloric acid. *Brit. J. Cancer*, **37**, 171-182.

Kerr, M. A. (1990). The structure and function of human IgA. *Biochem. J.* **271,** 285-296.

Kerr, M. A. & Woof, J. M. (1998). Fc_α receptors. In: *Mucosal Immunology. 2nd edition.* Editors, P. L. Ogra, J. Mestecky, M. E. Lamm, W. Strober, J. R. McGhee and J. Bienenstock. Academic Press Inc., San Diego. In press.

Kerr, M. A., Loomes, L. M., Bonner, B. C., Hutchings, A. B. & Senior, B. W. (1997). Purification and characterization of human serum and secretory IgA1 and IgA2 using jacalin. *Methods Mol. Med.* **9,** 265-278.

Kessler, M. J., Shively, J. E., Pritchard, D. G. & Todd, C. W. (1978). Isolation, immunological characterization and structural studies of a tumour antigen releated to carcinoembryonic antigen. *Cancer Res.* **38**, 1041-1048.

Khan, W. N., Frängsmyr, L., Teglund, S., Israelsson, A., Bremer, K. & Hammaström, S. (1992a). Identification of three new genes and estimation of the size of the carcinoembronic antigen superfamily. *Genomics,* **14,** 384-390.

Khan, W. N., Teglund, S., Bremer, K. & Hammaström, S. (1992b). The pregnancy-specific glycoprotein family of the immunoglobulin superfamily: identification of new members and estimation of family size. *Genomics,* **12,** 780-787.

Killeen, N., Moessner, R., Arvieux, J., Willis, A. & Williams, A. F. (1988). The MRC OX-45 antigen of rat leukocytes and endothelium is in a subset of the immunoglobulin superfamily with CD2, LFA-3 and carcinoembryonic antigens. *EMBO J.* **7,** 3087-3091.

Killian, M. & Russell, M. W. (1994). Function of mucosal immunoglobulins. In: *Handbook of Mucosal Immunology.* Editors, P. I. Ogra, M. E. Lamm, J. R. McGhee, J. Mestecky, W. Strober, & J. Bienenstock. Academic Press, San Diego. pp 127-137.

Kimball, P. M. & Brattain, M. G. (1978). A comparison of methods for the isolation of carcinoembryonic antigen. *Cancer Res.* **38**, 619-623.

King, R. D. (1996). Prediction of secondary structure. In: *Protein structure prediction.* Editor, M. J. E. Sternberg. Oxford University Press, Oxford. pp 79-99.

Kirszbaum, L., Sharpe, J. A., Goss, N., Lahnstein, J. & Walker, I. D. (1989). The α-2 chain of murine CD8 lacks an invariant Ig-like disulphide bond but contains a

unique intrachain loop instead. *J. Immunol.* **142,** 3931-3936.

Kleywegt, G. J. & Jones, T. A. (1995). XdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Cryst.* **D52,** 826-828.

Kleywegt, G. J. & Jones, T. A. (1995b). Braille for pugilists. In: *Making the most of your model. Proceedings of the CCP4 study weekend 6-7 January 1995.* Editors, W. N. Hunter, J. M. Thornton & S. Bailey. CCL Daresbury Laboratory, UK. pp 11-24.

Kleywegt, G. J. & Jones, T. A. (1996). Efficient rebuilding of protein structures. *Acta Cryst.* **D52,** 829-832.

Kleywegt, G. J. & Jones, T. A. (1997). Model building and refinement practice. *Methods. Enzymol.* **277,** 208-230.

Klickstein, L. B., York, M. B., de Fougerolles, A. R. & Springer, T. A. (1996). Localization of the binding site on intercellular adhesion molecule-3 (ICAM-3) for lymphocyte function-associated antigen-1 (LFA-1). *J. Biol. Chem.* **271,** 23920-23927.

Koch, C. A., Anderson, D., Moran, M. F., Ellis, C. & Pawson, T. (1991). SH2 and SH3 domains - elements that control interactions of cytoplasmic signaling domains. *Science,* **252,** 668-674.

Konishi, H., Ochiya, T., Chester, K. A., Begent, R. H. J., Muto, T., Sugimura, T. & Terada, M. (1998). Targeting strategies for gene delivery to carcinoembryonic antigen-producing cancer cells by retrovirus displaying a single-chain variable fragment antibody. *Hum. Gene Ther.* **9,** 235-248.

Kortt, A. A., Malby, R. L., Caldwell, J. B., Gruen, L. C., Ivancic, N., Lawrence, M. C., Howlett, G. J., Webster, R. G., Hudson, R. G. & Colman, P. M. (1994). Recombinant anti-sialidase single-chain variable fragment antibody. Characterization, formation of dimer and higher-molecular-mass multimers and the solution of the crystal structure of the single-chain variable fragment/sialidase complex. *Eur. J. Biochem.* **221,** 151-157.

Kraehenbuhl, J.-P. & Neutra, M. P. (1992). Molecular and cellular basis of immune protection of mucosal surfaces. *Physiol. Rev.* **72,** 853-879.

Krajči, P., Solberg, R., Sandberg, M., Øyen, O., Jahnsen, T. & Brandtzaeg, P. (1989). Molecular cloning of the human transmembrane secretory component (poly-Ig receptor) and its mRNA expression in human tissues. *Biochem. Biophys. Res. Commun.* **158,** 783-789.

Kratky, O. (1963). X-ray small angle scattering with substances of biological interest in diluted solutions. *Prog. Biophys. Chem.* **13,** 105-173.

Krop-Watorek, A., Oikawa, S., Oyama, Y. & Nakazato, H. (1998). Oligomerization of N-terminal domain of carcinoembryonic antigen (CEA) expressed in *Escherichia coli. Biochem. Biophys. Res. Commun.* **242,** 79-83.

Krueger, N. X. & Saito, H. (1992). *Proc. Natl. Acad. Sci. USA,* **89,** 7417-7421.

Krupey, J., Gold, P. & Freedman, S. O. (1968). Physicochemical studies of the carcinoembryonic antigens of the human digestive system. *J. Exp. Med.* **128,** 387-398.

Kuby, J. (1996). *Immunology. 3rd edition.* W. H. Freeman and company, New York.

Kuhn, L. C. & Kraehenbuhl, J. P. (1979). Interaction of rabbit secretory component with rabbit IgA dimer. *J. Biol. Chem.* **254,** 11066-11071.

Kuwahara, M., Kuroki, M., Arakawa, F., Senba, T., Matsuoka, Y., Hideshima, T.,

Yamashita, Y. & Kanda, H. (1996). A mouse/human-chimeric bispecific antibody-reactive with human carcinoembryonic antigen-expressing cells and human T-lymphocytes. *Cancer Res.* **16,** 2661-2667.

Kwong, P. D., Ryu, S. E., Hendrickson, W. A., Axel, R., Sweet, R. M., Folenawasserman, G., Hensley, P. & Sweet, R. W. (1990). Molecular characteristics of recombinant human CD4 as deduced from polymorphic crystals. *Proc. Natl. Acad. Sci. USA,* **87,** 6423-6427.

Labeit, S. & Kolmerer, B. (1995). Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science,* **270,** 293-296.

Lan, K. H., Kanai, F., Shiratori, Y., Okabe, S., Yoshida, Y., Wakimoto, H., Hamada, H., Tanaka, T., Ohashi, M. & Omata, M. (1996). Tumor-specific gene-expression in carcinoembryonic antigen-producing gastric-cancer cells using adenovirus vectors. *Gastroenterol.* **111,** 1241-1251.

Lan, K. H., Kanai, F., Shiratori, Y., Ohashi, M., Tanaka, T., Okudaira, T., Yoshida, Y., Hamada, H., & Omata, M. (1997). In vivo selective gene expression and therapy mediated by adenoviral vectors for human carcinoembryonic antigen-producing gastric carcinoma. *Cancer Res.* **57,** 4279-4284.

Lane, D. M., Eagle, K. F., Begent, R. H. J., Hope-Stone, L. D., Green, A. J. Green, Casey, J. L., Keep, P. A., Kelly, A. M. B., Ledermann, J. A., Glaser, M. G. & Hilson, A. J. W. (1994). Radioimmunotherapy of metastatic colorectal tumours with iodine-131-labelled antibody to carcinoembryonic antigen: phase I/II study with comparative biodistribution of intact and $F(ab')_2$ antibodies. *Brit. J. Cancer,* **70,** 521-525.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26,** 283-291.

Lattman, E. E. (1995). (Editor) Protein structure prediction: a special issue. *Proteins,* **23.**

Lea, S. & Stuart, D. (1995). Analysis of anitgenic surfaces of proteins. *FASEB J.* **9,** 87-93.

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992). Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2.6 Å resolution. *Cell,* **68,** 1145-1162.

Lee, B. & Richards, F. M. (1971). An interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55,** 379-400.

Lemke, G. & Axel, R. (1985). *Cell,* **40,** 501-508.

Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J. Mol. Biol. 136, 225-270.

Lesk, A. M. & Chothia, C. (1982). Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domain. *J. Mol. Biol.* **160,** 325-342.

Leusch, H. G., Hefta, S. A., Drzeniek, Z., Hummel, K., Markos-Pusztai, Z. and Wagener, C. (1990). *Escherichia coli* of human origin binds to carcinoembryonic antigen (CEA) and non-specific crossreacting antigen (NCA). *FEBS Lett.* **261,** 405-409.

Leusch, H. G., Drzeniek, Z. and Markos-Pusztai, Z. (1991). Binding to *Escherichia coli* and *Salmonella* strains to member of the carcinoembryonic antigen family: Differential binding inhibition by aromatic alpha-glycosides of mannose. *Infect. Immunity.* **59,** 2051-2057.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein folding: the endgame. *Annu. Rev. Biochem.* **66,** 549-579.

Lin, S. H. & Guidotti, G. (1989). Cloning and expression of a cDNA coding for a rat liver plasma membrane ecto-ATPase. The primary structure of the ecto-ATPase is similar to that of the human biliary glycoprotein I. *J. Biol. Chem.* **264,** 14408-14414.

Lindner, P., May, R. P. & Timmins, P. A. (1992). Upgrading of the SANS instrument D11 at the ILL. *Physica B*, **180**, 967-972.

Lisowska, E., Krop-Watorek, A. & Sedlaczek, P. (1983).The dimeric structure of carcinoembryonic antigen (CEA). *Biochem. Biophys. Res. Comm.* **115,** 206-211.

Loomes, L. M., Stewart, W. W., Mazengera, R. L., Senior, B. W. & Kerr, M. A. (1991). Purification and characterisation of human immunoglobulin IgA1 and IgA2 isotypes from serum. *J. Immunol. Methods,* **141,** 209-218.

Lucisano-Valim, Y. M. & Lachmann, P. J. (1991). The effect of antibody isotype and antigenic epitope density on the complement-fixing activity of immune complexes: a systematic study using chimaeric anti-NIP antibodies with human Fc regions. *Clin. Exp. Immunol.* **84,** 1-8.

MacArthur, M. W., Driscoll, P. C. & Thornton, J. M. (1994). NMR and crystallography - complementary approaches to structure determination. *Trends Biotech.* **12,** 149-153.

Madden, D. R., Gorga, J. C., Strominger, J. L. & Wiley, D. C. (1992). The three-dimensional structure of HLA-1327 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC. *Cell,* **70,** 1035-1048.

Maliezewski, C. R., March, C. J., Schoenborn, M. A., Gimpel, S. & Shen, L. (1990). Expression cloning of a human Fc receptor for IgA. *J. Exp. Med.* **172,** 1665-1672.

Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. (1980). Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0 and 1.9 Å resolution. *J. Mol. Biol.* **141,** 369-391.

Martin, A. C. R., Cheetham, J. C. & Rees, A. R. (1989). Modelling antibody hypervariable loops: A combined algorithm. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 9268-9272.

Matthews, B. W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33,** 491-497.

Mattu, T. J., Pleass, R. J., Willis, A. C., Kilian, M., Wormald, M. R., Lellouch, A. C., Rudd, P. M., Woof, J. M. & Dwek, R. A. (1998). The glycosylation and structure of human serum IgA1, Fab and Fc regions and the role of N-glycosylation on Fcα receptor interactions. *J. Biol. Chem.* **273,** 2260-2272.

Max, E. E. & Korsmeyer, S. J. (1985). Human J chain gene. *J. Exp. Med.* **161,** 832-849.

Mayans, M. O., Coadwell, W. J., Beale, D., Symons, D. B. A. & Perkins, S. J. (1995). Demonstration by pulsed neutron scattering that the arrangement of the Fab and Fc fragments in the overall structures of bovine IgG1 and IgG2 in solution is similar. *Biochem. J.* **311**, 283-291.

Mazanec, M. B., Nedrud, J. G. & Lamm, M. E. (1987). Immunoglobulin A monoclonal antibodies protect against Sendai virus. *J. Virol.* **61,** 2624-2626.

Mazanec, M. B., Kaetzel, C. S., Lamm, M. E., Fletcher, D. & Nedrud, J. G. (1992). Intracellular neutralization of virus by immunoglobulin A antibodies. *Proc.*

Natl. Acad. Sci. USA, **89,** 6901-6905.

Mazanec, M. B., Nedrud, J. G., Kaetzel, C. S. & Lamm, M. E. (1993). A three-tiered view of the role of IgA in mucosal defense. *Immunol. Today,* **14,** 430-435.

McAneny, D., Ryan, C. A., Beazley, R. M. & Kaufman, H. L. (1996). Results of a phase I trial of a recombinant vaccinia virus that expresses carcinoembryonic antigen in patients with advanced colorectal cancer. *Annals Surgical. Oncol.* **3,** 495-500.

McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature,* **348,** 552-554.

Merritt, E. A., Sarfaty, S., Vandenakker, F., Lhoir, C., Martial, J. A. & Hol, W. G. J. (1994). Crystal structure of cholera-toxin B-pentamer bound to receptor G(M1) pentasaccharide. *Protein Sci.* **3,** 166-175.

Meyer, D. F. (1994). Analysis of the Structural Changes that Occur During the Oxidation of Human Low Density Lipoproteins. *Ph.D Thesis*, University of London.

Mian, I. S., Bradwell, A. R. & Olson, A. J. (1991). Structure, function and properties of antibody binding sites. *J. Mol. Biol.* **217,** 133-151.

Michael, N. P., Chester, K. A., Melton, R. G., Robson, L., Nicholas, W., Boden, J. A., Pedley, R. B., Begent, R. H. J., Sherwood, R. F. & Minton, N. P. (1996). In vitro and in vivo characterisation of a recombinant carboxypeptidase G2: anti-CEA scFv fusion protein. *Immunotechnology,* **2,** 47-57.

Michetti, P., Mahan, M. J., Slauch, J. M., Mekalanos, J. J. & Neutra, M. R. (1992). Monoclonal secretory immunoglobulin-a protects mice against oral challenge with the invasive pathogen salmonella-typhimurium. *Infect. Immun.* **60,** 1786-1792.

Montelione, G. T., Wuthrich, K., Nice, E. C., Burgess, A. W. & Scheraga, H. A. (1987). Solution structure of murine epidermal growth factor: determination of the polypeptide backbone chain-fold by nuclear magnetic resonance and distance geometry. *Proc. Natl. Acad. Sci. USA,* **84,** 5226-5230.

Monteiro, R. C., Kubagawa, H., & Cooper, M. D. (1990). Cellular distribution, regulation and biochemical nature of an Fcα receptor in humans. *J. Exp. Med.* **148,** 597-613.

Monteiro, R. C., Hostoffer, R. W., Cooper, M. D., Bonner, J. R., Gartland, G. L. & Kubagawa, H. (1993). Definition of IgA receptors on eosinophils and their enhanced expression in allergic individuals. *J. Clin. Invest.* **92,** 1681-1685.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein-structure coordinates. *Proteins,* **12,** 345-364.

Morton, H. C., Atkin, J. D., Owens, R. J. & Woof, J. M. (1993). Purification and characterization of chimeric human IgA1 and IgA2 expressed in COS and Chinese hamster ovary cells. *J. Immunol.* **151,** 4743-4752.

Morton, H. C., van Egmond, M. & van de Winkel, J. G. J. (1996). Structure and function of human IgA Fc receptors (FcαR). *Crit. Rev. Immunol.* **16,** 423-440.

Murakami, M., Kuroki, M., Arakawa, F., Kuwahara, M., Oikawa, S., Nazakato, H. & Matsuoka, Y. (1995). A reference of the GOLD classification of monoclonal antibodies against carcinoembryonic antigen to the domain structure of the carcinoembryonic antigen molecule. *Hybridoma*, **14**, 19-28.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and

structures. *J. Mol. Biol.* **247,** 536-540.

Nagel, G., Grunert, F., Kuijpers, T. W., Watt, S. M., Thompson, J. & Zimmermann, W. (1993). Genomic organization, splice variants and expression of CGM1, a CD66-related member of the carcinoembryonic antigen gene family. *Eur. J. Biochem.* **214,** 27-35.

Nair, S. K., Boczkowski, D., Morse, M., Cumming, R. I., Lyerly, H. K. & Gilboa, E. (1998). Induction of primary carcinoembryonic antigen (CEA)-specific cytotoxic T lymphocytes *in vitro* using human dendritic cells transfected with RNA. *Nature Biotech.* **16,** 364-369.

Nap, M., Hammarström, M.-L., Börmer, O., Hammarström, S., Wagener, C., Handt, S., Schreyer, M., Mach, J.-P., Buchegger, F., von Kleist, S., Grunert, F., Seguin, P., Fuks, A., Holm, R & Lamerz, R. (1992).Specificity and affinity of monoclonal antibodies against carcinoembryonic antigen. *Cancer Res.* **52,** 2329-2339.

Navaza, J. (1994). AMoRe: an automated package for molecular replacement. *Acta Cryst.* **A50,** 157-163.

Nikula, T. K., Bocchia, M., Curcio, M. J., Sgouros, G., Ma, Y., Finn, R. D. & Scheinberg, D. A. (1995). Impact of high tyrosine fraction in complementary determining regions of radioiodination on IgG immunoreactivity. *Molec. Immunol.* **32,** 865-872.

Nikolova, E. B., Tomana, M. & Russell, M. W. (1994). All forms of human IgA antibodies bound to antigen interfere with complement (C3) fixation induced by IgG or by antigen alone. *Scand. J. Immunol.* **39,** 275-280.

Nissim, A. & Eshar, Z. (1992). The human mast cell receptor binding site maps to the third constant domain of immunglobulin E. *Mol. Immunol.* **9,** 1065-1072.

Oikawa, S., Nakazato, H. and Kosaki, G. (1987). Primary structure of human carcinoembryonic antigen (CEA) deduced from cDNA sequence. *Biochem. Biophys. Res. Commun.*. **142,** 511-518.

Oikawa, S., Inusuka, C., Kuroki, M., Matsuoka, Y., Kosaki, G. and Nakazato, H. (1989). Cell adhesion of non-specific cross-reacting antigen (NCA) and carcinoembryonic antigen (CEA) expressed on CHO cell surface: Homophilic and heterophilic adhesion. *Biochem. Biophys. Res. Commun.* **164,** 39-45.

Oikawa, S., Inuzuka, C., Kuroki, M., Arakawa, F., Matsuoka, Y., Kosaki, G. and Nakazato, H. (1991). A specific cell adhesion activity between members of carcinoembryonic antigen family, W272 and NCA, is mediated by N-domains. *J. Biol. Chem.* **266,** 7995-8001.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372,** 631-634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. (1997). CATH - a hierarchic classification of protein domain structures. *Structure,* **5,** 1093-1108.

Orlandi, R., Güssow, D. H., Jones, P. T. & Winter, G. (1989). Cloning immunoglobulin variable domains for expression by the polymerase chain reaction. *Proc. Natl. Acad. Sci. USA,* **86,** 3833-3837.

Orr, H. T., Lancet, D., Robb, R. J., Lopez de Castro, J. A. & Strominger, J. L. (1979). The heavy chain of human histocompatibility antigen HLA-B7 contains an Ig-like region. *Nature,* **282,** 266-270.

Osborn, L., Vassallo. C., Browning, B. G., Tizard, R., Haskard, D. O., Benjamin, C. D., Dougas, I. & Kirchhausen, T. (1994). Arrangement of domains, and amino-acid

residues required for binding of vascular cell-adhesion molecule-1 to its counter-receptor VLA-4 (alpha-4-beta-1). *J. Cell Biol.* **124,** 601-608.

Otwinowski, Z. (1993). Oscillation data reduction program. In: *Data collection and processing. Proceedings of the CCP4 study weekend 29-30 January 1993.* SERC Daresbury Laboratory, Warrington WA4 4AD,UK. ISSN 0144-5677.

Padlan, E. A. (1990). On the nature of antibody combining sites: unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins*, **7,** 112-124.

Padlan, E. A. (1991). A possible procedure for reducing the immunogenicity of antibody variable domains while preserving their ligand-binding properties. *Molec. Immunol.* **28,** 489-498.

Padlan, E. A. (1994). Anatomy of the antibody molecule. *Mol. Immunol.* **31,** 169-217.

Padlan, E. A., Abergel, C. & Tipper, J. P. (1995). Identification of specificity-determining residues in antibodies. *FASEB J.* **9,** 133-139.

Parker, M. J., Dempsey, C. E., Hosszu, L. L. P., Waltho, J. P. & Clarke, A. R. (1998). Topology, sequence evolution and folding dynamics of an immunoglobulin domain. *Nature Struct. Biol.* **5,** 194-198.

Pathy, L. (1993). Modular design of proteases of coagulation, fibrinolysis, and complement activation: implications for protein engineering and structure-function studies. *Methods. Enzymol.* **222,** 10-21.

Pavlenko, A. F., Chikalovets, I. V., Kurika, A. V., Glasunov, V. P., Mikhalyuk, L. V. & Ovodov, Yu. S. (1990). Carcinoembryonic antigen, its spatial structure and localisation of antigenic determinants. *Tumour Biol.* **11**, 306-318.

Paxton, R. J., Mooser, G., Pande, H., Lee, T. D. and Shively, J. E. (1987). Sequence analysis of carcinoembryonic antigen: identification of glycosylation sites and homology with the immunoglobulin supergene family. *Proc. Natl. Acad. Sci. USA,* **84,** 920-924.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA,* **85**, 2444-2448.

Pedersen, J. T. & Rees, A. R. (1993). Molecular modelling of anti-CEA antibodies. Unpublished data.

Pedersen, J. T., Searle, S., Henry, A. & Rees, A. R. (1992). Antibody modelling: Beyond homology. *Immunomethods*, **1**, 126-136.

Perisic, O., Webb, P. A., Holliger, P., Winter, G. & Williams, R. L. (1994). Crystal structure of a diabody, a bivalent antibody fragment. *Structure*, **2,** 1217-1226.

Perkins, S. J. (1985). Molecular modelling of human-complement subcomponent C1q and its complex with $C1r_2C1s_2$ derived from neutron-scattering curves and hydrodynamic analyses. *Biochem. J.* **228,** 13-26.

Perkins, S. J. (1986). Protein volumes and hydration effects: the calculation of partial specific volumes, neutron scattering matchpoints and 280 nm absorption coefficients for proteins and glycoproteins from amino acid sequences. *Eur. J. Biochem.* **157**, 169-180.

Perkins, S. J. (1988). X-ray and neutron solution scattering. In: *Modern physical methods in biochemistry, part B.* Editors A. Neuberger, and L. L. van Deenen. Elsevier science publishers B. V. pp143-265.

Perkins, S. J. (1994). High-flux X-ray and neutron solution scattering. *Methods Mol. Biol.* **22,** 39-60.

Perkins, S. J. & Weiss, H. (1983). Low resolution structural studies of mitochondrial

ubiquinol-cytochrome c reductase in detergent solutions by neutron scattering. *J. Mol. Biol.* **168**, 847-866.

Perkins, S. J., Chung, L. P. & Reid, K. B. M. (1986). Unusual ultrastructure of complement component C4b-binding protein of human complement by synchrotron X-ray scattering and hydrodynamic analysis. *Biochem. J.* **233**, 799-807.

Perkins, S. J., Smith, K. F., Amatayakul, S., Ashford, D., Rademacher, T. W., Dwek, R. A., Lachmann, P. J. & Harrison, R. A. (1990). The two-domain structure of the native and reaction centre cleaved forms of C1 inhibitor of human complement by neutron scattering. *J. Mol. Biol.* **214**, 751-763.

Perkins, S. J., Nealis, A. S., Sutton, B. J. & Feinstein, A. (1991). Solution structure of human and mouse immunoglobulin-M by synchrotron X-ray scattering and molecular graphics modelling - a possible mechanism for complement activation. *J. Mol. Biol.* **221,** 1345-1366.

Perkins, S. J., Smith, K. F., Kilpatrick, J. M., Volanakis, J. E. & Sim, R. B. (1993). Modelling of the serine protease fold by X-ray and neutron scattering and sedimentation analyses: its occurrence in factor D of the complement system. *Biochem. J.* **295**, 87-99.

Perkins, S. J., Ashton, A. W., Boehm, M. K. & Chamberlain, D. C. (1998a). Molecular structures from low angle X-ray and neutron scattering strudies. *Int. J. Biol. Macromol.* **22,** 1-16.

Perkins, S. J., Ullman, C. G., Brisset, N. C., Chamberlain, D. C. & Boehm, M. K. (1998b). Analogy and solution scattering modelling: new structural strategies for the multidomain proteins of complement, cartilage and the immunoglobulin superfamily. *Immunol. Rev.* **163,** 237-250.

Pervin, S., Chakraborty, M., Bhattacharya-Chatterjee, M., Zeytin, H., Foon, K. A. & Chatterjee, S. K. (1997). Induction of antitumour immunity by an anti-idiotype antibody mimicking carcinoembryonic antigen. *Cancer Res.* **57,** 728-734.

Peterson, P. A., Cunningham, B. A., Berggard, I. & Edelman, G. M. (1972). β_2 microglobulin - a free Ig domain. *Proc. Natl. Acad. Sci. USA,* **82,** 1697-1701.

Pfuhl, M. & Pastore, A. (1995). Tertiary structure of an immunoglobulin-like domain from the giant muscle protein titin: a new member of the I set. *Structure,* **3,** 391-401.

Pfuhl, M., Improta, S., Politon, A. S. & Pastore, A. (1997). When a module is also a domain: the role of the N terminus in the stability and the dynamics of immunoglobulin domains from titin. *J. Mol. Biol.* **265,** 242-256.

Pilz, I., Kratky, O., Licht, A. & Sela, M. (1973). Shape and volume of anti-poly(D-alanyl) antibodies in the presence and absence of tetra-D-alanine as followed by small-angle X-ray scattering. *Biochemistry* **12**, 4998-5005.

Pluckthun, A. (1990). Antibodies from *Eschericia coli. Nature,* **347,** 497-.

Poljak. R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P. & Saul, F. (1973). Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2.8 Å resolution. *Proc. Natl. Acad. Sci. USA,* **70,** 3305-3310.

Ponder, & Richards, F. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193,** 775-791.

Powell, M. J. D. (1977). Restart procedures for the conjugate gradient method.

Mathematical Programming, **12,** 241-254.
Prahl, J. W., Abel, C. A. & Grey, H. M. (1971). Carboxy-terminal structure of the α chain of human IgA myeloma proteins. *Biochemistry*, **10**, 1808-1812.
Putnam, F. W., Lu, Y.-S. V. & Low, T. L. K. (1979). Primary structure of human IgA1 immunoglobulin. *J. Biol. Chem.* **254,** 2865-2874.
Raag, R. & Whitlow, M. (1995). Single-chain Fvs. *FASEB J.* **9,** 73-80.
Rao, S. T. & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76,** 241-256.
Ramachandran, G. N. & Sassiekharan, V. (1968). Conformation of polypeptides and proteins. *Advan. Protein Chem.* **23,** 283-437.
Read, D. A., Chester, K. A., Keep, P. A., Begent, R. H. J., Pedersen, J. T. & Rees, A. R. (1995). Mutagenesis of single-chain antibody MFE-23 and its effect on affinity for CEA. *Brit. J. Cancer*, **71**, 57-57. Supplement XIV: abstract P132.
Rebstock, S., Lucas, K., Thompson, J. A. & Zimmermann, W. (1990). cDNA and gene analyses imply a novel structure for a rat carcinoembryonic antigen-related protein. *J. Biol. Chem.* **265,** 7872-7879.
Recny, M. A., Luther, M. A., Knoppers, M. H., Neidhardt, E. A., Khandekar, S. S., Concino, M. F., Schmike, P. A., Francis, M. A., Moebius, U., Reinhold, B. B., Reinhold, V. N. & Reinherz, E. L. (1992). N-glycosylation is required for human CD2 immunoadhesion functions. *J. Biol. Chem.* **267,** 22428-22424.
Rees, A. R., Searle, S. J., Henry, A. H., Whitelegg, N. & Pedersen, J. (1996). Antibody combining sites: structure and prediction. In: *Protein structure prediction.* Editor, M. J. E. Sternberg. Oxford University Press, Oxford. pp141-172.
Renegar, K. B. & Small Jr, P. A. (1991a). Passive transfer of local immunity to influenza virus infection by IgA antibody. *J. Immunol.* **146,** 1972-1978.
Renegar, K. B. & Small Jr, P. A. (1991b). Immunoglobulin A mediation of murine nasal anti-influenza virus immunity. *J. Virol.* **65,** 2146-2148.
Renegar, K. B., Jackson, G. D. F. & Mestecky, J. (1998). In vitro comparison of the biologic activities of monoclonal monomeric IgA, polymeric IgA, and secretory IgA. *J. Immunol.* **160,** 1219-1223.
Reth, M. (1989). Antigen receptor tail clue. *Science,* **338,** 383-384.
Riechmann, L., Clark, M., Waldmann, H. & Winter, G. (1988). Reshaping human antibodies for therapy. *Nature,* **332,** 323-327.
Rhodes, G. (1993). *Crystallography made crystal clear.* Academic Press Ltd., London.
Richards, C. A., Austin, E. A. & Huber, B. E. (1995). Transcriptional regulatory sequences of carcinoembryonic antigen - identification and use with cytosine deaminase for tumor-specific gene therapy. *Hum. Gene Ther.* **6,** 881-893.
Richardson, J. & Richardson, D. C. (1989). Principles and patterns of protein conformation. In: *Prediction of protein structure and the principles of protein conformation.* Editor G. Fasman. Plenum, New York. pp1-98.
Richmond, T. J. & Richards, F. M. (1978). Packing of α-helices: geometrical constraints and contact areas. *J. Mol. Biol.* **119**, 537-555.
Robson, B. & Pain, R. H. (1971). Analysis of the code relating sequence to the conformation in protteins: possible implications for the mechanism of formation of helical regions. *J. Mol. Biol.* **58**, 237-259.
Robson, B. & Suzuki, E. (1976). Conformational properties of amino acids residues in globuar proteins. *J. Mol. Biol.* **107**, 327-356.
Roguska, M. A., Pederson, J. T., Keddy, C. A., Henry, A. H., Searle, S. J., Lambert, J.

M., Goldmacher, V. S., Blattler, W. A., Rees, A. R. & Guild, B. C. (1994). Humanization of murine monoclonal antibodies through variable domain resurfacing. *Proc. Natl. Acad. Sci. USA,* **91,** 969-973.

Rojas, M., Fuks, A. & Stanners, C. P. (1990). Biliary glycoprotein (BGP), a member of the immunoglobulin supergene family, functions in vitro as a Ca^{++}-dependent intercellular adhesion molecule. *Cell Growth Differ.* **1,** 527-533.

Rojas, M., Demarte, L., Screaton, R. A. & Stanners, C. P. (1996). Radical differences in functions of closely-related members of the human carcinoembryonic antigen gene family. *Cell Growth Differ.* **7,** 655-662.

Rossmann, M. G., Moras, D., and Olsen, K. W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature* **250**, 194-199.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599.

Rost, B., Sander, C. & Schneider, R. (1994). PHD - An automatic mail server for protein secondary structure prediction. *CABIOS*, **10**, 53-60.

Rudd, P. M., Fortune, F., Patel, T., Parekh, R. B., Dwek, R. A., & Lehner, T. (1994). A human T-cell receptor recognises O-linked sugars from the hinge region of human IgA1 and IgD. *Immunology,* **83,** 99-106.

Ruoslahti, E. & Öbrink, B. (1996). Common principles in cell adhesion. *Exp. Cell Res.* **227,** 1-11.

Russell, M. W., Reinholdt, W. J. & Kilian, M. (1989). Anti-inflammatory activity of human IgA antibodies and their Fab alpha fragments: inhibition of IgG-mediated complement activation. *Eur. J. Immunol.* **19,** 2243-2249.

Russell, S. J., Hawkins, R. E. & Winter, G. (1993). Retroviral vectors displaying functional antibody fragments. *Nucl. Acids Res.* **21,** 1081-1085.

Ryu, S. -E., Kwong, P. D., Truneh, A., Porter, T. G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N. -h., Axel, R., Sweet, R. W. & Hendrickson, W. A. (1990). Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature,* **348,** 419-426.

Ryu, S. -E., Truneh, A., Sweet, R. W. & Hendrickson, W. A. (1994). Structures of an HIV and MHC binding fragment from human CD4 as refined in two crystal lattices. *Structure,* **2,** 59-74.

Šali, A. & Blundell, T. L. (1990). The definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212,** 403-428.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins,* **9,** 58-68.

Sauer-Eriksson, A. E., Kleywegt, G. J., Uhlen, M. & Jones, T. A. (1995). Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. *Structure,* **3,** 265-278.

Saul, F. A. & Poljak, R. J. (1993). Structural patterns at residue positions 9, 18, 67 and 82 in the V_H framework regions of human and murine immunoglobulins. *J. Mol. Biol.* **230,** 15-20.

Sarmay, G., Lund, G., Rozsnyay, Z., Gergely, J. & Jefferis, R. (1992). Mapping and comparison of the interaction sites on the Fc region of IgG responsible for triggering antibody dependent cellular cytotoxicity (ADCC) through different types of human Fcγ receptor. *Mol. Immunol.* **29,** 633-639.

Schlom, J., Eggensperger, D., Colcher, D., Molinolo, A., Houchens, D., Miller, L. S., Hinkle, G. & Siler, K. (1992). Therapeutic advantage of high-affinity anticarcinoma radioimmunoconjugates. *Cancer Res.* **52,** 1067-1072.

Schreuder, H., Tardif, C., Trump-Kallmeyer, S., Soffientini, A., Sarubbi, E., Akeson, A., Bowlin, T., Yanofsky, S. & Barrett, R. W. (1997). A new cytokine-receptor binding mode revealed by the crystal structure of the IL-1 receptor with an antagonist. *Nature,* **386,** 194-200.

Screaton, R. A., Penn, L. Z. & Stanners, C. P. (1997). Carcinoembryonic antigen, a human tumour marker, cooperates with Myc and Bcl-2 in cellular transformation. *J. Cell. Biol.* **137,** 939-952.

Schwarz, K., Mehnert-Solzer, C., von Kleist, S. & Grunert, F. (1988). Analysis of the specificity of CEA reactive monoclonal antibodies. Immunological support for the domain model of CEA. *Molec. Immun.* **25**, 889-898.

Semenyuk, A. V., & Svergun, D. I. (1991). GNOM - a program package for small-angle scattering data-processing. *J. Appl. Crystallogr.* **24,** 537-540.

Senior, B. W., Loomes, L. M. & Kerr, M. A. (1991). Microbial IgA proteases and virulence. *Revs. Med. Microbiol.* **2,** 200-207.

Shapiro, L., Fannon, A. M., Kwong, P. D., Thompson, A., Lehmann, M. S., Grubel, G., Legrand, J. F., Als-Neilsen, J., Colman, D. R., & Hendrickson, W. A. (1995). Structural basis of cell-cell adhesion by cadherins. *Nature,* **374,** 327-337.

Shapiro, L., Doyle, J. P., Hensley, P., Colman, D. R. & Hendrickson, W. A. (1996). Crystal structure of the extracellular domain from P_0, the major structural protein of peripheral nerve myelin. *Neuron,* **17,** 435-449.

Shen, L., Lasser, R. & Fanger, M. W. (1989). My43, a monoclonal antibody that reacts with human myeloid cells inhibits monocyte IgA binding and triggers function. *J. Immunol.* **143,** 4117-4122.

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. & Davies, D. R. (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci. USA,* **84,** 8075-8079.

Shogren, R., Gerken, T. A. & Jentoff, N. (1989). Role of glycosylation on the conformation and chain dimensions of O-linked glycoproteins - light-scattering studies of ovine submaxillary mucin. *Biochemistry,* **28,** 5525-5536.

Sippel, C. J., McCollum, M. J. & Perlmutter, D. H. (1994). Bile-acid transport by the rat-liver canalicular bile-acid transport ecto-ATPase protein is dependent on ATP but not on its own ecto-ATPase activity. *J. Biol. Chem.* **269,** 2820-2826.

Sippel, C. J., Shen, T. X. & Perlmutter, D. H. (1996). Site-directed mutagenesis within an ectoplasmic ATPase consensus sequence abrogates the cell aggregating properties of the rat liver canalicular bile acid transporter ecto-ATPase cell CAM 105 and carcinoembryonic antigen. *J. Biol. Chem.* **271,** 33095-33104.

Skubitz, K. M., Ducker, T. P. & Gouli, S. A. (1992). CD66 monoclonal antibodies recognize a phosphotyrosine-containing protein bearing a carcinoembryonic antigen cross-reacting antigen on the surface of human neutrophils. *J. Immunol.* **148,** 852-860.

Skubitz, K. M., Micklem, K. & van der Schoot, C. E. (1995a). Summary of the CD66 and CD67 cluster report. In: *Leukocyte typing Vol. 1.* Editors, S. Schlossma, L. Boumsell, W. Gilles, J. Hartan, T. Kishimoto, C. Morimoto, J. Ritz, S. Shaw, R. Silverstein, T. Springer, T. Tedder, & R. Todd. Oxford University Press, Oxford, UK. pp889-899.

Skubitz, K. M., Campbell, K. D., Ahmed, K. & Skubitz, A. P. N. (1995b). CD66 family members are associated with tyrosine kinase activity in human neutrophils. *J. Immunol.* **155,** 5382-5390.

Slayter, H. S. & Codington, J. F. (1973). Size and configuration of glycoprotein fragments cleaved from tumor cells by proteolysis. *J. Biol. Chem.* **248,** 3405-3410.

Slayter, H. S. & Coligan, J. E. (1975). Electron microscopy and physical characterization of the carcinoembronic antigen. *Biochemistry,* **14,** 2323-2330.

Smith, K. F., Harrison, R. A. & Perkins, S. J. (1990). Structural comparisons of the native and reaction centre cleaved forms of α_1-antitrypsin by neutron and X-ray solution scattering. *Biochem. J.* **267,** 203-212.

Srinivasan, N., Guruprasad, K. & Blundell, T. L. (1996). Comparative modelling of proteins. In: *Protein structure prediction.* Editor M. J. E. Sternberg. Oxford University Press, Oxford. pp111-140.

Stanfield, R. L., Fieser, T. M., Lerner, R. A. & Wilson, I. A. (1990). Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 Å. *Science,* **248,** 712-719.

Stanners, C. P., DeMarte, L., Rojas, M., Gold, P. and Fuks, A. (1995). Opposite functions for two classes of genes of the human carcinoembryonic antigen family. *Tumour Biol.* **16,** 23-31.

Staunton, D. E., Dustin, M. L., Erickson, H. P. & Springer, T. A. (1990). The arrangement of the immunoglobulin-like domains of ICAM-1 and the binding sites for LFA-1 and rhinovirus. *Cell,* **61,** 243-254.

Stefanova, I., Horejsi, V. Ansotegui, I. J., Knapp, W. & Stockinger, H. (1991). GPI-anchored cell-surface molecules complexed to protein tyrosine kinases. *Science,* **254,** 1016-1019.

Stern, L. J., Brown, J. H., Jandetzky, T. J., Gorga, J. C., Urban, R. G., Strominger, J. L. & Wiley, D. C. (1994). Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature,* **368,** 215-221.

Streydio, C., Lacka, K., Swillens, S. & Vassart, G. (1988). The human pregnancy-specific beta 1-glycoprotein (PS beta G) and the carcinoembryonic antigen (CEA)-related proteins are members of the same multigene family. *Biochem. Biophys. Res. Commun.* **154,** 130-137.

Suh, S. W., Bhat, T. N., Navia, M. A., Cohen, G. H., Rao, D. N., Rudikoff, S. & Davis, D. R. (1986). The galactan-binding immunoglobulin Fab J539. An X-ray diffraction study at 2.6 Å resolution. *Proteins Struct. Funct. Genet.* **1,** 74-80.

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). Knowledge-based modelling of homologous proteins. Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engin.* **1,** 377-384.

Sutton, B. J. & Phillips, D. C. (1983). The three-dimensional structure of the carbohydrate within the Fc fragment of immunoglobulin G. *Biochem. Soc. Transact.* **11,** 130-132.

Svehag, S. E. & Bloth, B. (1970). Ultrastructure of secretory and high-polymer imuunoglobulin A of human and rabbit origin. *Science,* **168,** 847-850.

Svenberg, T. (1976). Carcinoembryonic antigen-like substances from human bile. Isolation and partial characterization. *Int. J. Cancer,* **17,** 588-596.

Svergun, D. I (1992). Determination of the Regularization Parameter in Indirect-

Transform Methods Using Perceptual Criteria. *J. Appl. Cryst.* **25,** 495-503.

Svergun, D. I., Semenyuk, A. V., & Feigin, L. A. (1988). Small-angle-scattering-data treatment by the regularization method. *Acta Crystallogr.* **A44,** 244-250.

Svergun, D. I., Richard, S., Koch, M. H. J., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). Protein hydration in solution: Experimental observation by X-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA,* **95**, 2267-2272.

Takahashi, H., Tamura, H., Shimba, N., Shimada, I. & Arata, Y. (1994). Role of the domain-domain interaction in the construction of the antigen combining site - a comparative study by H-1-N-15 shift correlation nmr spectroscopy of the Fv and Fab fragments of antidansyl mouse monoclonal-antibody. *J. Mol. Biol.* **243,** 494-503.

Tanaka, T., Kanai, F., Okabe, S., Yoshida, Y., Wakimoto, H., Hamada, H., Shiratori, Y., Lan, K. H., Ishitobi, M. & Omata, M. (1996). Adenovirus-mediated prodrug gene-therapy for carcinoembryonic antigen-producing human gastric-carcinoma cells *in vitro*. *Cancer Res.* **56,** 1341-1345.

Tanaka, T., Kanai, F., Lan, K. H., Ohashi, M., Shiratori, Y., Yoshida, Y., Hamada, H. & Omata, M. (1997). Adenovirus-mediated gene therapy of gastric carcinoma using cancer-specific gene expression *in vivo*. *Biochem. Biophys. Res. Commun.* **231,** 775-779.

Taylor, W. R. (1988). A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* **28**, 161-169.

Teglund, S., Olsen, A., Khan, W. N., Frängsmyr, L. & Hammarström, S. (1994). The pregnancy-specific glycoprotein (PSG) gene cluster on human chromosome 19: fine structures of the 11 PSG genes and identification of 6 new genes forming a third subgroup within the carcinoembryonic antigen (CEA) family. *Genomics,* **23,** 669-684.

Thompson, J. A. (1995). Molecular cloning and expression of carcinoembryonic antigen gene family members. *Tumour Biol.* **16,** 10-16.

Thompson, J. A., Mauch, E.-M., Chen, F.-S., Hinoda, Y., Schrewe, H., Berling, B., Barnert, S., von Kleist, S., Shively, J. E. & Zimmermann, W. (1989). Analysis of the size of the carcinoembronic antigen (CEA) gene family: isolation and sequencing of N-terminal domain exons. *Biochem. Biophys. Res. Commun.* **158,** 996-1004.

Thompson, J., Koumari, R., Wagner, K., Barnert, S., Schleussner, C., Schrewe, H., Zimmermann, W., Müller, G., Schempp, W., Zaninetta, D., Ammaturo, D. & Hardman, N. (1990). The human pregnancy-specific glycoprotein genes are tightly linked on the long arm of chromosome 19 and are coordinately expressed. *Biochem. Biophys. Res. Commun.* **167,** 848-859.

Thompson, J. A., Grunert, F. and Zimmermann, W. (1991). The carcinoembryonic antigen gene family: Molecular biology and clinical perspectives. *J. Clin. Lab. Anal.* **5,** 344-366.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.***22,** 4673-4680.

Thomsen, N. K., Soroka, V., Jensen, P. H., Berezin, V., Kiselyov, V. V., Bork, E. & Porben, F. M. (1996). The three-dimensional structure of the first domain of neural cell adhesion molecule. *Nat. Struct. Biol.* **3,** 581-585.

Torano, A. & Putnam, F. W. (1978). Complete amino acid sequence of the α2 heavy chain of a human IgA2 immunoglobulin of the A2m(2) allotype. *Proc. Natl. Acad. Sci. USA,* **75,** 966-969.

Towns-Andrews, E., Berry, A., Bordas, J., Mant, G. R., Murray, P. K., Roberts, K., Sumner, I., Worgan, J. S., Lewis, R. & Gabriel, A. (1989). Time-resolved X-ray diffraction station: X-ray optics, detectors and data acquisition. *Rev. Sci. Instrum.* **60,** 2346-2349.

Tsuzukida, Y., Wang, C. C. & Putnam, F. W. (1979). Structure of the A2m(1) allotype of human IgA-A recombinant molecule. *Proc. Natl. Acad. Sci.USA,* **76,** 1104-1108.

Underdown, B. J. & Schiff, M. J. (1986). Immunoglobulin A: strategic defense initiative at the mucosal surface. *Annu. Rev. Immunol.* **4,** 389-417.

van der Merwe, P. A., Brown, M. H., Davis, S. J. & Barclay, A. N. (1993). *Biochem. Soc. Trans.* **21,** 340S.

Vasudevan, S. G., Armargeo, W. L. F., Shaw, D. C., Lilley, P. E., Dixon, N. E., and Poole, R. K. (1991). Isolation and nucleotide-sequence of the hmp gene that encodes a hemoglobin-like protein in *Escherichia coli* K-12. *Molec. Gen. Genetics* **226,** 49-58.

Verhaar, M. J., Chester, K. A., Keep, P. A., Robson, L., Pedley, R. B., Boden, J. A., Hawkins, R. E. & Begent, R. H. J. (1995). A single chain Fv derived from a filamentous phage library has distinct tumour targeting advantages over one derived from a hybridoma. *Int. J. Cancer,* **61,** 497-501.

Verhaar, M. J., Keep, P. A., Hawkins, R. E., Robson, L., Casey, J. L., Pedley, B., Boden, J. A., Begent, R. H. J. & Chester, K. A. (1996). ^{99m}Tc radiolabelling using a phage-derived single chain Fv with C-terminal cysteine for colorectal tumour imaging. *J. Nucl. Med.* **37,** 868-872.

Vigers, G. P. A., Anderson, L. J., Caffes, P. & Brandhuber, B. J. (1997). Crystal structure of the type-I interleukin-1 receptor complexed with interleukin-1β. *Nature,* **386,** 190-194.

Virji, M., Makepeace, K., Ferguson, D. J. P. & Watt, S. M. (1996). Carcinoembryonic antigens (CD66) on epithelial cells and neutrophils are receptors for Opa proteins of pathogenic Neisseriae. *Mol. Microbiol.* **22,** 941-950.

von Kleist, S., Chavanel, G. & Burtin, P. (1972). Identification of an antigen from normal human tissue that crossreacts with the carcinoembryonic antigen. *Proc. Natl. Acad. Sci. USA,* **69,** 2492-2494.

Wagner, H. E., Toth, C. A., Steele, G. D. & Thomas, P. (1992). Invasive and metastatic potential of human colorectal cancer cell lines: relationship to cellular differentiation and carcinoembryonic antigen production. *Clin. Exp. Metastasis,* **10,** 25-31.

Wako, H. & Blundell, T. L. (1994*a*). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins I. solvent accessibility classes. *J. Mol. Biol.* **238,** 682-692.

Wako, H. & Blundell, T. L. (1994*b*). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins II. secondary structures. *J. Mol. Biol.* **238,** 693-708.

Waksman, G., Kominos, D., Robertson, S. C., Pant, N., Baltimore, D., Birge, R. B.,

Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., Resh, M. D., Rios, C. B., Silverman, L. & Kuriyan, J. (1992). Crystal structure of the phosphotyrosine recognition domains SH2 of v-src complexed with tyrosine- phosphorylated peptides. *Nature,* **358,** 646-653.

Wang, J., Yan, Y., Garrett, T. P. J., Liu, J., Rodgers, D. W., Garlick, R. L., Tarr, G. E., Husain, Y., Reinherz, E. L. & Harrison, S. C. (1990). Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature,* **348,** 411-419.

Wang, J. H., Pepinsky, B., Stehle, T., Lin, J. H., Karpusas, M., Browning, B. & Osborn, L. (1995). The crystal structure of an N-terminal two-domain fragment of vascular cell adhesion molecule 1 (VCAM-1): a cyclic peptide based on the domain 1 C-D loop can inhibit VCAM-1-α4 integrin interaction. *Proc. Natl. Acad. Sci. USA,* **92,** 5714-5718.

Wang, J., Stehle, T., Pepinsky, B., Liu, J., Karpusas, M. & Osborn, L. (1996). Structure of a functional fragment of VCAM-1 refined at 1.9 Å resolution. *Acta Cryst.* **D52,** 369-379.

Watanabe, S. and Chou, J. Y. (1988). Isolation and characterization of complementary DNAs encoding human pregnancy-specific beta 1-glycoprotein. *J. Biol. Chem.* **263,** 2049-2054.

Weisbart, R. H., Kacena, A., Schuh, A. & Golde, D. W. (1988). GM-CSF induces human neutrophil-mediated phagocytosis by an IgA Fc receptor activation mechanism. *Nature,* **332,** 647-648.

Weiss, A. (1993). T-cell antigen receptor signal transduction - a tale of tails and cytoplasmic protein-tyrosine kinases. *Cell,* **73,** 209-212.

Wetlaufer, D. B. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proc. Natl. Acad. Sci. USA* **70,** 697-701.

Wignall, G. D. & Bates, F. S. (1987). Absolute calibration of small angle neutron scattering data. *J. Appl. Crystallogr.* **20,** 28-40.

Williams, A. (1987). A year in the life of the immunoglobulin superfamily. *Immunol. Today,* **8,** 298-303.

Williams, A. F. & Barclay, A. N. (1988). The immunoglobulin superfamily - domains for cell surface recognition. *Annu. Rev. Immunol.* **6,** 381-405.

Williams, A. F. & Gagnon, J. (1982). Neuronal cell Thy-1 glycoprotein: homology with Ig. *Science,* **216,** 696-703.

Williamson, M. P. (1994). The structure and function of proline-rich regions in proteins. *Biochem. J.* **297,** 249-260.

Wilson, I. A. & Stanfield, R. L. (1993). Antibody-antigen interactions. *Curr. Opin. Struct. Biol.* **3,** 113-118.

Wilson, I. A. & Stanfield, R. L. (1994). Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.* **4,** 857-867.

Winner, L. S., Mack, J., Weltzin, R. A., Mekalanos, J. J., Kraehenhubl, J.-P. & Neutra, M. R. (1991). New model for analysis of mucosal immunity: intestinal secretion of specific monoclonal immunoglobulin A from hybridoma tumors protects against *Vibrio cholerae* infection. *Infect. Immun.* **59,** 977-982.

Withka, J. M., Wyss, D. F., Wagner, G., Arulanandam, A. R. N., Reinherz, E. L. & Recny, M.M A. (1993). Structure of the glycosylated adhesion domain of human T lymphocyte glycoprotein CD2. *Structure,* **1,** 69-81.

Wittekind, C. (1995). Carcinoembryonic antigen family members as diagnostic tools

in immunohistopathology. *Tumour Biol.* **16,** 42-47.

Worgan, J. S., Lewis, R., Fore, N. S., Sumner, I. L., Berry, A., Parker, B., D'Annunzio, F., Martin-Fernandez, M. L., Towns-Andrews, E., Harries, J. E., Mant, G. R., Diakun, G. P., & Bordas, J. (1990). The application of multiwire X-ray detectors to experiments using synchrotron radiation. *Nuclear Instruments and Methods in Physics Research*, **A291**, 447-454.

Worcester, D.L. (1988). Contrast variation and the versatility of deuterium in structural studies of biological macromolecules. *J. Appl. Cryst.* **21**, 669-674.

Wu, T. T. & Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132,** 211-249.

Wu, H., Myszka, D. G., Tendian, S. W., Brouillette, C. G., Sweet, R. W., Chaiken, I. M. & Hendrickson, W. A. (1996). Kinetic and structural analysis of mutant CD4 receptors that are defective in HIV gp120 binding. *Proc. Natl. Acad. Sci. USA,* **93,** 15030-15035.

Wyss, D. F., Choi, J. S., Li, J, Knoppers, M. H., Willis, K. J., Arulanandam, A. R. N., Smolyar, A., Reinherz, E. L. & Wagner, G. (1995). Conformation and function of the N-linked glycan in the adhesion domain of human CD2. *Science,* **269,** 1273-1278.

Yamashita, K., Totani, K., Kuroki, M., Matsuoka, Y., Ueda I. & Kobata, A. (1987). Structural studies of the carbohydrate moieties of carcinoembryonic antigens. *Cancer Res.* **47**, 3451-3459.

Yamashita, K., Totani, K., Iwaki, Y., Kuroki, M., Matsuoka, Y., Endo, T. & Kobata, A. (1989). Carbohydrate structures of nonspecific cross-reacting antigen-2, a glycoprotein purified from meconium as an antigen cross-reacting with anticarcinoembryonic antigen antibody. *J. Biol. Chem.* **264**, 17873-17881.

Yan, Z. F., Deng, X. B., Chen, M. X., Xu, Y., Ahram, M., Sloane, B. F. & Friedman, E. (1997). Oncogenic c-Ki-ras but not oncogenic c-Ha-ras up-regulates CEA expression and disrupts basolateral polarity in colon epithelial cells. *J. Biol. Chem.* **272,** 27902-27907.

Yang, C., Kratzin, H., Götz, H. & Hilschmann, N. (1979). Die Primästruktur eines monoklonalen IgA1-Immunoglobulins (Myelomprotein Tro). VII. Darstellung, Reinigung und Charackterisierung der Disulfidbrücken. *Hoppe-Seyler's Z. Physiol. Chem.* **360**, 1919-1940.

Yokota, T., Milenic, D. E., Whitlow, M. & Schlom, J. (1992). Rapid tumor penetration of a single-chain Fv and comparison with other immunoglobulin forms. *Cancer Res.* **52,** 3402-3408.

Yoshioka, T., Masuko, T., Kotanagi, H., Aizawa, O., Saito, Y., Nakazato, H., Koyama, K. & Hashimoto, Y. (1998). Homotypic adhesion through carcinoembryonic antigen plays a role in hepatic metastasis development. *Japan. J. Cancer Res.* **89,** 177-185.

Young, A. C. M., Zhang, W., Sacchettini, J. C. & Nathenson, S. G. (1994). The three-dimensional structure of H-2D^b at 2.4 Å resolution - implications for antigen-determinant selection. *Cell,* **76,** 39-50.

Zbar, A. P., Lemoine, N. R., Wadhwa, M., Thomas, H., Snary, D. & Kmiot, W. A. (1998). Biological therapy: approaches in colorectal cancer. Strategies to enhance carcinoembryonic antigen (CEA) as an immunogenic target. *Br. J. Cancer,* **77,** 683-693.

Zdanov, A., Li, Y., Bundle, D. R., Deng, S-J., MacKenzie, C. R., Narang, S. A., Young, N. M. & Cygler, M. (1994). Structure of a single-chain antibody variable domain (Fv) fragment complexed with a carbohydrate antigen at 1.7-Å resolution. *Proc. Natl. Acad. Sci. U.S.A.* **91,** 6423-6427.

Zeng, Z.-H., Castaño, A. R., Segelke, B. W., Stura, E. A., Peterson, P. A. & Wilson, I. A. (1997). Crystal structure of mouse CD1: an MHC-like fold with a large hydrophobic binding groove. *Science,* **277,** 339-345.

Zhou, H., Fuks, A., Alcaraz, G., Bolling, T. J. and Stanners, C. P. (1993). Homophilic adhesion between Ig superfamily carcinoembryonic antigen molecules involves double reciprocal bonds. *J. Cell Biol.* **122,** 951-960.

Zikan, J., Mestecky, J., Kulhavy, R. & Bennet, J. C. (1986). The stochiometry of J chain in human dimeric IgA. *Mol. Immunol.* **23,** 541-544.

Zimmermann, W., Weber, B., Ortlieb, B., Rudert, F., Schempp, W., Fiebig, H.-H., Shively, J. E., von Kleist, S. & Thompson, J. A. (1988). Chromosomal localization of the carcinoembryonic antigen gene family and differential expression in various tumors. *Cancer Res.* **48,** 2550-2554.

Zimmermann, W., Weiss, M. & Thompson, J. A. (1989). cDNA cloning demonstrates the expression of pregnancy-specific glycoprotein genes, a subgroup of the carcinoembryonic antigen gene family, in fetal liver. *Biochem. Biophys. Res. Commun.* **163,** 1197-1209.

Publications

Boehm, M. K., Mayans, M. O., Thornton, J. D., Begent, R. H. J., Keep, P. A. & Perkins, S. J. (1996). Extended glycoprotein structure of the seven domains in human carcinoembryonic antigen by X-ray and neutron solution scattering and an automated curve fitting procedure: implications for cellular adhesion. *J. Mol. Biol.* **259,** 718-736.

Ashton, A. W., Boehm, M. K., Gallimore, J. R., Pepys, M. B. & Perkins, S. J. (1997). Pentameric and decameric structures in solution of the serum amyloid P component by X-ray and neutron scattering and molecular modelling analyses. *J. Mol. Biol.* **272,** 408-422.

Ashton, A. W., Boehm. M. K., Johnson, D. J. D., Kemball-Cook, G. & Perkins, S. J. (1998). The solution structure of human coagulation factor VIIa in its complex with tissue factor is similar to free factor VIIa: a study of a heterodimeric receptor-ligand complex by X-ray and neutron scattering and computational modeling. *Biochemistry,* **37,** 8208-8217.

Perkins, S. J., Ashton, A. W., Boehm, M. K., & Chamberlain, D. (1998). Molecular structures from low angle X-ray and neutron scattering studies. *Int. J. Biol. Macromol.* **22**, 1-16.

Perkins, S. J., Ullman, C. G., Brissett, N. C., Chamberlain, D. C. & Boehm, M. K. (1998). Analogy and solution scattering modelling: new structural strategies for the multidomain proteins of complement, cartilage and the immunoglobulin superfamily. *Immunol. Rev.* **163,** 237-250.

Boehm, M. K., Woof, J. M., Kerr, M. A. & Perkins, S. J. (1999). The Fab and Fc fragments of IgA1 exhibit a different arrangement from that in IgG: a study by X-ray and neutron solution scattering and homology modelling. *Submitted for publication.*

Boehm M. K., Corper, A. L., Wan, T., Sohi, M., Sutton, B. J., Thornton, J. D., Keep, P. A., Chester, K. A., Begent, R. H. J. & Perkins, S. J. (1999). Crystal structure of the anti-carcinoembryonic antigen single-chain Fv antibody MFE-23 by X-ray crystallography: the lattice contacts provide a model for interactions with carcinoembryonic antigen. *Submitted for publication.*

Boehm, M. K., Chester, K. A., Begent, R. H. J. & Perkins, S. J. (1999). Structural model for the complex between an single-chain Fv MFE-23 and carcinoembryonic antigen by neutron and X-ray scattering and homology modelling: implications for tumour targeting. *Submitted for publication.*

Abstracts and Poster Presentations

Mark K. Boehm, M. Olga Mayans, Jeremy D. Thornton, Richard H. J. Begent, Pat A. Keep and Stephen J. Perkins. (1995). Molecular modelling of the seven domain structure of human carcinoembryonic antigen by X-ray and neutron scattering. Poster presented at Neutron Scattering 1995, Hulme Hall, Manchester University 3-4 April 1995. **Awarded the prize for the best poster in the biological sciences section.**

Mark K. Boehm, M. Olga Mayans, Jeremy D. Thornton, Richard H. J. Begent, Pat A. Keep and Stephen J. Perkins. (1995). Automated modelling of the multidomain structure of carcinoembryonic antigen. Poster and abstract presented at the CCP13/NCD Workshop, Daresbury Laboratory, Warrington 9-11 May 1995.

Mark K. Boehm, Adam L. Corper, Tommy Wan, Maninda Sohi, Brian J. Sutton, Jeremy D. Thornton, Pat A. Keep, Richard H. J. Begent and Stephen J. Perkins. (1996). Crystal structure of MFE-23, a clinically important anti-CEA scFv antibody fragment, at 0.28 nm. Poster and abstract presented at the International Union of Crystallography XVII, Seattle, U.S.A., 8-17 August 1996.

Mark K. Boehm, Jeremy D. Thornton, Pat A. Keep, Kerry A. Chester, Richard H. J. Begent and Stephen J. Perkins. (1997). The solution structure of MFE-23, a clinically-important antibody fragment, by a comparison of neutron scattering data with its crystal structure. Poster and abstract presented at the Institute of Physics Condensed Matter and Materials Physics Conference, University of Exeter, 17-19 December 1997.

Mark K. Boehm, Jenny M. Woof, Michael A. Kerr and Stephen J. Perkins. (1997). Solution structure of human immunoglobulin A1 by X-ray and neutron scattering. Poster and abstract presented at the Institute of Physics Condensed Matter and Materials Physics Conference, University of Exeter, 17-19 December 1997.

Reprinted from *J. Mol. Biol.* (1996) **259**, 718–736

JMB

Extended Glycoprotein Structure of the Seven Domains in Human Carcinoembryonic Antigen by X-ray and Neutron Solution Scattering and an Automated Curve Fitting Procedure: Implications for Cellular Adhesion

Mark K. Boehm, M. Olga Mayans, Jeremy D. Thornton Richard H. J. Begent, Pat A. Keep and Stephen J. Perkins

0022–2836/96/000000–00 $18.00/0

J. Mol. Biol. (1996) 259, 718–736

JMB

Extended Glycoprotein Structure of the Seven Domains in Human Carcinoembryonic Antigen by X-ray and Neutron Solution Scattering and an Automated Curve Fitting Procedure: Implications for Cellular Adhesion

Mark K. Boehm[1,2], M. Olga Mayans[1], Jeremy D. Thornton[2] Richard H. J. Begent[2], Pat A. Keep[2]* and Stephen J. Perkins[1]*

[1]*Department of Biochemistry and Molecular Biology and* [2]*Department of Clinical Oncology, Royal Free Hospital School of Medicine Rowland Hill Street, London NW3 2PF, UK*

**Corresponding authors*

Carcinoembryonic antigen (CEA) is one of the most widely used cell-surface tumour markers for tumour monitoring and for targeting by antibodies. It is heavily glycosylated (50% carbohydrate) and a monomer is constructed from one V-type and six C2-type fold domains of the immunoglobulin superfamily. The solution arrangement at low resolution of the seven domains in CEA cleaved from its membrane anchor was determined by X-ray and neutron scattering. Guinier analyses showed that the X-ray radius of gyration R_G of CEA was 8.0 nm. The length of CEA was 27 to 33 nm, and is consistent with an extended arrangement of seven domains. The X-ray cross-sectional radius of gyration R_{XS} was 2.1 nm, and is consistent with extended carbohydrate structures in CEA. The neutron data gave CEA a relative molecular mass of 150,000, in agreement with a value of 152,500 from composition data, and validated the X-ray analyses. The CEA scattering curves were analysed using an automated computer modelling procedure based on the crystal structure of CD2. The V-type and C2-type domains in CD2 were separated, and the C2-type domain was duplicated five times to create a linear seven-domain starting model for CEA. A total of 28 complex-type oligosaccharide chains in extended conformations were added to this model. By fixing the six interdomain orientations to be the same, three-parameter searches of the rotational orientations between the seven domains gave 4056 possible CEA models. The best curve fits from these corresponded to a family of zig-zag models. The long axis of each domain was set at $160(\pm 25)°$ relative to its neighbour, and the two perpendicular axes were orientated at $10(\pm 30)°$ and $-5(\pm 35)°$. Interestingly, the curve fit from this model is within error of that calculated from a CEA model generated directly from the CD2 crystal structure by the superposition of adjacent domains. Zig-zag models of this type imply that the protein face of the GFCC' β-sheet in neighbouring CEA domains lie on alternate sides of the CEA structure. Such a model has implications for the adhesion interactions between CEA molecules on adjacent cells or for the antibody targeting of CEA.

Keywords: carcinoembryonic antigen; glycoprotein; X-ray scattering; neutron scattering; molecular modelling

Abbreviations used: CEA, carcinoembryonic antigen; Ig, immunoglobulin; CPU, central processing unit; PSG, pregnancy-specific glycoprotein.

0022–2836/96/240718–19 $18.00/0

Introduction

Targeted cancer therapy depends on efficient ligand binding to a target on tumour cells. Both rapid on-rate and slow off-rate are required as components of such ligands. Antibodies through their great potential for diversity provide ligands for targeting to most common types of cancer. Manipulation of antibody structures to provide slower off-rates should facilitate tumour retention of antibody, and will offer a powerful means of cancer therapy. Carcinoembryonic antigen (CEA), which is typically expressed on colon carcinoma cells, also appears in other tumours of epithelial origin (e.g. breast, lung, pancreas; Thompson & Zimmerman, 1988; Thompson *et al.*, 1991). CEA is a cell-surface protein and as such can be exploited as a target protein for anti-tumour antibodies linked either to radiolabels or to cytotoxic substances. Clinical applications of the existence of CEA on tumour cells are complicated by two factors, the lack of tumour specificity, and the existence of a family of related antigens including non-specific cross-reacting antigen and biliary glycoprotein (Thompson *et al.*, 1991). In order to improve the reliability of CEA as a tumour marker, the different epitopes in CEA require identification, together with those in non-specific cross-reacting antigen and biliary glycoprotein in order to suggest improvements to existing anti-CEA antibodies by the use of molecular biological methods.

CEA is a heavily glycosylated protein with a commonly assumed M_r of 180,000, and over 50% is carbohydrate. The CEA sequence contains 668 amino acid residues in one variable-type and six constant-type immunoglobulin fold domains (V-type and C2-type Ig: Williams & Barclay, 1988). There are three similar repeats of Ig fold pairs of 178 residues, each denoted I, II and III in Figure 1, together with a hydrophobic C-terminal peptide of 26 residues (Oikawa *et al.*, 1987; Paxton *et al.*, 1987). The C-terminal peptide is post-translationally replaced by a glycosyl phosphatidyl-inositol anchor for attachment to membranes (Jean *et al.*, 1988; Hefta *et al.*, 1988). The locations of 28 putative carbohydrate sites in CEA are also known, and the large number of these sites may influence considerably the exposure of antigenic surfaces.

No atomic structure for CEA is known at present. For reason of its glycosylation, it is most unlikely that the extracellular domains could be crystallised intact, and alternative strategies are required for a structure determination. Electron microscopy has suggested that CEA is rodshaped (Slayter & Coligan, 1975), although this may be limited by the non-physiological conditions of measurement. An extended CEA model was predicted on the basis of homology with crystal structures for the Bence-Jones dimer REI and the cell-surface proteins CD4 and CD2 (Bates *et al.*, 1992). Such a model, however, requires experimental validation. X-ray and neutron scattering will provide structural information to a resolution of approximately 3 nm, and is well suited for such a study of the CEA domains. No special preparation of CEA is required, apart from its solubilisation from membranes, and data are obtained in near-physiological conditions. The joint use of X-ray scattering in H_2O buffers and neutron scattering in ^2H_2O buffers enables the contributions of protein and carbohydrate to the scattering curve to be analysed. In X-ray scattering, the CEA carbohydrate makes a greater contribution to the intensity of the observed curve than the CEA protein. The scattering data will show whether CEA contains an extended or compact arrangement of domains, and whether it is monomeric or dimeric.

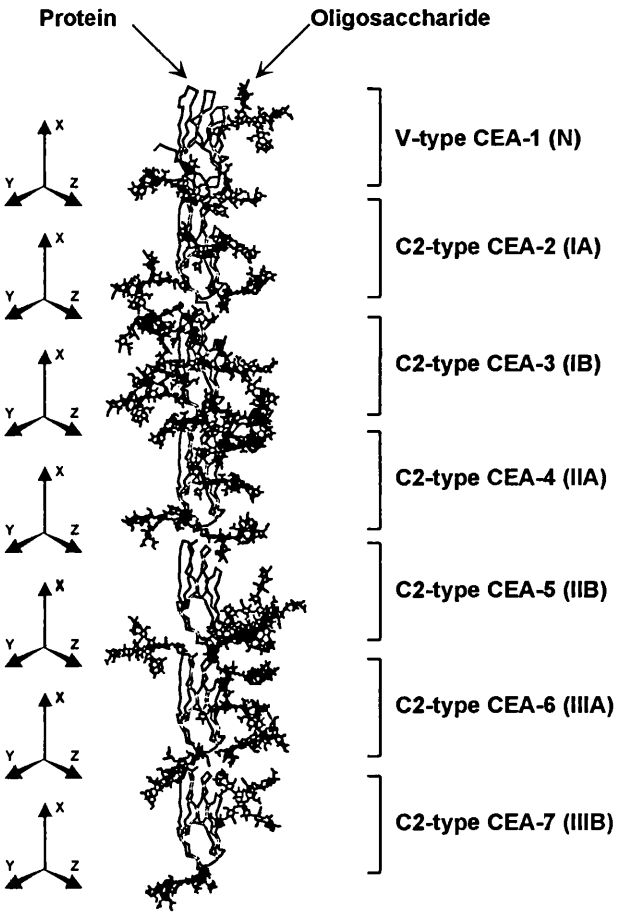


Figure 1. A linear model of CEA derived from the crystal structure of human CD2. The CEA domains are numbered 1 to 7, and as N to IIIB to follow Oikawa *et al.* (1987). Seven separate CD2 C2-type domains (shown as α-carbon traces) were orientated such that the N terminus and C terminus of each C2-type domain (α-carbon atoms of Glu104 and Pro180 in 1hnf) were aligned on the same *X*-axis, with a 0.42 nm spacing between adjacent C2-type domains. A variable (V)-type CD2 domain was superimposed on the N-terminal C2-type domain. *X*, *Y* and *Z*-axes were assigned to each domain at the α-carbon atom of Pro180 as origin so that each domain could be rotated independently of each other. Twenty-eight oligosaccharide chains were added in full at the putative N-linked glycosylation sites (see the text).

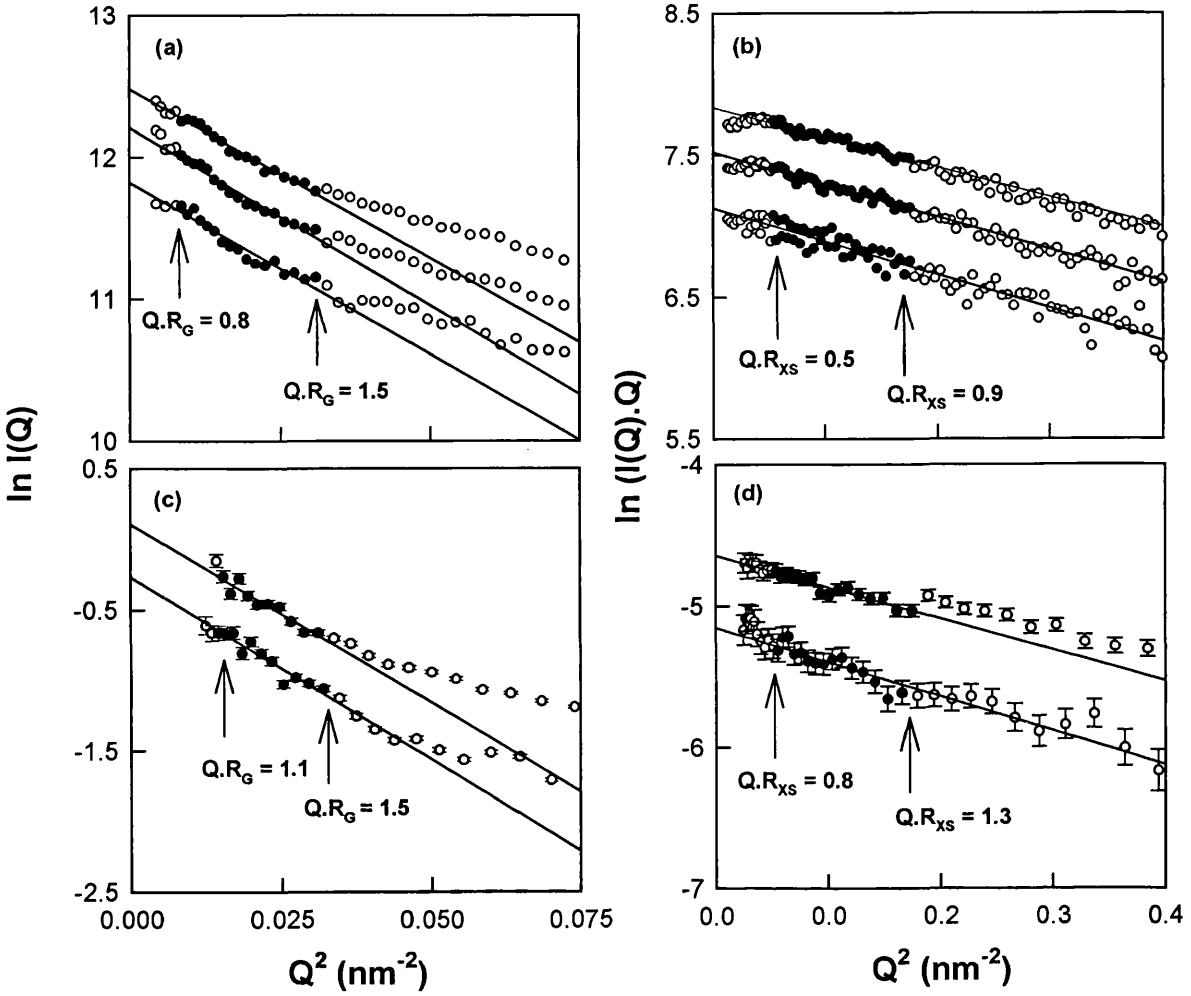


Figure 2. X-ray and neutron Guinier R_G and R_{XS} plots for CEA. (a) and (b) The X-ray plots for CEA concentrations of 3.3, 2.5 and 1.6 mg/ml. The filled circles between the QR_G and QR_{XS} ranges as arrowed show the data points used to determine R_G and R_{XS} values. The R_G values were extracted using a Q range of 0.09 to 0.18 nm^{-1}, and the R_{XS} values likewise from a Q range of 0.23 to 0.42 nm^{-1}. Data outside these ranges are denoted by open circles. (c) and (d) The neutron plots for CEA concentrations of 7.3 and 3.6 mg/ml, with other details as for (a) and (b) except that the Q range used for the R_G values was 0.12 to 0.18 nm^{-1}. Error bars correspond to the statistical errors of LOQ neutron data collection.

Scattering analyses have been enhanced by a recently developed automated method of constrained curve modelling that permits a systematic evaluation of possible models for CEA, and places limits on the precision of such models (Perkins *et al.*, 1991; Mayans *et al.*, 1995; Beavil *et al.*, 1995). We have shown that scattering curves are fully calculable from atomic coordinates (Smith *et al.*, 1990; Perkins *et al.*, 1993). The CEA domains can be represented using known atomic structures from homologous proteins from the Ig superfamily. Starting from recent crystal structures for related cell-surface proteins such as CD2 and CD4 (Wang *et al.*, 1990; Ryu *et al.*, 1990; Jones *et al.*, 1992; Brady *et al.*, 1993; Bodian *et al.*, 1994) and other known constraints, it is possible to derive and analyse four families of CEA structures to determine which is the most compatible with the experimental scattering curves. This constrained conformational search was used to propose an extended low-resolution model for the glycoprotein structure of CEA and the precision of such a structure. The resulting model permitted an assessment of the steric accessibility of the Ig fold domains in CEA in terms of their physiological role in cellular adhesion events (Benchimol *et al.*, 1989; Zhou *et al.*, 1993), and the binding of antibodies and scFv antibody fragments to CEA (Hammarström *et al.*, 1989; Schwarz *et al.*, 1988; Ikeda *et al.*, 1992; Nap *et al.*, 1992; Murakami *et al.*, 1995).

Results and Discussion

Synchrotron X-ray scattering measurements on CEA

CEA was solubilised to cleave it from its membrane anchor by treatment with perchloric acid (see Materials and Methods), and pretreated

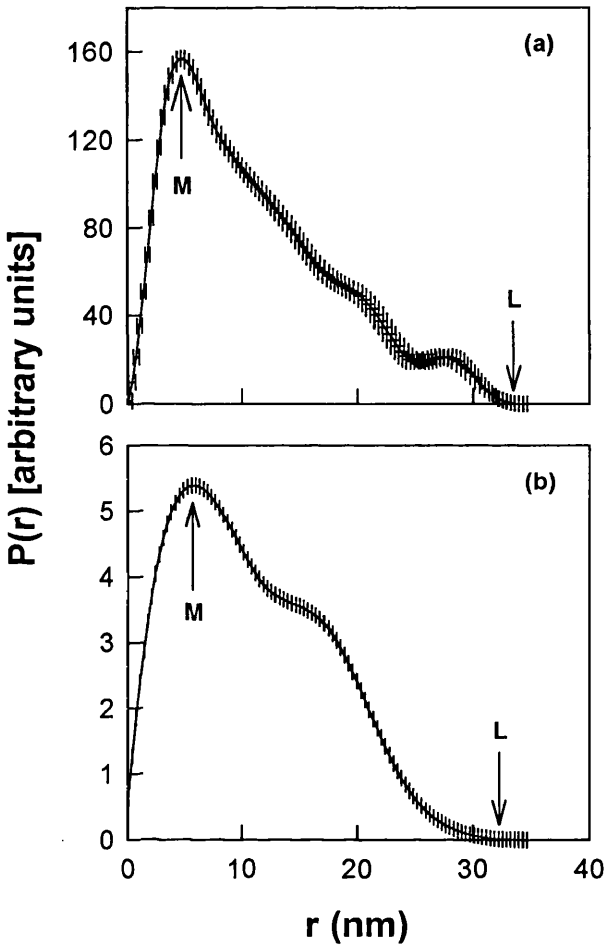


Figure 3. X-ray and neutron distance distribution functions $P(r)$ for CEA. (a) The X-ray $P(r)$ curve at a CEA concentration of 3.3 mg/ml as a continuous line with error bars as calculated using ITP-91, and a maximum M at 4.9 nm. (b) The neutron $P(r)$ curve at a CEA concentration of 3.6 mg/ml, with M at 6.0 nm. The maximum dimension of CEA was determined to be 28 to 32 nm from (a) and (b).

by gel filtration to remove aggregates. Using synchrotron X-ray scattering data, linear Guinier plots could be obtained for ten different CEA samples in the concentration range 1.6 to 7.0 mg/ml (Figure 2(a)). Occasionally, steeply curved Guinier regions at low Q were observed in place of the linear plot seen in Figure 2(a), and these data were rejected as these correspond to CEA aggregates. Time-frame analysis of the Guinier region showed that CEA was resistant to the radiation damage effects that often occur with synchrotron X-ray beams. Guinier analyses gave a consistent R_G value of 8.0($\pm$0.6) nm in a QR_G range of 0.8 to 1.5. Since CEA is expected to be rod-like in structure, the R_G values will underestimate the true value for CEA as the available Q range for measurement is not sufficiently low. R_G values cited below are the apparent values from Guinier fits except when specified otherwise. Calculation of the elongation or anisotropy ratio of CEA from the ratio R_G/R_O (where R_O is the R_G of the sphere with the same hydrated volume of 241.1 nm^3 as CEA) gave a minimum value of 2.7($\pm$0.2). Since R_G/R_O values for typical globular proteins are close to 1.28 (Perkins, 1988), it is concluded that CEA possesses a highly elongated structure in solution.

CEA gave satisfactory linear cross-sectional R_{XS} analyses (Figure 2(b)) in an acceptable QR_{XS} range of 0.5 to 0.9 and beyond to higher Q. The ten CEA curves gave a consistent R_{XS} value of 2.1($\pm$0.2) nm. The combination of the R_G and R_{XS} analyses (see Materials and Methods) gave a length L of 27($\pm$2) nm for CEA. The combination of the $I(0)$ and $[I(Q)Q]_{Q\to 0}$ analyses gave a similar length L of 31($\pm$4)nm. If the hydrated volume of CEA is 241.1 nm^3, combination of the L and R_{XS} values showed that, to a first approximation, CEA can be represented by a cylinder of dimensions $L \times 2A \times 2B$ of 29 nm $\times$ 8 nm $\times$ 1 nm. This showed that CEA has an elongated cross-section. While the cross-sectional dimensions of 8 nm $\times$ 1 nm are not physically realistic, they can be attributed to the consequence of highly extended carbohydrate chains on the protein surface.

The $I(Q)$ curves in reciprocal space for CEA were transformed into distance distribution functions $P(r)$ in real space (Figure 3(a)). The R_G and $I(0)$ values calculated from $P(r)$ analysis were within 6% and 2%, respectively, of the Guinier values (Figure 2(a)), which indicates that these analyses are self-consistent. The point at which the $P(r)$ curve intersects the zero axis at large inter-vector distances R gave the maximum dimension L of CEA as 33 nm. This is comparable with the values calculated from the two Guinier analyses above, although the precision of this determination is not high for reason of the low intensity of $P(r)$ at large r. The maximum M in $P(r)$ gives the most frequently occurring distance within CEA and this was determined as r = 4.9 nm.

Pulsed neutron scattering measurements on CEA

Two CEA samples were studied by neutron scattering as a control for X-ray-induced radiation damage and contrast-dependent properties. In common with other glycoproteins, CEA was prone to aggregation in ^2H_2O solvents. All samples yielding curved Guinier plots were not considered further. The neutron Guinier R_G analysis of Figure 2(c) showed that CEA in ^2H_2O buffers resulted in an R_G value of 8.8($\pm$0.5) nm. This is within error of the X-ray data, although the precision of measurement on LOQ is less for reason of a reduced QR_G range of fit. The neutron Guinier $I(0)/c$ values are standardised relative to a standard deuterated polymer, and the M_r of CEA can be determined by comparison with other proteins in ^2H_2O buffers measured on LOQ. The mean $I(0)/c$ value for CEA was determined to be 0.183($\pm$0.043). Simulations show that the systematic error of this determination is maximally 5%. The $I(0)/c$ value for

bovine IgG1 and IgG2 were determined to be 0.180 and 0.182(±0.006; Mayans *et al.*, 1995). As IgG1 and IgG2 have values of 144,000, the M_r for CEA was determined as 150,000(±35,000). This agrees well with the M_r of 152,500 calculated from the CEA composition (see Materials and Methods). The M_r calculation validated the CEA scattering data and showed that CEA as prepared is monomeric, thus ruling out dimer models for CEA.

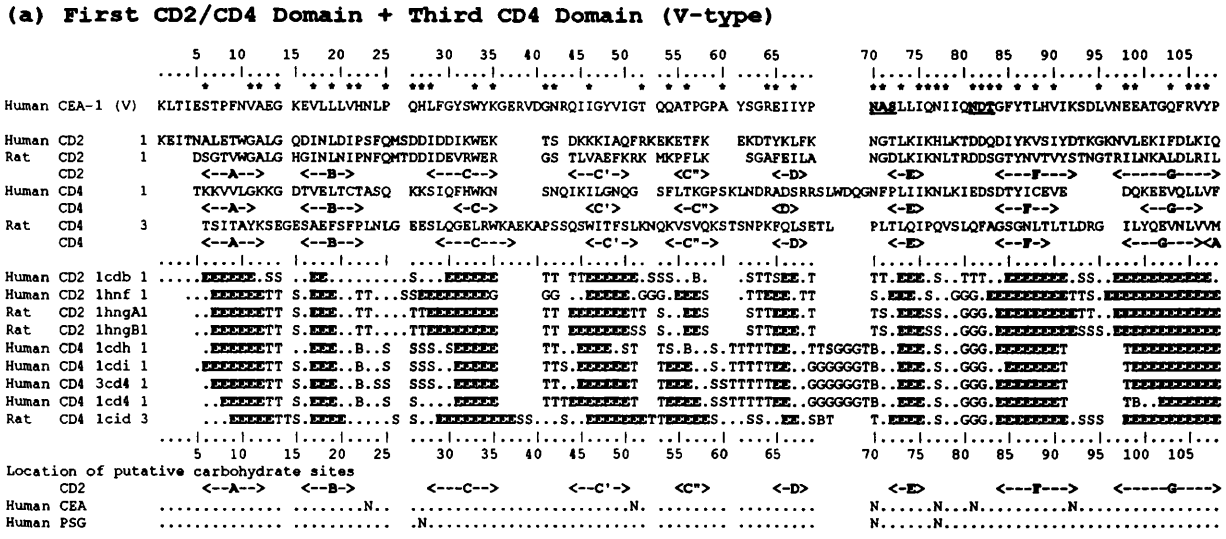
While less precise than the X-ray data, the neutron analyses validated the results from X-ray scattering. The neutron Guinier cross-sectional R_{XS} analysis of Figure 2(d) resulted in an R_{XS} value of 2.3(±0.3) nm for CEA in ^2H_2O buffers. The neutron R_G and R_{XS} analyses gave a length L of 37(±2) nm for CEA, while the $I(0)$ and $[I(Q)Q]_{Q\to 0}$ analyses gave L as 29(±1) nm. The dimensions $L \times 2A \times 2B$ of CEA were determined to be 33 nm × 9 nm × 1 nm from a dry volume of 178.9 nm^3. Interestingly, the R_{XS} values did not exhibit a contrast dependence as reported elsewhere for glycoproteins (Perkins *et al.*, 1990). Such a dependence was in fact found at larger Q beyond the R_{XS} region in the X-ray and neutron scattering curves of CEA (see below). The neutron distance distribution function $P(r)$ for CEA in Figure 3(b) gave R_G and $I(0)$ values within 5% and 1%, respectively, of those found in the Guinier analyses, and showed that the neutron $P(r)$ and Guinier analyses were self-consistent. The maximum M in $P(r)$ was 6.0 nm, and the maximum dimension L was 29 nm. All these analyses gave values similar to the X-ray values.

Initial molecular graphics model for CEA

As the CEA domains belong to the Ig superfamily (Williams & Barclay, 1988; Killeen *et al.*, 1988; Bates *et al.*, 1992), the experimental X-ray and neutron curves of CEA were modelled using atomic coordinates for one V-type and six C2-type Ig fold domains. The most relevant crystal structures were those for the related cell-surface proteins CD2 and CD4. These were examined further by means of manual and MULTAL automated sequence alignments (Taylor, 1988). Of the 108 residues in the CEA V-type sequence, 42% showed >60% residue type conservation with the CD2 and CD4 V-type sequences. If the C2-type sequences of CEA were compared one-by-one with the CD2 and CD4 sequences, 31 to 34% of residues showed >60% conservation. Higher similarities were found with CD2 than with CD4, and this analysis confirms and extends the earlier observation that CEA belongs to a subset of the Ig superfamily (Killeen *et al.*, 1988). The DSSP secondary structure analysis program (Kabsch & Sander, 1983) was used to locate the β-strands in the crystal structures of CD2 and CD4 in order to position these in the CEA domains. These β-strands were consistently located in all eight or nine structures (Figure 4), and their positions could be assigned in CEA from the sequence alignment. The DSSP analysis shows that only CD2 has the linker sequences ERVS and EMVS between the V- and C2-type domains, which are analogous in length to the six links between the seven CEA domains. For these two reasons, the human CD2 crystal structure was selected for modelling the protein core of CEA.

Two different protein models for CEA were created. The "CD2-derived" model was based on the interdomain orientation of CD2 throughout CEA. The β-strands C, F and G (Figure 4) are structurally equivalent in the V- and C2-type domains (Jones *et al.*, 1992). The α-carbon atoms located within β-strand C (Glu131, Leu132 and Leu134), β-strand F (Lys159, Lys161 and Thr163) and β-strand G (Val174, Glu175 and Pro176) of the CD2 C2-type domain were readily superimposed upon those in the V-type domain (Asp32, Ile33 and Trp35; Ile80, Lys82 and Ser84; Ile97, Phe98 and Asp99). After linking 12 V-type and C2-type domains in six CD2 structures, the five surplus V-type domains were deleted to leave a seven-domain CEA model. The C-terminal α-carbon atom of Pro180 and the N-terminal α-carbon atom of Glu104 between two adjacent C2-type domains were found to be separated by 0.42 nm. The "linear" CEA model was based on treating the seven domains in the CD2-derived model as independent objects, in which the first and last α-carbon atoms of Glu104 and Pro180 defined the X-axis (Figure 1; see Materials and Methods). These α-carbon atoms were separated by 0.42 nm in neighbouring domains. The Y-plane was defined by the α-carbon atom of Phe160. The origin was set at the C-terminal residue Pro180 (Figure 1; see Materials and Methods). All seven domains were aligned in the same orientation, and CEA models for curve fitting were generated by rotations of each domain about its axes.

Using the known carbohydrate composition of CEA and the similar sizes of each oligosaccharide as constraints, two types of oligosaccharide conformations were constructed. Examination of over 50 glycoprotein coordinate files in the Brookhaven database showed that the three largest oligosaccharide structures were those of Fc Kol, human leucocyte elastase and glucoamylase (Deisenhofer, 1981; Bode *et al.*, 1989; Aleshin *et al.*, 1992, 1994). The oligosaccharides found in the database showed that the first three residues GlcNAc-GlcNAc-Man have a common near-linear conformation by virtue of β1 → 4 links between them. Despite the further addition of carbohydrate residues, the resulting oligosaccharide structures have similar spatial dimensions. Each oligosaccharide was represented by a triantennary complex-type structure $Man_3GlcNAc_6Gal_3Fuc_3NeuNAc_1$, whose structure was based on that found in Fc Kol (Figure 5; see Materials and Methods). Each one was added at known putative sites in the CEA model according to the alignment of Figure 4. However, while the carbohydrate chains in human leucocyte elastase and CD2 have conformations extended from the protein surface (Bode *et al.*, 1989; Wyss *et al.*, 1995), glucoamylase is observed to



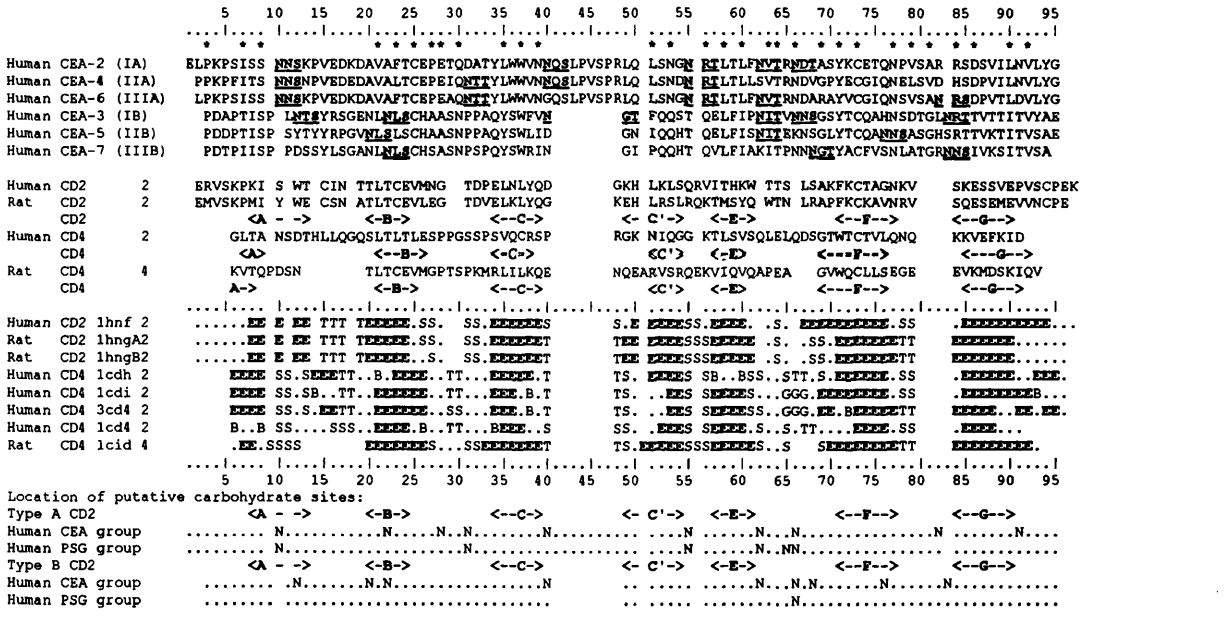


Figure 4. Sequence alignment of CEA with CD2 and CD4 and their known structures. (a) The V-type domain of CEA (CEA-1) is compared with V-type domains in CD2 and CD4. (b) The C2-type domains of CEA (CEA-2 to CEA-7) are compared with C2-type domains in CD2 and CD4. In (a) and (b), residues are marked with an asterisk if residue conservation is at least 60%, with conserved groups of residues as defined in parentheses: (G, A, S), (I, L, M, V), (F, H, W, Y), (D, E), (H, K, R), (S, T), (N, Q). Putative N-linked glycosylation sites are in bold and underlined. Sequences are labelled with the appropriate Brookhaven Protein Database code. The secondary structure elements identified by the DSSP program are labelled as follows: E, β-strand (bold); B, single residue β-ladders; T, turns; S, bends; G, 3_{10}-helix. In (b), the interdomain link residues in human and rat CD2 are ERVS and EMVS, respectively. At the bottom of both (a) and (b), the location of at least one putative Asn glycosylation site in sequence alignments of subgroups of the CEA gene family is denoted by N.

possess compact oligosaccharide structures positioned against the protein surface (Aleshin *et al.*, 1992, 1994). The two conformational types were therefore represented in CEA models by positioning all 28 chains either extended away from or in close proximity to the protein surface. This modelling procedure permits a realistic evaluation of the two possible extremes of carbohydrate structures in CEA.

Single-density X-ray scattering curve modelling for CEA

The objective of curve modelling was to show which family of domain structures best represented the solution structure of CEA. The automated curve-fitting procedure AUTOSCT (Beavil *et al.*, 1995) was applied to a range of rotamers calculated from the linear CEA model. This is based on the

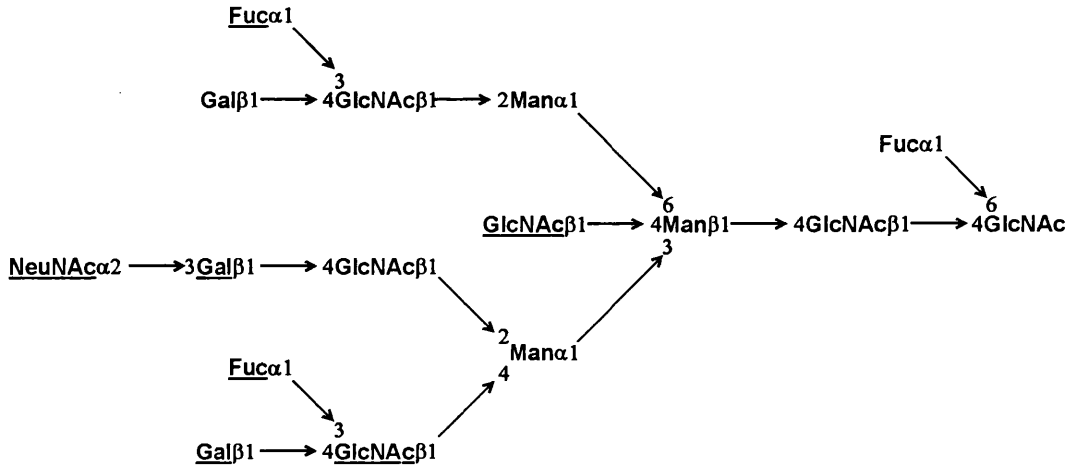


Figure 5. The averaged structure for a single oligosaccharide site on CEA. This was determined to be $Man_3GlcNAc_6Gal_3Fuc_3NeuNAc_1$ from the carbohydrate analysis of CEA (Yamashita *et al.*, 1987, 1989). A model of this was constructed using the nine-residue carbohydrate coordinates in the crystal structure of the Fc fragment of immunoglobulin G (Deisenhofer, 1981). Seven further residues (underlined) were added to the Fc coordinates to produce the complete oligosaccharide model for CEA.

procedure previously calibrated with known single-domain crystal structures (Smith *et al.*, 1990; Perkins *et al.*, 1993). A full three-axis rotational search in 30° steps over 360° for the six interdomain links in CEA would involve 18 parameters and $(12 \times 12 \times 12)^6$ steps = 3×10^{19} models. As 2.5 minutes were required for each curve fit, the conformational search was only feasible for 1×10^3 to 1×10^4 models. The similarity of both the link regions and domain types in CEA (Figure 4) suggested that a simplified three-parameter approach could be usefully applied to CEA in which the same *X*, *Y* and *Z*-axis rotations were applied to all six interdomain interfaces. This procedure excludes consideration of CEA models in which the six sets of rotations differ between the domains. It is adequate to cover a suitable survey of the major families of CEA structures while remaining within the precision of the technique.

Contour plots showed that a large proportion of possible CEA structural families could be rejected. Figure 6 illustrates how the calculated R_G and $R_{2.0}$

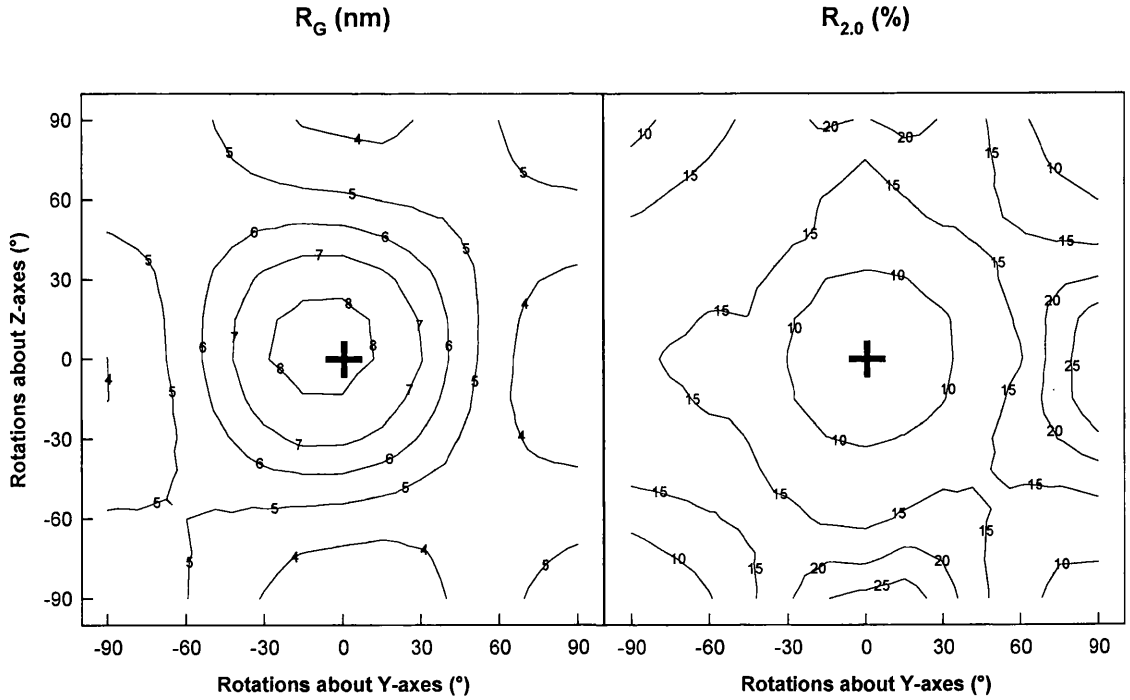


Figure 6. Contour maps of the dependence of the R_G and *R*-factor on domain rotations in the starting linear model of CEA. The *X*-axis rotation is set as 0°. The map was generated by domain rotations in 15° steps from –90° to 90° about the *Y* and *Z*-axes. The starting linear model is obtained with $Y = Z = 0°$ and is marked by a plus symbol.

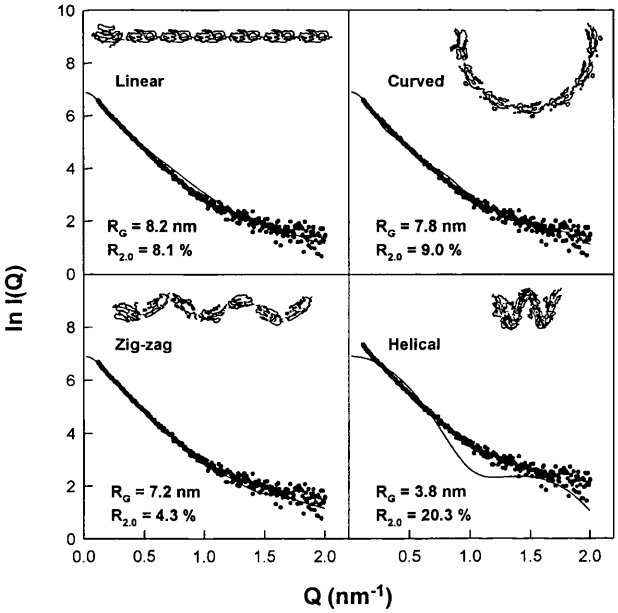


Figure 7. Comparison of the calculated scattering curves for four families of CEA structures with the X-ray scattering curves. Linear, curved, zig-zag and helical CEA models are depicted without the oligosaccharide chains in order to clarify the protein cores used to define each structural family. The *R*-factors were calculated using the X-ray experimental data out to $Q = 2.0$ nm^{-1}.

values depend on the *Y*-axis and *Z*-axis rotations between −90° and 90° in 15° increments when the *X*-axis rotation is fixed at 0°. The R_G values ranged from 8.2 nm for the starting linear model ($Y = 0°$, $Z = 0°$) to 3.1 nm for the most compact structure ($Y = 90°$, $Z = 0°$). The $R_{2.0}$ values ranged from 5.3% for the best curve fits ($Y = -90°$, $Z = -90°$) to 29.8% for the worst fit ($Y = 90°$, $Z = 0°$). The two contour plots showed that models with R_G values close to the experimental range of 8.0(±0.6) nm produced low $R_{2.0}$ values. Models with suitable $R_{2.0}$ values at the four corners of the contour map could be ruled out as these were incompatible with the R_G data.

The full search of CEA domain rotations between 0° and 345° in 15° steps about the *X*-axes, and between −90° and 90° in 15° steps about the *Y* and *Z*-axes generated 24 × 13 × 13 = 4056 single-density models with extended carbohydrate chains. This calculation required seven days of R4600PC central processing unit (CPU) time. Histograms showed that the $R_{2.0}$ values ranged from 4.3% to 30.4% (mostly less than 9%), the R_G values ranged from 3.09 nm to 8.34 nm, and the R_{XS} values ranged from 0.13 nm to 4.14 nm (mostly between 1.7 and 2.3 nm). Four families of structures could be distinguished, as defined by the structure of the protein core (Figure 7). None of the models gave consistently good curve fits, although zig-zag models gave the most promising fits, as shown by comparisons of the calculated and observed curves in the *Q* range of 0.09 to 1.0 nm^{-1} in Figure 7.

(1) The linear model ($X = Y = Z = 0°$) gave a poor X-ray curve fit, as shown also from the low R_{XS} value of 1.62 nm and high $R_{2.0}$ value of 8.1%, even though the R_G value is reasonable at 8.2 nm. Such a CEA structure was too elongated with too narrow a cross-section.

(2) Curved CEA models resulted from rotations only about the *Y* or *Z*-axes (domain tilting). While the curved model of Figure 7 ($X = Y = 0°$, $Z = 30°$) had reasonable R_G and R_{XS} values of 7.8 nm and

Table 1. Summary of CD2, CD4 and CEA rotational angles that define their models

Model	Rotations (°) *X*-axis	*Y*-axis	*Z*-axis	R_G[a] (nm)	R_{XS} (nm)	$R_{2.0}$ (%)	$s^0_{20,w}$ (S)
Experimental values:				8.0(±0.6)	2.1(±0.2)	=	6.04(±0.22)
CEA scattering models:							
Sterically allowed ranges	0 to 345	−90 to 90	−90 to 90				
Single-density and linear	0	0	0	(8.2)	(1.62)	(8.1)	5.7
Single-density and curved	0	0	30	(7.8)	(1.82)	(9.0)	5.7
	0	0	(−90 to 90)				
Single-density zig-zag	120	30	60	(7.2)	(2.12)	(4.3)	6.1
(Figure 7)	(90 to 270)	(−90 to 90)	(−90 to 90)				
Single-density and helical	30	90	0	(3.8)	(3.54)	(20.3)	8.2
	(0 to 90 and 270 to 345)	(−90 to 90)	(−90 to 90)				
Double-density zig-zag (Figure 8)	165	30	15	8.0	1.99	4.7	5.9
Double-density CD2-derived (Figure 9)	=	=	=	7.9	2.10	4.9	6.1
Crystal structures (PDB code)[b]							
Human CD2 (1hnf)	210	−10	−40	8.0	2.14	5.0	5.9
Rat CD2 (1hngA)	220	10	−50	7.7	2.16	5.0	5.9
Rat CD2 (1hngB)	220	5	−50	7.7	2.22	5.4	6.0
Human CD4[c] (1cdh)	200	−45	−40	7.6	2.20	5.9	5.8
Rat CD4 (1cid)	230	−44	15	7.9	1.99	4.6	6.0

[a] The R_G R_{XS} and $R_{2.0}$ values correspond to two-density models for X-ray curve fits except when single-density models (in parentheses) are specified.

[b] Rotational angles were derived from the following crystal structures to apply to the linear CEA model (see Materials and Methods) to generate these CEA structures.

[c] Other human CD4 structures (1cdi, 3cd4 and 1cd4) give similar rotational angles.

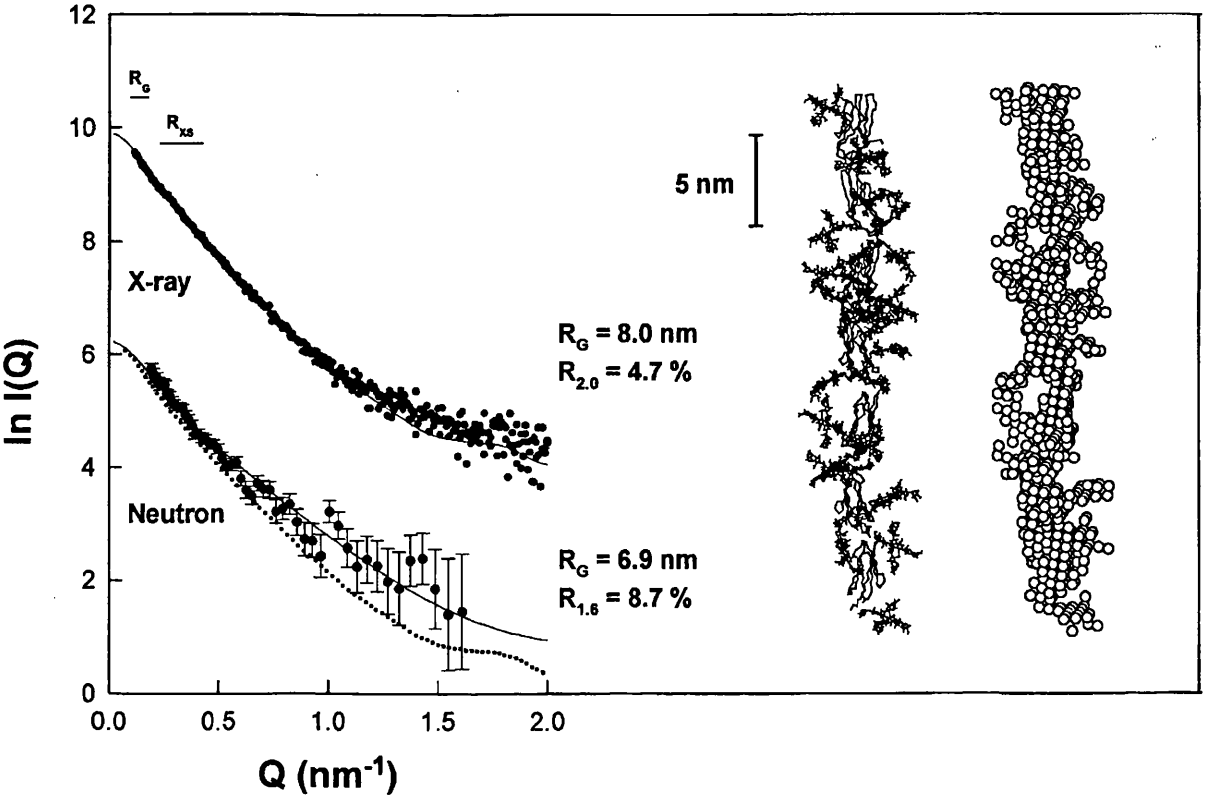


Figure 8. Comparison of the simulated X-ray and neutron scattering curves for the best-fit CEA model with experimental X-ray and neutron data. The seven CEA domains in each model are shown as α-carbon traces, whereas the carbohydrate chains are represented in full. The corresponding sphere model (sphere diameter, 0.572 nm) is shown. A two-density hydrated sphere model with a protein:carbohydrate density ratio of 2:3 was used to calculate the X-ray scattering curve, and a single-density unhydrated model was used for the neutron curve. The calculated neutron curve was corrected for wavelength resolution and beam divergence. For the experimental X-ray curve (●), the R_G value is 8.0 nm and the $R_{2.0}$ value for the X-ray data is 4.7%. For the experimental neutron curve (●), the R_G value is 6.9 nm and the $R_{1.6}$ value is 8.7%. The calculated X-ray curve is shown as a broken line for comparison purposes with the neutron curve. The Q ranges used for the R_G and R_{XS} analyses (Figure 2) are denoted by horizontal bars.

1.82 nm respectively, the $R_{2.0}$ value was high at 9.0%.

(3) The combination of domain tilting with domain twisting about the X-axes produced zig-zag and helical CEA models. The zig-zag models have X-axes rotations between 90° and 270° for all Y-axes and Z-axes rotations (Table 1), and produced elongated structures. That in Figure 7 (X = 120°, Y = 30°, Z = 60°) gave the most reasonable curve fit with a satisfactory R_{XS} of 2.12 nm, and a low $R_{2.0}$ value of 4.3%, although the R_G value of 7.2 nm was low.

(4) Helical models were primarily generated from X-axes rotations in the remaining ranges of 0° to 90° and 270° to 345°, and had more compact structures. The helical model (X = 30°, Y = 90°, Z = 0°) was too compact, as evidenced by the low R_G value of 3.8 nm, the high R_{XS} of 3.54 nm and the high $R_{2.0}$ value of 20.3%.

Control X-ray and neutron scattering curve modelling for CEA

Control calculations were performed to assess the 4056 CEA models further.

(5) Based on previously tested procedures (Smith *et al.*, 1990; Perkins *et al.*, 1993), curve fitting was now applied jointly to the X-ray and neutron scattering curves of CEA. This required consideration of the different scattering densities of the protein and carbohydrate of CEA. Their electron densities were 419 e nm^{-3} and 492 e nm^{-3}, respectively, compared with that of water at 334 e nm^{-3} (Perkins, 1986), and the ratio of electron densities of protein: carbohydrate was 1.00:1.86. The neutron scattering densities were equivalent to 42.3% and 46.0% ^2H_2O, respectively (Perkins, 1986), which gives a ratio of 1.00:0.93. The protein and carbohydrate spheres in the CEA models were assigned weights of 2:3 and 1:1 in 4056 X-ray and 4056 neutron curve fits, respectively. The joint analysis also involved hydrated models in the X-ray modelling, but no corrections for beam effects, while dry models were used in neutron modelling but the curves require corrections for beam wavelength spread and divergence.

To follow Beavil *et al.* (1995), the two-density model curve fits were first sorted in order of the R-factors, then the models were filtered to ensure that at least 460 protein and 460 carbohydrate

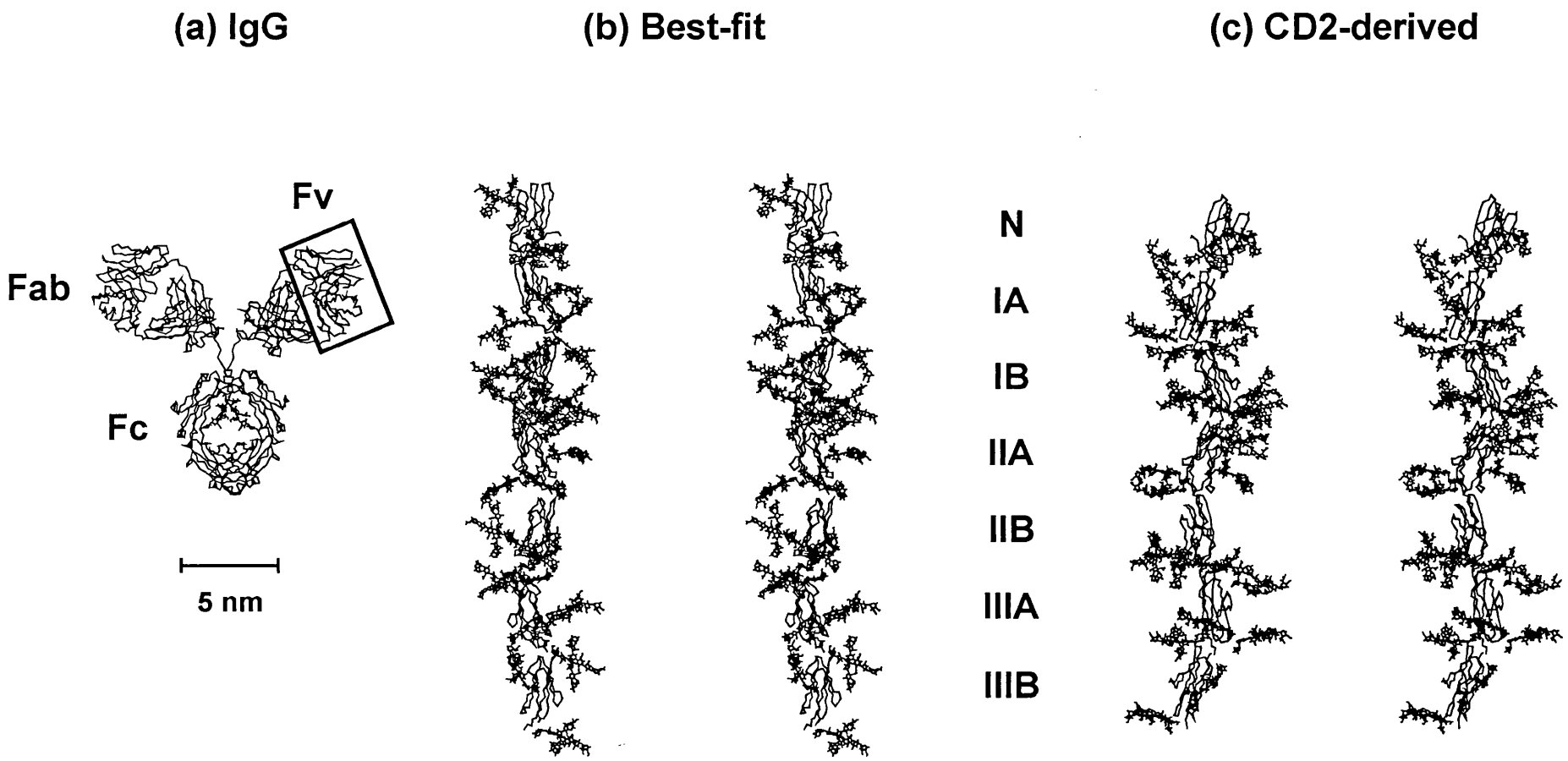


Figure 9. Molecular graphics stereoviews of the final zig-zag and CD2-derived models for CEA. The domains are shown as α-carbon traces, whereas the carbohydrate chains are represented in full. The immunoglobulin G model from Mayans *et al.* (1995) is shown on the same scale for comparison, with its Fv region boxed.

spheres were present in each model (i.e. no steric overlap), the R_G values were between 7.4 nm and 8.6 nm, and the R_{XS} values were between 1.9 nm and 2.3 nm. The filtering procedure resulted in a limited family of models. Comparison of the 100 and 400 best-fit models gave similar outcomes, and this showed that the analysis was stable. The 100 best-fit models corresponded to the zig-zag family with a mean X-axis rotation of 160($\pm$25)°, Y-axis rotation of 10($\pm$30)° and Z-axis rotation of −5($\pm$35)°. Interestingly, the mean R_G of 7.8($\pm$0.2) nm and R_{XS} of 2.02 ($\pm$0.06) nm were now both within the error of experiment (Table 1). Figure 8 shows that a two-density zig-zag CEA model close to the mean with $X = 165°$, $Y = 30°$ and $Z = 15°$ (see Figure 9(b)) gave a good fit to the experimental X-ray data ($R_G = 8.00$ nm, $R_{XS} = 1.99$ nm, $R_{2.0} = 4.7\%$). The neutron model had $R_G = 6.9$ nm, $R_{XS} = 1.73$ nm, $R_{1.6} = 8.7\%$. While the quality of the neutron curve is poorer, and the R_G value is less than that observed, it should be noted that the neutron and X-ray curves are noticeably different in Figure 8. Despite this difference, the two-density zig-zag model is able to offer a good curve fit for the neutron data. Comparison of the single and two-density models in Table 1 shows that the X, Y and Z-axis rotations were within 45° of each other. Allowance for two densities has now selected CEA models for which both the X-ray R_G and R_{XS} values agreed well with observation, unlike the single-density model, which gave too low an R_G value. In other words, the single-density CEA models with the lowest $R_{2.0}$ values satisfied the extended cross-sectional dimensions of CEA by coiling to increase the diameter at the expense of the total length of CEA (smaller R_G). The use of two-density models in which the carbohydrate has an increased scattering density relative to the protein avoided this effect to offer good X-ray R_G and R_{XS} values and low $R_{2.0}$ values.

(6) A two-density CEA model was derived directly from the crystal structure of human CD2 (Figure 9(c); see Materials and Methods) for comparison with the zig-zag model. This had interdomain rotations equivalent to $X = 210°$, $Y = -10°$ and $Z = -40°$ (Table 1), which are within one to two standard deviations of the best-fit two-density zig-zag model, and therefore was structurally similar to this (Figure 9(b)). Table 1 shows that the double-density CD2-derived model and five models derived from human and rat CD2 all gave comparable X-ray R_G, R_{XS} and $R_{2.0}$ values to those of the double-density zig-zag model (curve fits not shown). The rotations for human and rat CD4 are comparable with those for CD2, despite the small differences noted by Jones *et al.* (1992), and Table 1 shows that these CD4-derived models also generated comparable good agreements. The success of CEA models based on the CD2 and CD4 crystal structures is further support for a family of zig-zag structures for CEA (Figure 9(b) and (c)).

(7) The premise that CEA has extended oligosaccharide chains on its surface was tested by performing curve simulations with 4056 single and double-density models in which the oligosaccharide chains were positioned close to the protein surface. The sorting of the curve fits based on $R_{2.0}$ values and filtering based on the observed R_G and R_{XS} parameters showed that it was not possible to match the R_G and R_{XS} values simultaneously. The best X-ray models from these searches exhibited low R_G values of 5.9 to 6.1 nm, satisfactory R_{XS} values of 2.21 to 2.25 nm, and poorer $R_{2.0}$ values of 6.9 to 7.8%. The recalculation of 4056 neutron curve fits confirmed this result. This shows that the CEA carbohydrate structures in general extend into solution, rather than being compact against the protein surface.

Sedimentation velocity of CEA and its hydrodynamic modelling

Sedimentation coefficients $s^\circ_{20,w}$ are a monitor of macromolecular elongation akin to R_G data, and provide a control for the scattering data. In sedimentation velocity runs, the mean $s^\circ_{20,w}$ value from three CEA samples between 0.42 and 0.83 mg/ml was determined to be 6.04($\pm$0.22) S. From this, a high frictional ratio f/f_o of 1.80 was calculated, where f_o is the frictional coefficient of a sphere with the same hydrated volume as that of CEA. The value of $s^\circ_{20,w}$ is comparable with the range of previously published values of 5.47 to 8.04 S (Krupey *et al.*, 1968) and 6.8 S (Slayter & Coligan, 1975).

Procedures for hydrodynamic simulations based on sphere models have been tested by Smith *et al.* (1990) and Perkins *et al.* (1993). The GENDIA program was used to calculate the $s^\circ_{20,w}$ values from the 4056 hydrated X-ray CEA models, each with about 959 spheres. This required 30 days of R3000 CPU time. The 100 best two-density X-ray models gave $s^\circ_{20,w}$ values between 5.82 and 6.26 S. The mean $s^\circ_{20,w}$ value was 6.00($\pm$0.10) S. The two CEA models of Figure 9 gave $s^\circ_{20,w}$ values of 5.9 and 6.1 S (Table 1). As these agreements were within the acceptable precision of $\pm$0.3 S for these calculations (Perkins *et al.*, 1993), the hydrodynamic data and their modelling provide further support for the CEA zig-zag model derived from X-ray curve fits. Table 1 showed that the zig-zag model offered better agreement than the linear, curved and helical models for CEA.

Conclusions

Low-resolution models for CEA by scattering

In its natural state, CEA occurs primarily in the colonic epithelium, with increased expression in colon cancer. It functions as a cell-adhesion molecule (Benchimol *et al.*, 1989; Oikawa *et al.*, 1991; Zhou *et al.*, 1993). Solution scattering has resulted in an improved understanding of its solution structure. Up to now, it was not clear that CEA is

monomeric. The neutron molecular mass agreed well with the molecular mass for a CEA monomer. The success of the scattering curve fits likewise indicated that CEA is monomeric. Gel filtration and SDS-PAGE studies on CEA have generally found that it is monomeric (Slayter & Coligan, 1975), although Lisowska *et al.* (1983) have suggested that CEA may dimerise depending on the conditions of sample preparation. The monomeric state of CEA as studied here implies that the homotypic cell-adhesion interactions between individual CEA molecules on different cells are weak, and that a large number of CEA molecules are required to generate a significant adhesive interaction between cells.

The present X-ray scattering data are more precise than the neutron data. These indicate that CEA is 27 to 33 nm in length and up to 8 nm in width, and contains extended carbohydrate structures. Knowledge of the CEA composition and curve modelling based on the structural homology with CD2 resulted in two rod-like models of length 27 nm and width 8 nm (Figure 9). In agreement with this, electron microscopy has visualised CEA as distinctive rod or cruller-shaped macromolecules with larger dimensions of 9 nm × 40 nm (Slater & Coligan, 1975), although the rotary-shadowing method in use can overestimate macromolecular dimensions (Slayter & Codington, 1973). Rod-shaped structures have also been reported for related members of the Ig superfamily such as the neural cell adhesion molecule (N-CAM) and intercellular adhesion molecule-1 (ICAM-1) by electron microscopy (Becker *et al.*, 1989).

The application of automated curve modelling to the domains of CEA demonstrates the utility of this method to multidomain proteins that cannot in all probability be crystallised. The present curve modelling was constrained on the basis of the protein sequence, the CD2 crystal structure and the distance between its two domains, the carbohydrate composition, and the carbohydrate conformations seen in known crystal structures. The curve fits were implemented on the basis of tested scattering densities and procedures for curve modelling. In combination with the experimental X-ray curve, these constraints place limits on molecular structures for CEA. For CEA, the definition of a full range of interdomain rotations enables a basic set of 4056 models to be tested against experiment. The advantages of this procedure are that a full range of rotational structures is tested automatically, the assumptions in CEA curve modelling could be tested systematically, and the statistical precision of the curve fitting can be defined. In this way, four major conformational families from 4056 CEA models could be identified, of which only one resulted in good curve fits. The repetitive evaluation of 4056 models to test the effect of one or two-density modelling, the use of the CD2 crystal structure and hydrodynamic sedimentation coefficients consistently indicated that the zig-zag family of CEA structures offered the most satisfactory account of the CEA scattering curve.

In relation to the final low-resolution structure of CEA, it should be noted that solution scattering shows only CEA structures that are compatible with scattering curves, and do not determine a unique structure. The X-ray curves close to Q of 1 nm^{-1} were sensitive to the relative orientations of the seven domains in CEA. Analyses based on R_G or R_{XS} values alone were restricted to the Q range of 0.09 to 0.18 nm^{-1} and 0.23 to 0.42 nm^{-1}, and were therefore less discriminatory than the R-factor (Figure 7). All three parameters, together with sedimentation coefficients, were nonetheless useful to assess stereochemically correct models. The existence of six interdomain links in CEA implies that as many as 1×10^{19} models should be tested to explore all interdomain orientations, of which a significant small proportion will give good curve fits. While flexible interdomain conformations may exist, as observed in the two molecules in the crystal structure of rat CD2 (Jones *et al.*, 1992), Table 1 shows that the extent of this is relatively limited in the five crystal structures. It is not possible to allow for independent interdomain links in the simulations; however, it was reasonable to presume that all six interdomain links are similar in CEA in order to permit a realistic survey of allowed conformations. In summary, by two approaches, the present modellings were able to demonstrate that the two-domain crystal structures of CD2 (and CD4 as well) offer a good explanation of the overall solution structure of CEA.

Implications of the CEA structure for biological activity

The extended zig-zag structure of CEA in solution (Figure 9) implies that such a structure on cell membranes would extend from the membrane into solution. All seven domains would then be accessible for protein-protein interactions. A CEA N-terminal V-type domain and one of the C2-type domains are both required to be present for homotypic cell-adhesion interactions (Oikawa *et al.*, 1991; Zhou *et al.*, 1993; Hashino *et al.*, 1993). The IIIA-IIIB domain pair of CEA is involved, although interactions with the IA-IB and IIA-IIB domain pairs have not been excluded.

Incorporation of the CD2 and CD4 crystal structures in the CEA model suggests that each C2-type Ig fold is approximately half-rotated about its X-axis relative to its neighbour such that the EBA and GFCC' β-sheets (Figure 4(b)) in the β-sheet sandwich of each C2-type Ig fold will present alternate faces along one side of the long axis of CEA. Note that this steric relationship cannot be deduced at the resolution of solution scattering, and an atomic structure determination of pairs of CEA domains will be required to verify this. Such a half-rotation is seen in the crystal structure of tissue factor and the growth hormone receptor, both of which have two adjacent C2-type Ig folds,

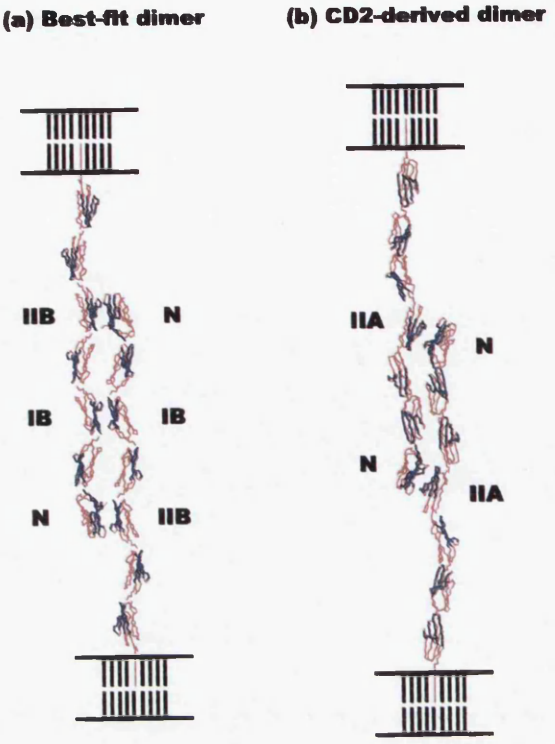


Figure 10. Schematic possible models for the homotypic interaction between two CEA molecules from different cells. The CEA α-carbon trace of its polypeptide chain is depicted in red, except for the GFCC' β-sheet which is shown in blue. No carbohydrate is shown for reasons of clarity. The C-terminal membrane anchor of CEA is shown in red. The best-fit model from Figure 9(b) shows GFCC' β-sheet interactions between the N-IIB pair of domains and also between the IB-pair of domains. The CD2-derived model from Figure 9(c) shows an interaction between the N-IIA pair of domains.

although these folds are closer to fibronectin type III domains (de Vos *et al.*, 1992; Harlos *et al.*, 1994; Campbell & Spitzfaden, 1994). This alternation is attributed to the manner in which the last β-strand G of the first domain runs almost unchanged in direction into the first β-strand A of the second domain, and this packing scheme causes the two Ig folds to become twisted relative to one another (see Campbell & Spitzfaden (1994) for a discussion).

In members of the Ig superfamily, the AGFCC'C'' β-sheet in the V-type domain has been associated with the CD2 and CD4 ligand-binding faces (Wang *et al.*, 1990; Ryu *et al.*, 1990; Jones *et al.*, 1992; Bodian *et al.*, 1994), the GFCC' β-sheet in the C2-type second domain of the α-chain of the IgE FcεRI receptor is the determinant of the IgE-receptor interaction (Beavil *et al.*, 1993), and the AGFCC'C'' face in VCAM-1 is likewise involved in ligand binding (Jones *et al.*, 1995). A special case of the GFCC' β-sheet in the structurally related C2-type second domain of tissue factor has been implicated in interactions with factor VIIa (Harlos *et al.*, 1994). Computer modelling has suggested that the GFCC' β-sheet in the C2-type first domain of the intercellular adhesion molecule-1 (ICAM-1, CD54) is involved in binding to the integrin LFA-1 (Staunton *et al.*, 1990; Berendt *et al.*, 1992). It is thus tempting to propose that the GFC β-sheet forms possible protein ligand sites for the interactions between different CEA molecules from different cells, although mutagenesis experiments will be required to prove this. If this is true, such adhesion sites on the IA-IB, IIA-IIB and IIIA-IIIB domains would be equally presented on two or three sides of the long axis of CEA through rotational twisting to facilitate these interactions.

The location of the 28 putative carbohydrate sites in CEA supports the possible role of the GFCC' β-sheet of the C2-type domains in adhesion. From the sequence alignment of Figure 4, 20 such sites are predicted to lie on exposed loops and only eight are positioned on β-strands. Carbohydrate sites are not found on the β-strands of the V-type domain. They are found at the centre of either strand A or B in all six of the EBA β-sheets in the C2-type domains of human CEA. These EBA β-sheets are unlikely to present protein ligand sites to a CEA V-type domain. In contrast, five of the six GFCC' β-sheets have no carbohydrate site and are potentially available for antiparallel homotypic CEA interactions. The single exception is an oligosaccharide site at the end of strand F in domain IIB. Molecular graphics inspection of the best-fit and CD2-derived models in Figure 9 shows that the six GFCC' faces in human CEA were free of steric hindrance caused by the extended carbohydrate structures in CEA.

The location of putative carbohydrate sites in further sequence analyses of the human CEA gene family, which includes CEA, non-specific cross-reacting antigen, biliary glycoprotein and CEA-group members, shows that the GFCC' β-sheet is likewise accessible (Thompson *et al.*, 1991; Entrez CD-ROM database Release 17.0, June 1995). In 58 V-type sequences, none of three or four oligosaccharide sites occurs in the β-sheet AGFCC'C'', apart from two exceptions at the end of strand F (bottom of Figure 4(a)). This holds for 70 sequences from the related human pregnancy-specific glycoprotein (PSG) gene family. In 67 C2-type sequences, a single type A or type B domain of the human CEA gene family contains between three and six oligosaccharide sites. The 128 sequences generally showed carbohydrate-free GFCC' β-sheets, with the occasional exception of the end of strand F in the type B sequences (bottom of Figure 4(b)). While the human PSG gene family showed reduced glycosylation levels in the C2-type domains, this also conforms to carbohydrate-free GFCC' β-sheets in 130 type A and 79 type B sequences.

Two hypothetical models for an antiparallel homotypic interaction of CEA molecules from different cells are shown in Figure 10. Two CEA structures can be positioned such that the AGFCC'C'' face of a V-type domain is close to a GFCC' face of a C2-type domain. The domain pair in question is dependent on the assumed rotational twist along the long axis of CEA, and this is not known. Figure 10 shows how an interaction can

occur with either the IB or IIB or IIIB domains, or with the IIA or IIIA domains. The models show that the GFCC' face of each C2-type domain is angled at about 45° to the vertical. This presents an upward-angled surface that can match with its complementary V-type domain. Note that the EBA faces in the C2-type domains of CEA are angled downwards to face the cell membrane.

The CEA models can be compared with a model for immunoglobulin G in order to assess binding sites for anti-CEA antibodies (Figure 9). Epitopes to monoclonal antibodies are present on all seven CEA domains (Schwarz *et al.*, 1988; Ikeda *et al.*, 1992; Murakami *et al.*, 1995). Of 52 anti-CEA monoclonal antibodies, 43 were shown to bind to one of five independent non-interacting protein epitopes GOLD 1 to 5, while the rest were directed against carbohydrate epitopes or were inactive (Hammarström *et al.*, 1989). This implies that an exposed protein surface on CEA of area at least 4 nm × 5 nm must be available to antibody. As the size of the antibody Fv fragment is comparable with a single CEA domain face, it is conceivable that the non-glycosylated GFCC' faces on the CEA domains are able to present many of these protein epitopes.

Materials and Methods

Preparation of CEA for solution scattering

CEA was prepared in two batches, one from a pool of five liver metastases of colon cancer and one from a single liver metastasis. In each case, the extraction procedure was the same (Keep *et al.*, 1978). Slices of frozen liver tissue were homogenised with 0.01 M phosphate-buffered saline at 4°C to give a thick slurry, which was stored at –20°C. To portions of this was added an equal volume of cold 2 M perchloric acid with stirring for five minutes. The resulting homogenate was centrifuged (30,000 *g* for 30 minutes at 4°C) and the supernatant dialysed against four changes of distilled water for 20 hours at 4°C. The extract was concentrated by Amicon PM10 ultrafiltration for the gel filtration step. For the second batch, the extraction buffer also contained the protease inhibitors pepstatin A, chymostatin, leupeptin and antipain, each at 10 mg/l. Sodium azide (0.02% w/v) was also added.

The crude extract from the first batch was purified by gel filtration on Sepharose 6B and Sephacryl S-300 in 0.05 M sodium phosphate buffer (pH 7.5), with 0.02% sodium azide. The CEA-containing fractions were identified by radioimmunoassay and by double diffusion in agar against rabbit anti-CEA. The peak corresponding to an M_r of 180,000 was taken in each case. A 4 mg sample of CEA as determined by radioimmunoassay (goat polyclonal anti-CEA) was purified further by immunoadsorption on a column of A5B7 monoclonal anti-CEA (43 mg) coupled to Sepharose 4B. Non-bound material was recovered by elution with phosphate-buffered saline (12 mM sodium, potassium phosphate (pH 7.4), 140 mM NaCl) and the bound material by elution with 3 M ammonium thiocyanate (40 ml). The bound fraction was concentrated by ultrafiltration at 4°C in an Amicon stirred cell with a PM10 membrane to 3.7 ml, then dialysed overnight against phosphate-buffered saline and filtered using a 0.2 μm filter. Aggregates were removed by fast protein liquid chromatography gel filtration on a Superose-12 (Pharmacia) column with two runs, each with 200 μl of CEA applied to the column and eluted with phosphate-buffered saline. Fractions (20 drops) were collected and tested against PK1G goat anti-CEA antiserum by double diffusion in agar. Fractions 9 and 10 from each run were pooled, filtered (0.2 μm filter) and concentrated by Centricon-30 (Amicon) centrifugation to 380 μl. A 30 μl aliquot was removed for characterisation and the remainder (0.90 mg CEA/ml by Hybritech CEA assay) was used for scattering or sedimentation analyses. The second batch of perchloric acid-extracted CEA was processed as for the first batch. Each batch gave a single diffuse band on non-reduced SDS-PAGE at a molecular mass of 200,000 Da. Whilst perchloric acid cleavage reduces CEA yields (Kimball & Brattain, 1978), antibodies raised to such CEA preparations have been successfully used to locate colonic tumours in patients (Lane *et al.*, 1994).

Preparations were cleared of aggregates shortly before scattering data collection by gel filtration through Superose-12 using a fast protein liquid chromatography system (Pharmacia). Proteins were dialysed at 6°C with four changes of buffer over at least 36 hours into phosphate-buffered saline in H_2O (X-rays) or 99.8% ^2H_2O (neutrons). The amino acid composition of CEA was taken from Oikawa *et al.* (1987) after exclusion of the signal peptide and membrane-spanning regions. Yamashita *et al* (1987, 1989) showed that 92% of CEA oligosaccharides are the complex type with a standard $Man_3GlcNAc_2$ core. In this core structure, 40% contain an additional bisecting GlcNAc residue, and 86% contain a further fucose moiety on the proximal GlcNAc residue. On average, there are 2.7 outer chain branches attached to the core, which are mainly Galβ1 → 4GlcNAc (43.8%) or Galβ1 → 4(Fucα1 → 3)GlcNAc (35.1%) repeats. The full analysis indicated that an average oligosaccharide chain contained 3.1 Man, 5.5 GlcNAc, 3.1 Gal and 2.7 Fuc residues. The sialic acid content per oligosaccharide chain was estimated to be 0.6 (Chandrasekaran *et al.*, 1983; Slayter & Coligan, 1975; Hammarström *et al.*, 1975; Kessler *et al.*, 1978; Pavlenko *et al.*, 1990). The CEA carbohydrate was therefore represented by 28 triantennary complex-type oligosaccharide chains of composition $Man_3GlcNAc_6Gal_3Fuc_3NeuNAc_1$, which were located at putative NXT or NXS sites in the CEA sequence (except when X = Pro). This composition is 53% (w/w) of CEA and is within error of the carbohydrate analyses reported by Kessler *et al.* (1978) and Chandrasekaran *et al.* (1983). The total M_r of CEA is calculated as 152,500, which is 15% less than that of the commonly reported value of 180,000 for CEA from SDS-PAGE (Slayter & Coligan, 1975). This M_r difference is often observed for heavily glycosylated glycoproteins (Gordon, 1975). Protein concentrations for M_r calculations were calculated from absorbance measurements at 280 nm using an absorption coefficient (1%, 1 cm) of 6.41 calculated from the CEA composition by the corrected Wetlaufer procedure (Perkins, 1986).

Synchrotron X-ray data collection at Station 8.2 at SRS

X-ray scattering data were obtained in six independent sessions using the low-angle solution scattering camera at Station 8.2 (Towns-Andrews *et al.*, 1989) at the SRS Daresbury Laboratory, Warrington, UK. Experiments were performed with beam currents in a range of 120 to 230 mA and a ring energy of 2.00 GeV. Samples were

measured for ten minutes in ten time frames for protein concentrations that ranged between 1.2 and 7.0 mg/ml. The use of a 500-channel quadrant detector (Worgan *et al.*, 1990) with sample-detector distances of 3.26 m to 3.54 m resulted in a usable Q range between 0.07 and 2.2 nm^{-1}. ($Q = 4\pi \sin\theta/\lambda$; 2θ = scattering angle, λ = wavelength). The Q range was calibrated using fresh, wet, slightly stretched rat tail collagen, based on a diffraction spacing of 67.0 nm. Samples were held in Perspex cells of sample volume 20 µl, contained within mica windows of thickness between 10 and 15 µm, and cooled at 15°C. Buffers and samples were measured in alternation for equal times to minimise background subtraction errors. Data were accepted only if the subsequent Guinier plots were linear and reproducible in repeated measurements. Each of the ten time frames was individually checked using Guinier analysis to check for the absence of time-dependent radiation-damage effects. Data reduction was performed using the standard Daresbury software package OTOKO (P. Bendall, J. Bordas, M. H. C. Koch & G. R. Mant, EMBL Hamburg and Daresbury Laboratory, unpublished software). Curves were normalised using an ion chamber monitor positioned after the sample for individual runs, and a detector response measured for at least eight hours using a uniform ^{55}Fe radioactive source. Reduced curves were calculated by subtraction of the buffer runs from those of the samples.

Pulsed neutron data collection at Instrument LOQ at ISIS

Neutron scattering data were obtained in four different beam sessions on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, UK (Heenan & King, 1993). The pulsed neutron beam was derived from proton beam currents of 160 to 190 µA. Monochromatisation was achieved using time-of-flight techniques. A ^3He ORDELA wire detector was employed to record intensities at a fixed sample-to-detector distance of 4.3 m. The samples and ^2H_2O buffers were measured in 2 mm thick rectangular Hellma cells positioned in a thermostatted rack at 15°C. Data acquisitions were for fixed totals of 4.0×10^6 monitor counts in runs lasting 50 to 60 minutes each for protein concentrations of 3.6 to 7.3 mg/ml. Spectral intensities were normalised relative to the scattering from a partially deuterated polystyrene standard (Wignall & Bates, 1987). Transmissions were measured for all samples and backgrounds. Reduction of the raw data collected in 100 time frames of 64×64 cells utilised the standard ISIS software package COLETTE (Heenan *et al.*, 1989). Scattered intensities were binned into individual diffraction patterns based on the wavelength range from 0.22 nm to 1.00 nm, and were corrected for a linear wavelength-dependence of the transmission measurements. The patterns were merged to give the full curve $I(Q)$ in a maximal Q range of 0.05 to 2.2 nm^{-1}. The Q range was based on 0.04% logarithmic increments, which was optimal both for Guinier analyses at low Q, and for better signal-noise ratios at large Q.

Analysis of reduced X-ray and neutron data

In a given solute-solvent contrast, the radius of gyration R_G is a measure of structural elongation if the internal inhomogeneity of scattering densities has no effect. Guinier analyses at low Q give the R_G and the forward scattering at zero angle $I(0)$ (Glatter & Kratky, 1982):

$$\ln I(Q) = \ln I(0) - R_G^2 Q^2/3$$

This expression is valid in a QR_G range up to 0.7 for extended rod-like particles, and is approximate in a QR_G up to 1.5 in which it underestimates the true R_G. The relative $I(0)/c$ values (c = sample concentration) for samples measured in the same buffer during a data session gives the relative molecular mass (M_r values) of the proteins when referenced against a suitable standard (Kratky, 1963; Wignall & Bates, 1987). If the structure is elongated, the mean radius of gyration of the cross-sectional structure R_{XS} and the mean cross-sectional intensity at zero angle $[I(Q)Q]_{Q\to 0}$ (Hjelm, 1985) can be obtained from:

$$\ln[I(Q)Q] = \ln[I(Q)Q]_{Q\to 0} - R_{XS}^2 Q^2/2$$

The R_G and R_{XS} analyses lead to the triaxial dimensions of the macromolecule. If the structure can be represented by an elongated elliptical cylinder, $L = [12(R_G^2 - 2R_{XS}^2)]^{1/2}$, where L is its length (Glatter & Kratky, 1982). Alternatively, L is given by $\pi I(0)/[I(Q)Q]_{Q\to 0}$ (Perkins *et al.*, 1986). The two semi-axes, A and B, of the elliptical cylinder are calculated by combining the dry or hydrated volume V ($V = \pi$ ABL) with the R_{XS} value ($R_{XS}^2 = (A^2 + B^2)/4$). The hydrated volume is obtained on the basis of a hydration of 0.3 g of water/g of glycoprotein and 0.0245 nm^3 per water molecule (Perkins, 1986). Data analyses employed an interactive graphics program SCTPL4 (A. S. Nealis & S. J. Perkins, unpublished software) on a Silicon Graphics 4D35S workstation.

Indirect transformation of the scattering data $I(Q)$ in reciprocal space into real space to give $P(r)$ was carried out using the ITP-91 program (Glatter & Kratky, 1982):

$$P(r) = \frac{1}{2\pi^2}\int_0^\infty I(Q)Qr\sin(Qr)\mathrm{d}Q$$

$P(r)$ corresponds to the distribution of distances r between any two volume elements within one particle weighted by the product of their respective electron or nuclear densities relative to the solvent density. This offers an alternative calculation of the R_G and $I(0)$ that is based on the full scattering curve, and gives the maximum dimension of the macromolecule L. For CEA, the X-ray $I(Q)$ contained 247 data points extending out to 2.04 nm^{-1} and was fitted with ten splines with D_{max} set as 35 nm. The neutron $I(Q)$ contained 78 data points extending out to 2.01 nm^{-1} and was fitted with six splines with D_{max} set as 35 nm. $P(r)$ was defined by 101 points. Criteria for the correct choice of parameters for $P(r)$ were: (1) $P(r)$ should exhibit positive values; (2) the R_G from ITP and Guinier analyses should agree; (3) $P(r)$ should be zero when r is zero; and (4) $P(r)$ should be stable and reproducible for different experimental $I(Q)$ curves when the number of splines and D_{max} is varied over a reasonable range. L was determined from $P(r)$ when this became zero at large r; however, errors in L can be significant as a result of the low intensity of $P(r)$ in this region.

Automated Debye scattering curve modelling of CEA

INSIGHT II V2.3.0 molecular graphics software (Biosym Technologies Inc.) on Silicon Graphics Indigo and Indy workstations (R3000 and R4000 series with 48

to 64 Mb memory) were utilised for all manipulations. Atomic coordinates for the CEA-related cell surface proteins CD2 and CD4 corresponded to pairs of V-type and C2-type Ig folds (Brookhaven database codes; human CD4 domains 1 and 2, 1cdh, 1cdi, 3cd4, 1cd4; rat CD4 domains 3 and 4, 1cid; human and rat CD2 domains 1 and 2, 1cdb, 1hnf, 1hng), except for 1cdb, which is a single V-type domain and corresponds to an NMR solution structure. The CEA oligosaccharide structure was based on the nine-residue structure in the Fc fragment of human IgG1 KOL (Brookhaven code 1fc1), to which seven additional carbohydrate residues were attached at appropriate positions to generate the full structure with composition as above. For the AUTOSCT automated curve fitting procedure (Beavil *et al.*, 1995), a Biosym Command Language macro was written to rotate each domain relative to one another to generate CEA conformations. The rotational centre of each domain was defined as the α-carbon atom of the C-terminal residue. For each C2-type domain, the *X*-axis was defined by the line joining the N-terminal and C-terminal α-carbon atoms, and the plane of its *Y*-axis was defined by an α-carbon atom. The corresponding *X*-axis and *Y*-axis for the V-type domain were defined by superimposition of the C2-type domain onto the V-type domain (see Results).

A scattering curve was calculated from each CEA coordinate model. A sphere model was created by placing the full atomic coordinate set within a three-dimensional array of cubes, each of side length 0.572 nm. This length is much less than the nominal resolution of $2\pi/Q_{max}$ of the scattering curves (3.1 nm for $Q_{max} = 2.0\ nm^{-1}$ in X-ray experiments; 3.7 nm for $Q_{max} = 1.7\ nm^{-1}$ in neutron experiments). If the number of atoms within a cube exceeded a user-defined cutoff, a sphere of the same volume as the cube (sphere diameter 0.710 nm) was placed at the centre of the cube. This cutoff was determined by the requirement that the total volume of spheres was within 1% of the dry volume of 178.9 nm^3 for CEA calculated from its composition (Chothia, 1975; Perkins, 1986). This cutoff was set as a minimum of three and four atoms to define each protein and carbohydrate sphere, respectively. The total of 485 protein and 474 carbohydrate spheres corresponded, respectively, to volumes of 90.6 nm^3 and 88.3 nm^3 calculated from the CEA composition. X-ray scattering experiments are influenced by a hydration shell surrounding the glycoprotein. A hydration of 0.3 g H_2O/g of glycoprotein and an electrostricted volume of 0.0245 nm^3 per bound water molecule (Perkins, 1986) were used to calculate a hydrated CEA volume of 241.1 nm^3 for X-ray curve fits based on sphere diameters of 0.783 nm. In neutron experiments, the hydration shell was not detectable and the dry sphere models were used for curve fits.

The scattering curve $I(Q)$ was calculated using Debye's Law adapted to spheres, essentially by computing all the distances r from each sphere to the remaining spheres and summing the results. The different scattering densities of protein and carbohydrate were incorporated in two-density scattering curves calculated from the methods of Glatter & Kratky (1982):

$$[I(Q)/I(0)] = g(Q)[n_1\rho_1^2 + n_2\rho_2^2 + 2\rho_1^2\Sigma A_j^{11}(\sin Qr_j/Qr_j) + 2\rho_2^2\Sigma A_j^{22}(\sin Qr_j/Qr_j) + 2\rho_1\rho_2\Sigma A1_j^{12}(\sin Qr_j/Qr_j)] \times (n_1\rho_1 + n_2\rho_2)^{-2}$$

The CEA model is constructed from n_1 and n_2 spheres of different densities ρ_1 and ρ_2; $g(Q) = 3(\sin QR - QR\cos QR)^2/Q^6R^6$ (the squared form factor of the spheres of radius R); A_j^{11}, A_j^{22} and A_j^{12} are the number of distances r_j for that increment of j between the spheres 1 and 1, 2 and 2, and 1 and 2, in that order; the summations are performed for $j = 1$ to m, where m is the number of different distances r_j. For the X-ray data, no correction was applied for wavelength spread or beam divergence as these are thought to be negligible. For the neutron data, a 16% wavelength spread for a nominal λ of 1.0 nm and a beam divergence of 0.016 radian were used as an approximation to correct the calculated neutron scattering curve for the reasons discussed by Mayans *et al.* (1995). The quality of the curve fits was assessed by calculations of the R_G and R_{XS} values of the model from the scattering curve in the same Q ranges used for Guinier fits, and the R-factor in the Q range extending to 1.6 nm^{-1} (denoted as $R_{1.6}$) or to 2.0 nm^{-1} ($R_{2.0}$; Smith *et al.*, 1990; Beavil *et al.*, 1995). Curve fits were assessed using Microsoft Excel spreadsheets in which the absence of steric overlap was verified from the volume of the model, which should be close to 959 spheres. The models were filtered and sorted on the basis of the observed and calculated R_G and R_{XS} values and the R-factor values.

Hydrodynamic analyses and modelling of CEA

Sedimentation coefficients $s^o_{20,w}$ for CEA were measured at 0.42, 0.59 and 0.83 mg/ml at 20°C on a Beckmann XL-A analytical ultracentrifuge operated at 70,000 *g* and equipped with scanning absorption optics at the National Centre for Macromolecular Hydrodynamics, Leicester. Traces were measured at 280 nm and analysed using a digitising pad interfaced with an Apple II computer to yield experimental $s^o_{20,w}$ values after correction for the density and viscosity of the buffer. Frictional coefficients f were calculated from the $s^o_{20,w}$ values for comparison with f values calculated directly from the hydrated models with approximately 959 spheres used for fits of the X-ray scattering curves (but now using non-overlapping spheres to satisfy the algorithm in use). These f values were derived using the modified Oseen tensor procedure in the program GENDIA (García de la Torre & Bloomfield, 1977a,b; Perkins *et al.*, 1993), and were imported into spreadsheets for joint analyses with the scattering fits. A more recent hydrodynamic modelling program HYDRO permitted the use of overlapping spheres (Garcia de la Torre, 1989). As this gave results that were very similar to those from GENDIA at the cost of much increased CPU time, HYDRO was used only as a control for the outcome of the GENDIA simulations.

Acknowledgements

M.K.B. gratefully acknowledges a Clement Wheeler-Bennett Trust studentship. We thank Professor K. D. Bagshawe for his interest and useful discussions. This work was supported by the Biotechnology and Biological Sciences Research Council (M.O.M.) and the Cancer Research Campaign (J.D.T., R.H.J.B. and P.A.K.). We gratefully thank Mr A. J. Beavil for generous computational advice, Dr W. Bras (Daresbury Laboratory), Dr R. K. Heenan and Dr S. M. King (Rutherford-Appleton Laboratory) for instrumental support, and Dr O. Byron (Leicester) for hydrodynamic measurements.

References

Aleshin, A., Golubev, A., Firsov, L. M. & Honzatko, R. B. (1992). Crystal structure of glucoamylase from *Aspergillus awamori* var. X100 to 2.2 Å resolution. *J. Biol. Chem.* **267**, 19291–19298.

Aleshin, A. E., Hoffman, C., Firsov, L. M. & Honzatko, R. B. (1994). Refined crystal structures of glucoamylase from *Aspergillus awamori* var. X100. *J. Mol. Biol.* **238**, 575–591.

Bates, P. A., Luo, J. & Sternberg, M. J. E. (1992). A predicted three-dimensional structure for the carcinoembryonic antigen (CEA). *FEBS Letters*, **301**, 207–214.

Beavil, A. J., Beavil, R. L., Chan, C. M. W., Cook, J. P. D., Gould, H. J., Henry, A. J., Owens, R. J., Shi, J., Sutton, B. J. & Young, R. J. (1993). Structural basis of the IgE-FcεRI interaction. *Biochem. Soc. Trans.* **21**, 968–972.

Beavil, A. J., Young, R. J., Sutton, B. J. & Perkins, S. J. (1995). Bent domain structure of recombinant human IgE-Fc in solution by X-ray and neutron scattering in conjunction with an automated curve fitting procedure. *Biochemistry*, **34**, 14449–14461.

Becker, J. W., Erickson, H. P., Hoffman, S., Cunningham, B. A. & Edelman, G. M. (1989). Topology of cell adhesion molecules. *Proc. Natl Acad. Sci. USA*, **86**, 1088–1092.

Benchimol, S., Fuks, A., Jothy, S., Beauchemin, N., Shirota, K. & Stanners, C. P. (1989). Carcinoembryonic antigen, a human tumor marker, functions as an intercellular adhesion molecule. *Cell*, **57**, 327–334.

Berendt, A. R., McDowall, A., Craig, A. G., Bates, P. A., Sternberg, M. J. E., Marsh, K., Newbold, C. I. & Hogg, N. (1992). The binding site on ICAM-1 for *Plasmodium falciparum*-infected erythrocytes overlaps but is distinct from the LFA-1 binding site. *Cell*, **68**, 71–81.

Bode, W., Meyer, E., Jr & Powers, J. C. (1989). Human leukocyte and porcine pancreatic elastase: X-ray crystal structures, mechanism, substrate specificity and mechanism-based inhibitors. *Biochemistry*, **28**, 1951–1963.

Bodian, D. L., Jones, E. Y., Harlos, K., Stuart, D. I. & Davis, S. J. (1994). Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2.5 Å resolution. *Structure*, **2**, 755–766.

Brady, R. L., Dodson, E. J., Dodson, G. G., Lange, G., Davis, S. J., Williams, A. F. & Barclay, A. N. (1993). Crystal structure of domains 3 and 4 of rat CD4: relation to the NH_2-terminal domains. *Science*, **260**, 979–983.

Campbell, I. D. & Spitzfaden, C. (1994). Building proteins with fibronectin type III modules. *Structure*, **2**, 333–337.

Chandrasekaran, E. V., Davila, M., Nixon, D. W., Goldfarb, M. & Mendicino, J. (1983). Isolation and structures of the oligosaccharide units of carcinoembryonic antigen. *J. Biol. Chem.* **258**, 7213–7222.

Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**, 304–308.

Deisenhofer, J. (1981). Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from *Staphylococcus aureus* at 2.9- and 2.8-Å resolution. *Biochemistry*, **20**, 2361–2370.

de Vos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992). Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science*, **255**, 306–312.

Garcia de la Torre, J. (1989). Hydrodynamic properties of macromolecular assemblies. In *Dynamic Properties of Biomolecular Assemblies* (Harding, S. E. & Rowe, A. J., eds), pp. 3–31, Royal Society of Chemistry, Cambridge.

Garcia de la Torre, J. & Bloomfield, V. A. (1977a). Hydrodynamic properties of macromolecular complexes. 1. Translation. *Biopolymers*, **16**, 1747–1761.

Garcia de la Torre, J. & Bloomfield, V. A. (1977b). Hydrodynamics of macromolecular complexes. 3. Bacterial viruses. *Biopolymers*, **16**, 1779–1793.

Glatter, O. & Kratky, O. (1982). Editors of *Small-angle X-ray Scattering*. Academic Press, New York.

Gordon, A. H. (1975). In *Electrophoresis of Proteins in Polyacrylamide and Starch Gels* (Work, T. S. & Work, E., eds), pp. 153s–164s, North Holland Publ. Co., Amsterdam.

Hammarström, S., Engvall, E., Johansson, B. G., Svensson, S., Sundblad, G. & Goldstein, I. J. (1975). Nature of the tumor-associated determinant(s) of carcinoembryonic antigen. *Proc. Natl Acad. Sci. USA*, **72**, 1528–1532.

Hammarström, S., Shively, J. E., Paxton, R. J., Beatty, B. G., Larsson, A., Ghosh, R., Bormer, O., Buchegger, F., Mach, J.-P., Burtin, P., Seguin, P., Darbouret, B., Degorce, F., Sertour, J., Jolu, J. P., Fuks, A., Kalthoff, H., Schmiegel, W., Arndt, R., Kloppel, G., von Kleist, S., Grunert, F., Schwarz, K., Matsuoka, Y., Kuroki, M., Wagener, C., Weber, T., Yachi, A., Imai, K., Hishikawa, N. & Tsujisaki, M. (1989). Antigenic sites in carcinoembryonic antigen. *Cancer Res.* **49**, 4852–4858.

Harlos, K., Martin, D. M. A., O'Brien, D. P., Jones, E. Y., Stuart, D. I., Polikarpov, I., Miller, A., Tuddenham, E. G. D. & Boys, C. W. G. (1994). Crystal structure of the extracellular region of human tissue factor. *Nature*, **370**, 662–666.

Hashino, J., Fukuda, Y., Iwao, K., Krop-Watorek, A., Oikawa, S., Nakazato, H. & Nakanishi, T. (1993). Production and characterization of monoclonal antibodies to N-domain and domain III of carcinoembryonic antigen. *Biochem. Biophys. Res. Commun.* **197**, 886–893.

Heenan, R. K. & King, S. M. (1993). Development of the small-angle diffractometer LOQ at the ISIS pulsed neutron source. In *Proceedings of an International Seminar on Structural Investigations at Pulsed Neutron Sources, Dubna, 1st–4th September 1992. Report E3-93-65*, Joint Institute for Nuclear Research, Dubna.

Heenan, R. K., King, S. M., Osborn, R. & Stanley, H. B. (1989). *COLETTE Users Guide*. Internal publication RAL-89-128, Rutherford Appleton Laboratory, Didcot, UK.

Hefta, S. A., Hefta, L. J. F., Lee, T. D., Paxton, R. J. & Shively, J. E. (1988) Carcinoembryonic antigen is anchored to membranes by covalent attachment to a glycosylphosphadidylinositol moiety: identification of the ethanolamine linkage site. *Proc. Natl Acad. Sci. USA*, **85**, 4648–4652.

Hjelm, R. P. (1985). The small-angle approximation of X-ray and neutron scatter from rigid rods of non-uniform cross section and finite length. *J. Appl. Crystallog.* **18**, 452–460.

Ikeda, S., Kuroki, M., Haruno, M., Oikawa, S., Nakazato, H., Kosaki, G. & Matsuoka, Y. (1992). Epitope

mapping of the carcinoembryonic antigen with various related recombinant proteins expressed in Chinese hamster ovary cells and 25 distinct monoclonal antibodies. *Mol. Immunol.* **29**, 229–240.

Jean, F., Malapert, P., Rougon, G. & Barbet, J. (1988). Cell membrane, but not circulating, carcinoembryonic antigen is linked to a phosphatidylinositol-containing hydrophobic domain. *Biochem. Biophys. Res. Commun.* **155**, 794–800.

Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. & Stuart, D. I. (1992). Crystal structure at 2.8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature*, **360**, 232–239.

Jones, E. Y., Harlos, K., Bottomley, M. J., Robinson, R. C., Driscoll, P. C., Edwards, R. M., Clements, J. M., Dudgeon, T. J. & Stuart, D. I. (1995). Crystal structure of an integrin-binding fragment of vascular cell adhesion molecule-1 at 1.8 Å resolution. *Nature*, **373**, 539–544.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Keep, P. A., Leake, B. A. & Rogers, G. T. (1978). Extraction of CEA from tumour tissue, foetal colon and patients' sera, and the effect of perchloric acid. *Brit. J. Cancer*, **37**, 171–182.

Kessler, M. J., Shively, J. E., Pritchard, D. G. & Todd, C. W. (1978). Isolation, immunological characterization and structural studies of a tumour antigen releated to carcinoembryonic antigen. *Cancer Res.* **38**, 1041–1048.

Killeen, N., Moessner, R., Arvieux, J., Willis, A. & Williams, A. F. (1988). The MRC OX-45 antigen of rat leukocytes and endothelium is in a subset of the immunoglobulin superfamily with CD2, LFA-3 and carcinoembryonic antigens. *EMBO J.* **7**, 3087–3091.

Kimball, P. M. & Brattain, M. G. (1978). A comparison of methods for the isolation of carcinoembryonic antigen. *Cancer Res.* **38**, 619–623.

Kratky, O. (1963). X-ray small angle scattering with substances of biological interest in diluted solutions. *Prog. Biophys. Chem.* **13**, 105–173.

Krupey, J., Gold, P. & Freedman, S. O. (1968). Physicochemical studies of the carcinoembryonic antigens of the human digestive system. *J. Exp. Med.* **128**, 387–398.

Lane, D. M., Eagle, K. F., Begent, R. H. J., Hope-Stone, L. D., Green, A. J., Casey, J. L., Keep, P. A., Kelly, A. M. B., Ledermann, J. A., Glaser, M. G. & Hilson, A. J. W. (1994). Radioimmunotherapy of metastatic colorectal tumours with iodine-131-labelled antibody to carcinoembryonic antigen: phase I/II study with comparative biodistribution of intact and $F(ab')_2$ antibodies. *Brit. J. Cancer*, **70**, 521–525.

Lisowska, E., Krop-Watorek, A. & Sedlaczek, P. (1983). The dimeric structure of carcinoembryonic antigen (CEA). *Biochem. Biophys. Res. Commun.* **115**, 206–211.

Mayans, M. O., Coadwell, W. J., Beale, D., Symons, D. B. A. & Perkins, S. J. (1995). Demonstration by pulsed neutron scattering that the arrangement of the Fab and Fc fragments in the overall structures of bovine IgG1 and IgG2 in solution is similar. *Biochem. J.* **311**, 283–291.

Murakami, M., Kuroki, M., Arakawa, F., Kuwahara, M., Oikawa, S., Nazakato, H. & Matsuoka, Y, (1995). A reference of the GOLD classification of monoclonal antibodies against carcinoembryonic antigen to the domain structure of the carcinoembryonic antigen molecule. *Hybridoma*, **14**, 19–28.

Nap, M., Hammarström, M.-L., Börmer, O., Hammarström, S., Wagener, C., Handt, S., Schreyer, M., Mach, J.-P., Buchegger, F., von Kleist, S., Grunert, F., Seguin, P., Fuks, A., Holm, R. & Lamerz, R. (1992). Specificity and affinity of monoclonal antibodies against carcinoembryonic antigen. *Cancer Res.* **52**, 2329–2339.

Oikawa, S., Nakazato, H. & Kosaki, G. (1987). Primary structure of human carcinoembryonic antigen (CEA) deduced from cDNA sequence. *Biochem. Biophys. Res. Commun.* **142**, 511–518.

Oikawa, S., Inuzuka, C., Kuroki, M., Arakawa, F., Matsuoka, Y., Kosaki, G. & Nakazato, H. (1991). A specific heterotypic cell adhesion between members of carcinoembryonic antigen family W272 and NCA is mediated by N-domains. *J. Biol. Chem.* **266**, 7995–8001.

Pavlenko, A. F., Chikalovets, I. V., Kurika, A. V., Glasunov, V. P., Mikhalyuk, L. V. & Ovodov, Y. S. (1990). Carcinoembryonic antigen, its spatial structure and localisation of antigenic determinants. *Tumour Biol.* **11**, 306–318.

Paxton, R. J., Mooser, G., Pande, H., Lee, T. D. & Shively, J. E. (1987). Sequence analysis of carcinoembryonic antigen: identification of glycosylation sites and homology with the immunoglobulin supergene family. *Proc. Natl Acad. Sci. USA*, **84**, 920–924.

Perkins, S. J. (1986). Protein volumes and hydration effects: the calculation of partial specific volumes, neutron scattering matchpoints and 280 nm absorption coefficients for proteins and glycoproteins from amino acid sequences. *Eur. J. Biochem.* **157**, 169–180.

Perkins, S. J. (1988). X-ray and neutron solution scattering. *New Comp. Biochem.* **11B**, 143–264.

Perkins, S. J., Chung, L. P. & Reid, K. B. M. (1986). Unusual ultrastructure of complement component C4b-binding protein of human complement by synchrotron X-ray scattering and hydrodynamic analysis. *Biochem. J.* **233**, 799–807.

Perkins, S. J., Smith, K. F., Amatayakul, S., Ashford, D., Rademacher, T. W., Dwek, R. A., Lachmann, P. J. & Harrison, R. A. (1990). The two-domain structure of the native and reaction centre cleaved forms of C1 inhibitor of human complement by neutron scattering. *J. Mol. Biol.* **214**, 751–763.

Perkins, S. J., Nealis, A. S., Sutton, B. J. & Feinstein, A. (1991). The solution structure of human and mouse immunoglobulin IgM by synchrotron X-ray scattering and molecular graphics modelling: a possible mechanism for complement activation. *J. Mol. Biol.* **221**, 1345–1366.

Perkins, S. J., Smith, K. F., Kilpatrick, J. M., Volanakis, J. E. & Sim, R. B. (1993). Modelling of the serine protease fold by X-ray and neutron scattering and sedimentation analyses: its occurrence in factor D of the complement system. *Biochem. J.* **295**, 87–99.

Ryu, S.-E., Kwong, P. D., Truneh, A., Porter, T. G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N.-H., Axel, R., Sweet, R. W. & Hendrickson, W. A. (1990). Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature*, **348**, 419–426.

Schwarz, K., Mehnert-Solzer, C., von Kleist, S. & Grunert, F. (1988). Analysis of the specificity of CEA reactive monoclonal antibodies. Immunological support for the domain model of CEA. *Mol. Immunol.* **25**, 889–898.

Slayter, H. S. & Codington, J. F. (1973). Size and configuration of glycoprotein fragments cleaved from tumor cells by proteolysis. *J. Biol. Chem.* **248**, 3405–3410.

Slayter, H. S. & Coligan, J. E. (1975). Electron microscopy and physical characterization of the carcinoembryonic antigen. *Biochemistry*, **14**, 2323–2330.

Smith, K. F., Harrison, R. A. & Perkins, S. J. (1990). Structural comparisons of the native and reaction centre cleaved forms of α_1-antitrypsin by neutron and X-ray solution scattering. *Biochem. J.* **267**, 203–212.

Staunton, D. E., Dustin, M. L., Erickson, H. P. & Springer, T. A. (1990). The arrangement of the immunoglobulin-like domains of ICAM-1 and the binding sites for LFA-1 and rhinovirus. *Cell*, **61**, 243–254.

Taylor, W. R. (1988). A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* **28**, 161–169.

Thompson, J. & Zimmermann, W. (1988). The carcinoembryonic antigen gene family: structure, expression and evolution. *Tumor Biol.* **9**, 63–83.

Thompson, J. A., Grunert, F. & Zimmermann, W. (1991). Carcinoembryonic antigen gene family: molecular biology and clinical perspectives. *J. Clin. Lab. Anal.* **5**, 344–366.

Towns-Andrews, E., Berry, A., Bordas, J., Mant, G. R., Murray, P. K., Roberts, K., Sumner, I., Worgan, J. S., Lewis, R. & Gabriel, A. (1989). Time-resolved X-ray diffraction station: X-ray optics, detectors and data acquisition. *Rev. Sci. Instrum.* **60**, 2346–2349.

Wang, J., Yan, Y., Garrett, T. P. J., Liu, J., Rodgers, D. W., Garlick, R. L., Tarr, G. E., Husain, Y., Reinherz, E. L. & Harrison, S. C. (1990). Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature*, **348**, 411–418.

Wignall, G. D. & Bates, F. S. (1987). Absolute calibration of small angle neutron scattering data. *J. Appl. Crystallog.* **20**, 28–40.

Williams, A. F. & Barclay, A. N. (1988). The immunoglobulin superfamily—domains for cell surface recognition. *Annu. Rev. Immunol.* **6**, 381–405.

Worgan, J. S., Lewis, R., Fore, N. S., Sumner, I. L., Berry, A., Parker, B., D'Annunzio, F., Martin-Fernandez, M. L., Towns-Andrews, E., Harries, J. E., Mant, G. R., Diakun, G. P. & Bordas, J. (1990). The application of multiwire X-ray detectors to experiments using synchrotron radiation. *Nucl. Instrum. Methods Phys. Res.* **A291**, 447–454.

Wyss, D. F., Choi, J. S., Li, J., Knoppers, M. H., Willis, K. J., Arulanandum, A. R. N., Smolyar, A., Reinherz, E. L. & Wagner, G. (1995). Conformation and function of the N-linked glycan in the adhesion domain of human CD2. *Science*, **269**, 1273–1278.

Yamashita, K., Totani, K., Kuroki, M., Matsuoka, Y., Ueda, I. & Kobata, A. (1987). Structural studies of the carbohydrate moieties of carcinoembryonic antigens. *Cancer Res.* **47**, 3451–3459.

Yamashita, K., Totani, K., Iwaki, Y., Kuroki, M., Matsuoka, Y., Endo, T. & Kobata, A. (1989). Carbohydrate structures of nonspecific cross-reacting antigen-2, a glycoprotein purified from meconium as an antigen cross-reacting with anticarcinoembryonic antigen antibody. *J. Biol. Chem.* **264**, 17873–17881.

Zhou, H., Fuks, A., Alcaraz, G., Bolling, T. J. & Stanners, C. P. (1993). Homophilic adhesion between Ig superfamily carcinoembryonic antigen molecules involves double reciprocal bonds. *J. Cell Biol.* **122**, 951–960.

Edited by R. Huber

(Received 28 September 1995; received in revised form 10 February 1996; accepted 27 March 1996)

ELSEVIER

International Journal of Biological Macromolecules
22 (1998) 1–16

INTERNATIONAL JOURNAL OF
Biological
Macromolecules
STRUCTURE, FUNCTION AND INTERACTIONS

Review article

Molecular structures from low angle X-ray and neutron scattering studies

S.J. Perkins *, A.W. Ashton, M.K. Boehm, D. Chamberlain

Department of Biochemistry and Molecular Biology, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, UK

Received 6 October 1997; accepted 17 October 1997

Abstract

Molecular structures can be extracted from solution scattering analyses of multidomain or oligomeric proteins by a new method of constrained automated scattering curve fits. Scattering curves are calculated using a procedure tested by comparisons of crystal structures with experimental X-ray and neutron data. The domains or subunits in the protein of interest are all represented by atomic coordinates in order to provide initial constraints. From this starting model, hundreds or thousands of different possible structures are computed, from each of which a scattering curve is computed. Each model is assessed for steric overlap, radii of gyration and *R*-factors in order to leave a small family of good fit models that corresponds to the molecular structure of interest. This method avoids the tedium of curve fitting by hand and error limits on the ensuing models can be described. For single multidomain proteins, the key constraint is the correct stereochemical connections between the domains in all the models. Successful applications to determine structures are summarised for the Fab and Fc fragments in immunoglobulin G, the three domain pairs in the Fc subunit of immunoglobulin E and the seven domains in carcinoembryonic antigen. For oligomeric proteins, the key constraint is provided by symmetry and successful analyses were performed for the association of the monomers of the bacterial amide sensor protein AmiC to form trimers and pentameric serum amyloid P component to form decameric structures. The successful analysis of the heterodimeric complex of tissue factor and factor VIIa required the use of constraints provided from biochemical data. The outcome of these analyses is critically appraised, in particular the biological significance of structures determined by these solution scattering curve fits. © 1998 Elsevier Science B.V.

Keywords: X-ray and neutron scattering; Molecular modelling; Multidomain proteins

1. Introduction

The structural arrangement of domains or subunits in multidomain or oligomeric proteins in dilute solutions can be determined by X-ray and neutron scattering studies at resolutions of 3 nm in near-physiological conditions [1–3], as a function of pH, temperature or another variable of interest. X-ray scattering using synchrotron radiation provides high quality curves that are minimally affected by instrumental geometry. X-rays visualise the macromolecule in a high positive solute–solvent contrast. This is analogous to seeing a glass rod in a beaker of water as the consequence of differences in the refractive indices of water and glass. Neutron scattering provides the means to visualise macromolecules in a range of positive and negative contrasts by the use of light and heavy water buffers. This now corresponds to a

Abbreviations: CEA, Carcinoembryonic antigen; FVIIa, Factor VIIa; IgG, Immunoglobulin G; IgE, Immunoglobulin E; IgM, Immunoglobulin M; SAP, Serum amyloid P component.

* Corresponding author. Tel.: +44 171 7940500/4210; fax: +44 171 7949645; e-mail: steve@rfhsm.ac.uk

PII S0141-8130(97)00088-3

range of different images of a multilayered plastic/glass rod with different refractive indices in the beaker viewed by the use of oils of higher and lower refractive indices than that of glass. Neutrons can therefore provide information on the structure of lipids, protein, carbohydrate and DNA/RNA within the macromolecule, as well as providing other advantages such as the absence of radiation damage effects sometimes seen with the use of synchrotron radiation. Scattering is complementary in scope to electron microscopy methods which directly visualise the structure of macromolecules in semi-crystalline forms or when flattened or stained on a template, although the conditions of measurements can potentially perturb the structure of interest. It is also complementary to analytical ultracentrifugation, which provides limited information on macromolecular elongation from sedimentation coefficients, as well as on molecular weights and macromolecular equilibria if relevant.

Traditionally solution scattering is seen as an enabling method that provides gross macromolecular information. Data collection to obtain scattering curves $I(Q)$ and their analysis to yield the overall radius of gyration R_G, the radius of gyration of the cross-section R_{XS} if applicable and the distance distribution function $P(r)$ will yield a set of dimensions on three axes for the macromolecule [4]. Molecular weight determinations from the forward scattering at zero scattering angle $I(0)/c$ (where c is the protein concentration in mg/ml) will identify the degree of oligomerisation if present. The modelling of the scattering curves by ellipsoids or assemblies of Debye spheres will verify the correct interpretation of the scattering data and enable the structure to be visualised. Such modelling is constrained by the known volume of the multidomain or multi-subunit protein in question, which determines the volume of the ellipsoids or spheres to be used and this can be calculated from its sequence. It can be refined by complementary information from the images visualised by electron microscopy or sedimentation coefficients from analytical ultracentrifugation. Recent developments in spherical harmonics show that this can create outline macromolecular shapes from scattering data and this represents an alternative and rapid means of interpreting scattering curves. The limitation of this approach is that no advantage is taken of relevant known atomic structures and consequently the biological significance of the outline shape is relatively restricted.

The impact of solution scattering on biology would be significantly improved if it were possible to derive molecular structures from the information contained in scattering curves. The availability of atomic structures from scattering would enable the biological significance of the structure to be perceived more readily. Recent developments based on the rapidly increasing numbers of atomic structures for small domains or subunits found in these structures from crystallography and NMR have begun to make this goal realisable. Thus these small structures can be assembled to reproduce the full macromolecule and used to calculate a scattering curve to determine whether it is compatible with the experimental curve. In other words, the modelling of the scattering curve is constrained by not only the known macromolecular volume, but also by the known atomic structures within the macromolecule, the known steric connections between these structures and any other known constraints. There is some analogy here with the fitting of amino acid coordinates to either a raw electron density map in a crystal structure or to the NMR parameters of assigned signals in 2D- and 3D-NMR spectroscopy in order to determine a protein structure. Such scattering curve fits accordingly require two developments, namely the verification of a reliable method to calculate scattering curves from atomic coordinates, together with an automated method to optimise and determine the best-fit macromolecular structure to a given scattering curve, as well as an estimation of the precision of this structure. Even though as much work again is required to model a scattering curve as it is to perform data collection, reduction and interpretation, the derivation of biologically useful information from the resulting best-fit model will make this worthwhile. This is especially important when it is not possible to crystallise a multidomain protein for reason of interdomain flexibility or high glycosylation.

The potential for the joint use of scattering data with atomic structures was first indicated by the modelling of the 71 domains in the structure of pentameric immunoglobulin M (IgM) [5]. There, the use of structurally homologous crystal structures based on those in immunoglobulin G (IgG) resulted in the assembly of models for four major fragments of IgM as well as for intact IgM that were able to replicate the five X-ray scattering curves in question. Molecular graphics examination of the ensuing IgM structure resulted in the

identification of residues involved in the binding of complement C1q to IgM, as well as permitting an evaluation of the conformational changes that occur in both C1q and IgM upon complexation to trigger complement activation. The IgM study was based on a manual trial-and-error strategy of generating likely structures for the domain fragments and assessing their compatibility with scattering data. This drawback prompted the development of a more automated approach for curve fitting starting from atomic structures, side-by-side with further tests to assess the validity of the curve fit procedures. The purpose of this review is to bring together results from these new calibration studies, together with a diverse range of applications to single unknown multidomain structures and oligomeric or heterodimeric structures (Table 1), in order to illustrate the utility of scattering, automated curve-fits and their limits [5–12].

2. Methods

2.1. X-ray and neutron scattering data collection and analysis

Experimental data collection was performed at European synchrotron and neutron facilities. Synchrotron X-ray scattering data using dilution series of samples in H_2O buffers were obtained at Stations 2.1 or 8.2 at the Synchrotron Radiation Source, Daresbury, UK, using a camera with a quadrant detector and with sample-detector distances of 3–3.5 m. They were reduced using OTOKO [13–15]. Neutron scattering data using dilution series of samples in ^2H_2O buffers (which avoids the high incoherent background of H_2O, as well as providing a high negative solute–solvent contrast) were obtained on the LOQ instrument in the wavelength range 0.2–1.0 nm and a sample-detector distance of 4.3 m at the pulsed neutron source ISIS, at the Rutherford Appleton Laboratory, Didcot, UK and reduced using COLETTE [16]. Neutron scattering data using H_2O or ^2H_2O buffer systems were obtained on Instruments D11 or D17 at the high-flux reactor at the Institut Laue-Langevin (ILL), Grenoble, France, using a wavelength of 1.0–1.1 nm and two different sample-detector distances at 1.4–2.0 m and 3.4–5.0 m. They were reduced using RNILS and SPOLLY [17]. The resulting scattering curves for modelling analyses typically cover a Q range of 0.06–2.3 nm^{-1}, where $Q = 4\pi \sin\theta/\lambda$, the scattering angle is 2θ and the wavelength is λ.

The experimental data were analysed in full prior to curve modelling. In a given solute–solvent contrast, the R_G is a measure of structural elongation if the internal inhomogeneity of scattering density within the macromolecule has no effect. Guinier analyses give the R_G and $I(0)$ values [4]:

$$\ln I(Q) = \ln I(0) - R_G^2 Q^2/3.$$

This expression is valid in a $Q \cdot R_G$ range up to 0.7–1.3, depending on the macromolecular shape. The relative $I(0)/c$ values (where c is the sample concentration) for samples measured in the same buffer during a data session gives the relative molecular weights M_r of the proteins when referenced against a suitable standard [18–20]. For an elongated structure, the R_{XS} and the cross-sectional intensity at zero angle $[I(Q)\cdot Q]_{Q\to 0}$ are obtained [21,22] from:

$$\ln[I(Q)\cdot Q] = \ln[I(Q)\cdot Q]_{Q\to 0} - R_{XS}^2 Q^2/2.$$

The combination of the R_G and R_{XS} analyses yields the maximum macromolecular dimension L in appropriate cases. X-ray and neutron Guinier analyses were processed using a common routine SCTPL5. Indirect transformation of the scattering data in reciprocal space $I(Q)$ into the distance distribution function $P(r)$ in real space was carried out using ITP-91 [4] and/or GNOM [23–25].

$$P(r) = \frac{1}{2\pi^2}\int_0^\infty I(Q)Qr\sin(Qr)\,\mathrm{d}(Q)$$

$P(r)$ offers an alternative calculation of R_G and $I(0)$ which is now based on the full scattering curve and also gives L.

2.2. Automated scattering curve modelling

The modelling of the X-ray and neutron scattering curves is conveniently achieved using small spheres of uniform density to represent the protein structure. The X-ray and neutron scattering curve $I(Q)$ were calculated by an application of Debye's Law adapted to spheres of a single density [4,26]:

$$\frac{I(Q)}{I(0)} = g(Q)\left(n^{-1} + 2n^{-2}\sum_{j=1}^{m} A_j \frac{\sin Qr_j}{Qr_j}\right)$$

$$g(Q) = (3(\sin QR - QR\cos QR))^2/Q^6R^6$$

where $g(Q)$ is the squared form factor for the sphere of radius R, n is the number of spheres

Table 1
Scattering curve fit analyses for six multidomain proteins

Five protein systems (a)–(e)	Molecular weight	Spheres	Cube side (nm)	Search parameters	Number of models	Instrument[a]	Observed R_G (nm)	Fitted R_G (nm)[b]	Q range (nm^{-1})	R-factor (%)
(a) Bovine IgG1	144 000	773–797	0.610	2	≈200	LOQ (neutrons)	5.64 ± 0.28	5.31	0.09–1.55	1.2
Bovine IgG2						LOQ (neutrons)	5.71 ± 0.51			
(b) IgE-Fc	75 300	371	0.658	5	4×9360	St 8.2 (X-rays)	3.52 ± 0.14	3.22	0.13–2.0	3.4
					1	LOQ (neutrons)	3.53 ± 0.15	3.22	0.13–1.5	6.3
(c) CEA	152 500	959[c]	0.572	3	3×4056	St 8.2 (X-rays)	8.0 ± 0.6	8.0	0.12–2.0	4.7
					2×4056	LOQ (neutrons)	8.8 ± 0.5	6.9	0.19–1.6	8.7
(d) AmiC trimers	127 900	1752	0.457	1	21	St 2.1 (X-rays)	3.35 ± 0.05	3.39	0.16–2.0	4.7
				4	176 851			3.39		6.3
				3	$2 \times 39\,041$			3.34, 3.32[d]		4.1, 3.9[d]
(e) SAP pentamer	127 000	2118	0.425	0	3	St 2.1 (X-rays)	3.99 ± 0.11	3.97	0.10–2.0	3.7
						D17 (neutrons)	3.69 ± 0.12	3.80	0.08–2.0	4.0
SAP decamer	254 000	4236	0.425	1	8×80	St 2.1 (X-rays)	4.23 ± 0.12	4.23	0.10–2.0	3.4
						D17 (neutrons)	4.09 ± 0.14	4.13	0.08–2.0	4.7
(f) Factor VIIa	51 400	666	0.452	6	15 625	St 8.2 (X-rays)	3.24 ± 0.08	3.22	0.10–2.0	4.4
					1	LOQ (neutrons)	3.22 ± 0.02		0.11–2.0	6.8
Tissue factor –factor VIIa complex	76 200	1020	0.452	3	4×9261	St 8.2 (X-rays)	3.20 ± 0.02	3.14	0.15–2.0	3.6
					1	LOQ (neutrons)	3.04 ± 0.08		0.15–2.0	7.8

[a] Neutron data correspond to 100% ^2H_2O buffers.
[b] The fitted R_G values correspond to the final model depicted in Fig. 3.
[c] Two-density models with 485 protein and 474 carbohydrate spheres were used for the final fit.
[d] Asymmetric trimers from [9].

filling the body, A_j is the number of distances r_j for that value of j, r_j is the distance between the spheres and m is the number of different distances r_j. The method has been tested with crystal structures for β-trypsin and α_1-antitrypsin [27,28] and more recently with one for pentameric serum amyloid P component [10]. The single density approach is applicable for proteins and for glycoproteins with low carbohydrate contents if equally good curve fits to the same model can be obtained with the X-ray data in positive contrasts and the neutron data in negative contrasts. If systematic curve fit deviations are observed in these two different solute–solvent contrasts, two-density modelling will be required, as exemplified below by carcinoembryonic antigen [8,26].

The stages of the modelling procedure are summarised in Fig. 1. Initial trial models were generated using INSIGHT II (Biosym/MSI, San Diego, USA) using the atomic structures for individual domains in order to determine how best to set up an automated procedure. Full coordinate models were used, except in the case of the IgE-Fc study where only α-carbon atoms were used to reduce the computational overhead of the large number of structures used in that analysis. If carbohydrate was present, the oligosaccharide chains were represented by a suitable structure adapted from the Brookhaven database [8] and added to Asn residues on the protein surface. For the analyses of single multidomain proteins, the domains were constrained in their relative positions by reasonable stereochemical links between their known structures (Fig. 2a,b and c). For the analyses of oligomers, symmetry constraints were used to define the location of the monomeric subunits (Fig. 2d and e).

The atomic coordinates of each glycoprotein model were converted to spheres (Table 1). The full coordinates were contained in a three-dimensional grid of cubes of side about 0.6 nm, this value being much less than the resolution $2\pi/Q_{max}$ of the scattering curves (2.7 nm for $Q_{max} = 2.3$ nm^{-1}). A cube was included in the sphere model if it contained sufficient coordinates above a cut-off value defined such that the total volume of all the cubes included in the model was equal to the dry protein and carbohydrate volume calculated from the sequence [29]. If the protein contained more residues than observed in the crystal structure for reason of crystallographic disorder, or the number of residues is altered when a homologous structure is used, the cut-off value for cube generation was adjusted accordingly to attain the correct volume. During a search, it is usually necessary to fix the position of the origin of the grid in order to ensure consistency of the grid conversion of coordinates into cubes. The use of α-carbon coordinates instead of the full coordinates for grid conversion is not preferred as the absence of the amino acid sidechains will influence the conversion, even though this should be compensated by the use of the full dry volume.

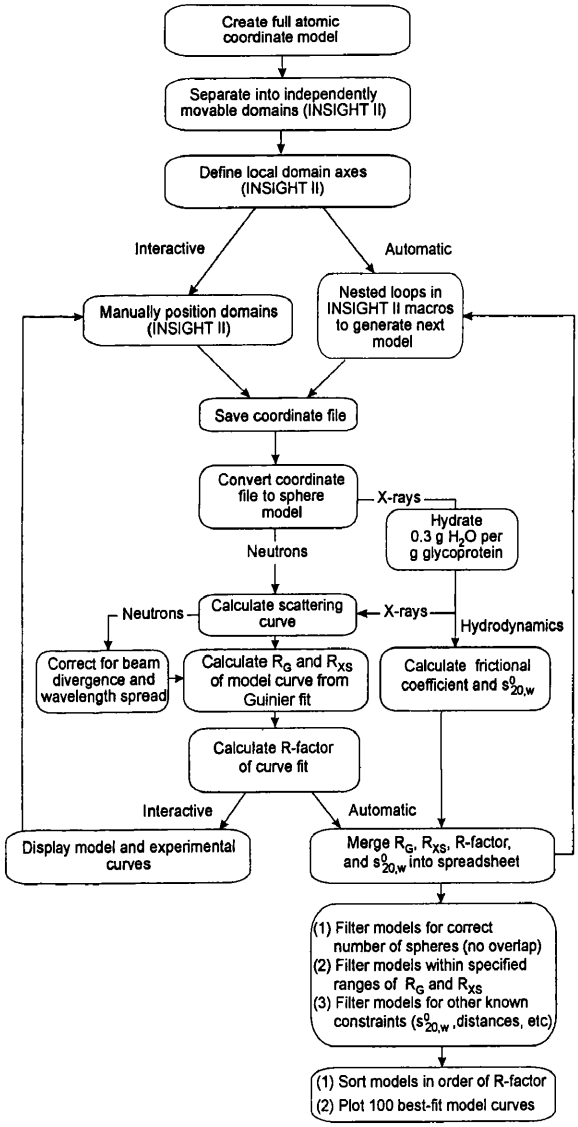


Fig. 1. Flow chart of two procedures for the initial manual and final automated analysis of multidomain models for scattering curve fits. Each box describes a stage in the two procedures, and further boxes show how additional information is included to evaluate the models. The automation of both procedures utilises INSIGHT II and Unix executable script files on Silicon Graphics workstations. The resulting parameters are filtered and sorted using Excel spreadsheets.

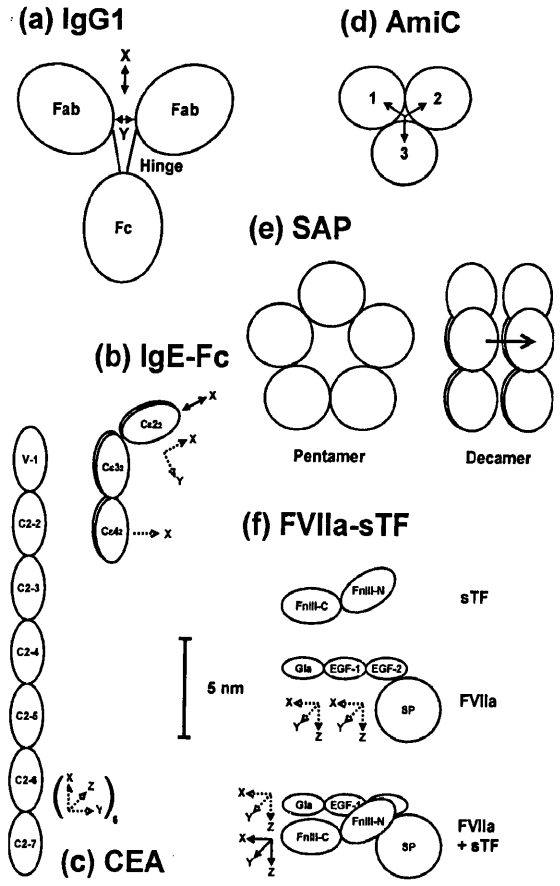


Fig. 2. Schematic outlines of six multidomain or oligomer structures to show how domain or subunit translations and rotations were implemented during the curve fit analyses. Translations are denoted by solid arrows, and rotations by dashed arrows: (a) For IgG1 and IgG2, the pair of Fab fragments were moved together in two-parameter translational searches along the X- and Y-axes relative to the Fc fragment; (b) For IgE-Fc, the $C\varepsilon2_2$ domain pair were translated along the X-axis twice, and rotated about the X- and Y-axes relative to the $C\varepsilon3$ and $C\varepsilon4$ domains. A further X-axis rotation involving the $C\varepsilon4_2$ domain pair resulted in a five-parameter search; (c) CEA models were evaluated using a three-parameter search in which the separation between the seven domains was fixed, and the domains were reorientated by the same X-, Y- and Z-axis angular increments applied to the six interdomain connections; (d) The formation of AmiC trimers was analysed using a one-parameter translational search of three AmiC monomers about a 3-fold axis of symmetry; (e) The formation of a SAP decamer from two pentamers was analysed using a one-parameter translational search of one pentamer relative to the other; (f) FVIIa was studied using a six-parameter search based on rotational movements of the single Gla and EGF-1 domains relative to the fixed EGF-2/SP domain pair. The complex between FVIIa and sTF was studied using a six-parameter translation and rotation of sTF relative to FVIIa.

The dry models do not have a hydration shell and are used for neutron curve modelling as neutron scattering observes unhydrated glycoprotein structures [10,27,28]. X-ray curve modelling requires hydrated structures and the dry volume was increased to allow for a hydration shell. This shell is well-represented by 0.3 g of water/g glycoprotein and an electrostricted volume of 0.0245 nm^3 per bound water molecule and corresponds to a water monolayer surrounding the protein surface [29], the volume of a free water molecule being 0.0299 nm^3. The simplest way to hydrate the cube models is to increase the length of the cube side to match the volume increase. This procedure is satisfactory for globular proteins of compact structure. However this will significantly distort the macromolecular structure if this contains a void space at its centre. In the case of the serum amyloid P component, an alternative algorithm HYPRO [10] was written to add a layer of hydration spheres evenly over the protein surface. Additional cubes were added in an uniform adjustable layer to the surface of the model in order to reach the required hydrated volume.

The Debye scattering curve simulations were based on overlapping spheres placed at the centre of each cube in the model, with the volume of each sphere set to be that of each cube. Scattering curves were calculated from the spheres for comparison with experimental data. No instrumental corrections to the calculated curves were applied for X-ray wavelength spread or beam divergence as synchrotron X-ray cameras utilise a pin-hole configuration that do not lead to geometrical distortion of the beam. Neutron cameras such as LOQ also use pin-hole geometries. However, as their dimensions are larger than X-ray cameras and longer wavelengths are used in order to maximise the available neutron flux, instrumental corrections are required. For D11 and D17, we often employed a Gaussian function based on a 16% wavelength spread $\Delta\lambda/\lambda$ (full-width-half-maximum) at λ of 1.0 or 1.1 nm and a beam divergence $\Delta\theta$ of 0.016 radians as an empirical correction. The theoretical values of $\Delta\lambda/\lambda$ for D11 and D17 are respectively, 8 and 10%, while that for $\Delta\theta$ depends on both the beam aperture (0.7×1.0 cm^2) and the detector cells (1 cm^2) and is arourd 0.01 radians. A reevaluation of $\Delta\lambda/\lambda$ for D17 data for serum amyloid P component gave 10% in good agreement with theory, although $\Delta\theta$ was larger at 0.024 radians [10]. The neutron fits deteriorate at large Q and this may indicate a small residual flat background that arises from incoherent scatter from the protons in the protein. The wavelength range of 0.2–1.0 nm used simultaneously on LOQ

(where time-of-flight techniques provide the necessary monochromatisation) complicates the beam corrections, however the use of a Gaussian function as for D17 data (10% for $\Delta\lambda/\lambda$ for a putative λ of 0.6 nm and 0.016 radians for $\Delta\theta$) gives reasonable curve fits [10].

Once trial curve fits indicated that analysis was possible, detailed model searches were run for several days, typically using a Silicon Graphics INDY R4400SC Workstation with 64 Mb of memory and a 4 Gb hard disk. Nested loops within INSIGHT II macro scripts (Fig. 1) are easily set up to generate hundreds or thousands of models based on two or more degrees of rotational and/or translational freedom between the domains or subunits in question. Each model was converted into spheres. An X-ray or neutron scattering curve was calculated from each model. The R_G and R_{XS} values were determined from the calculated curves in the same Q ranges used for Guinier fits of the experimental data. Three generous filters were used to remove unsatisfactory models: (1) The creation of models can result in physically unreasonable steric overlap between the subunits, accordingly the number of spheres in each model was compared to that expected from the dry volume calculated from the composition and the model was retained if the total was within 95% of that expected; (2) Next, models were retained if the modelled R_G and R_{XS} values were within 5% or $\pm$ 0.3 nm from the experimental values; and (3) Models were then assessed using a goodness-of-fit R-factor $= 100^*\Sigma|I(Q)_{exp}-I(Q)_{cal})|/\Sigma|I(Q)_{exp}|$ which was computed by analogy with the R-factor used in crystallography [7,27]. Note that the R-factor will depend on the Q range in use and the number of data points in that Q range and should be normalised against $I(Q)_{cal}$ for a given curve fitting exercise. For the purpose of automating the curve fit procedure, the R-factor was initially used in the low Q range out to 0.5 nm^{-1} in order to determine the scaling factor to match the experimental and calculated $I(Q)$ curves. Note that this is the Q range used for R_G and R_{XS} determinations. To define a working scale for curve comparisons, $I(0)_{cal}$ was arbitrarily set as 1000. The quality of the curve fits from each model in the search was then determined by computing the R-factor for successive Q ranges out to 0.8–2.0 nm^{-1} in 0.2 nm^{-1} steps (denoted $R_{0.8}$–$R_{2.0}$). While R-factors are not comparable between different curve fitting exercises and are primarily influenced by the large $I(Q)$ values at low Q, they provide a useful filter of models. A full list is prepared of each model, the geometrical steps used to define it, the number of spheres in it, its R_G and R_{XS} values and its $R_{0.8}$–$R_{2.0}$ values. The list is imported into a PC-based spreadsheet, which is used to set the cut-off filters, sort the models in order of their R-factors and identify the best curve fits for printing.

These procedures can also be used to calculate sedimentation coefficients from analytical ultracentrifugation experiments (Fig. 1). The same hydrated sphere models used for X-ray fits are used for this, even though the computing requirement becomes considerable. The comparison of calculated and experimental sedimentation coefficients provides further support for the scattering analysis.

3. Results and discussion

3.1. The bovine immunoglobulin subclasses IgG1 and IgG2

The bovine IgG isotypes IgG1 and IgG2 exhibit large differences in effector functions, where only IgG1 is selectively transported from blood plasma and ultimately into milk by specific cell receptors. IgG contains 12 immunoglobulin fold domains arranged within two Fab and one Fc fragment [30]. The four-domain Fab fragment of IgG recognises a vast array of antigens (foreign molecules), while receptor sites are located in the four-domain Fc fragment (Fig. 2a). Consequently there is much interest in the relative separation of the Fab and Fc fragments, yet there are known difficulties in crystallising and determining the structure of an intact antibody for reason of domain flexibility. The two Fab fragments and one Fc fragment in IgG1 and IgG2 are linked by a disulphide-linked polypeptide hinge at the centre. Receptor specificity for bovine IgG1 and not for IgG2 could result: (a) from a binding site present at the hinge region-Fc junction in IgG1 that is absent in IgG2; (b) from different hinge conformations in IgG1 and IgG2; or (c) from steric obstruction of the Fc site by the Fab fragments in IgG2. Earlier sequencing and structure prediction studies on bovine and ovine IgG1 and IgG2 showed that IgG2 had a seven-residue deletion in the hinge sequence, with the loss of the determinant motif for the receptor. Accordingly, IgG2 was predicted to have a short

hinge and steric hindrance of effector function was considered to be likely. Solution scattering provided a means to clarify the structural significance of these sequence differences between the two isotypes.

Neutron scattering on LOQ was used to study IgG1 and IgG2 [6]. Interestingly, the radii of gyration R_G were found to be similar at 5.64 and 5.71 nm for IgG1 and IgG2 respectively, in 100% ^2H_2O buffers. The two cross-sectional radii of gyration R_{XS} were also similar at 2.38–2.41 nm and 0.98–1.02 nm. It was concluded that both bovine IgG1 and IgG2 possessed similar overall solution structures, despite these sequence differences at the centre of their structures.

The availability of homologous crystal structures for the Fab and Fc fragments permitted an automated scanning search of possible IgG1 and IgG2 structures. Coordinates for the two Fab fragments were displaced in 0.25 nm steps in a two-dimensional $X-Y$ plane corresponding to the major plane of the Fc fragment (Fig. 2a). The hinge was omitted from these searches as this is small and not directly detectable by scattering. The use of stepwise $X-Y$ searches involving up to 200 planar arrangements of Fab and Fc fragments showed that the full IgG scattering curve in the Q ranges that correspond to the R_G and R_{XS} values were sensitive to the relative positions of the Fab and Fc fragments within IgG. In one search based on 56 models, four similar models were found to be consistent with the IgG1 and IgG2 scattering curves. In these models, the separation of the Fab C-terminus and Fc N-terminus α-carbon atoms ranged from 3.6 to 2.9 nm and the R-factor was determined to be 1.2% in the Q range of 0.09–1.55 nm^{-1} (Fig. 4a). Having optimised the location of the three fragments, the modelling analysis was completed by adding the hinge (Fig. 3a). A moderately extended hinge accounted for the solution structures of bovine IgG1 and IgG2. Energy refinements showed that the separation between the Fab and Fc fragments was stereochemically consistent with the different polypeptide length of the hinge in IgG1 and IgG2. The longer hinge in IgG1 appears to be present in a more coiled conformation than the shorter hinge in IgG2. In conclusion, the experimental data and their modelling supported hypothesis (a) in which sequence deletions in the hinge of IgG2 is the likely cause of the exclusion of this isotype from the transport process into milk.

3.2. The IgE-Fc fragment of immunoglobulin E

The plasma protein immunoglobulin E (IgE) is central for the immune response to foreign antigenic material and the development of an allergic, inflammatory response [31]. IgE contains 14 immunoglobulin fold domains, in which there is an additional pair of domains $(C\varepsilon2)_2$ in the Fc region (Fig. 2b) in place of the hinge in IgG. The interaction between IgE and its high affinity receptor FcεRI is central to allergic disease and involves the IgE-Fc fragment. While no crystal structure is known for the six-domain Fc fragment of IgE (IgE-Fc), it is possible to construct homology models for the four domains $(C\varepsilon3)_2$ and $(C\varepsilon4)_2$ by molecular graphics using the crystal structure of the corresponding four domains in IgG-Fc. The solution structure of the $(C\varepsilon2)_2$ domain pair relative to those of the $(C\varepsilon3)_2$ and $(C\varepsilon4)_2$ domain pairs is of great interest for understanding IgE-receptor interactions, so IgE-Fc was studied by X-ray and neutron scattering [7]. The upper limit on the R_G values was determined to be 3.52 ± 0.14 nm (X-rays) and 3.53 ± 0.05 nm (neutrons). The X-ray and neutron R_{XS} values were 1.89 ± 0.05 and 1.56 ± 0.09 nm, respectively. An upper limit on the maximum length of IgE-Fc was determined as 13 nm by both X-rays and neutrons.

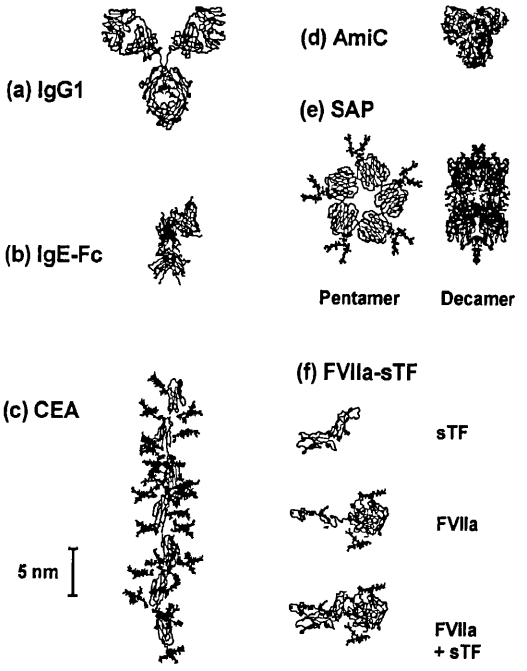


Fig. 3. The best-fit model from each curve fitting analysis to follow that of Fig. 2. The protein structure is denoted by an α-carbon trace, while oligosaccharides are shown in full if present.

The modelling of the IgE-Fc X-ray curves proved to be more complex than anticipated on the basis of the bovine IgG1 and IgG2 study. First, two available homology models for IgE-Fc in the Brookhaven database (codes 1ige and 2ige) that were based on alternative disulphide bridge connection schemes between the two heavy chains both gave poor agreement with experimental data. This was attributed to the unrefined position of the $(C\varepsilon2)_2$ domains in both models. Accordingly the IgE-Fc model with the correct disulphide bridging (2ige) was separated into four independent fragments, namely the $(C\varepsilon2)_2$ pair, the two $(C\varepsilon3)$ domains and the $(C\varepsilon4)_2$ pair. Trials were carried out in which X- and Y-axis rotations of the $(C\varepsilon2)_2$ pair relative to the remaining four domains were performed, together with two types of X-axis translation of the $(C\varepsilon2)_2$ pair to maintain domain connectivity (Fig. 2b). Even though this search covered all possible orientations and separations of the $(C\varepsilon2)_2$ pair, it also failed to give a good curve fit in the Q range of 0.5–1.0 nm^{-1}. Finally, starting from the best model from this $(C\varepsilon2)_2$ search, it was found that small rotations of the two $C\varepsilon3$ domains or large rotations of the $(C\varepsilon4)_2$ pair resulted in much improved curve fits.

The trial modelling of IgE-Fc enabled an automated five-parameter search to be initiated that applied rotations and translations to the six domains in order to fit the scattering data. Two different structures that differed slightly in the location of the $C\varepsilon3$ domains were used. The automated searches involved mainly movements in the $(C\varepsilon2)_2$ domains and the testing of over 37 000 models. Atypically, only the α-carbon coordinates were employed in the models to save computing time. The steric overlap filter eliminated 65% of the models if they contained less than 360 spheres as the result of the domains moving into each other prior to the grid transformation, 371 spheres being optimal. The use of further filters based on the R_G and R_{XS} values and the $R_{1.0}$ and $R_{2.0}$ values was examined. The $R_{2.0}$ values were more effective than the R_G values for selecting the best curve-fits. One reason is that $R_{2.0}$ monitored a larger Q range of $I(Q)$ intensities than R_G, which was advantageous when trace amounts of aggregates at the lowest Q values due to radiation damage caused slight increases in the R_G data (Table 1). Another advantage of $R_{2.0}$ is that a good curve fit corresponds to the lowest $R_{2.0}$ value obtained, while the modelled R_G value can be larger or smaller than the experimental R_G value so is less unequivocal as a filter. The disadvantage of $R_{2.0}$ is that it is not presented as an absolute value, so strictly its comparative usage is restricted to a single experimental curve.

The best fit model was defined as the mean structure of the 100 models with the smallest $R_{1.4}$ values. In this way, a bent IgE-Fc model with a $C\varepsilon2$ Y-axis rotation of 70° and an unchanged $C\varepsilon4$ X-axis rotation of 0° (Fig. 3b) was determined to give an excellent X-ray curve fit (Fig. 4b). The X-ray $R_{2.0}$ value was 3.4%, while the R_G value was 3.22 nm which is slightly less than the experimental X-ray value of 3.52 ± 0.12 nm for reason of trace aggregates. Comparison with the neutron curve gave a neutron $R_{1.5}$ value of 6.3%. The X-ray R_{XS} value of 1.93 nm agreed with the observed value of 1.89 ± 0.05 nm. Contour maps of $R_{2.0}$ values showed that a single best-fit minimum had been located by the searches and the maps enabled the experimental precision of the final model to be estimated. Using only those models for which $R_{2.0}$ was less than 4%, the precision of the IgE-Fc model was estimated to be between 40° and 90° for the $C\varepsilon2$–$C\varepsilon3$ bend angle and $\pm 50°$ for the $C\varepsilon3$–$C\varepsilon4$ bend angle. In conclusion, the modelling showed that IgE-Fc must adopt a bent structure at either the $C\varepsilon2$–$C\varepsilon3$ or $C\varepsilon3$–$C\varepsilon4$ junctions or at both if the observed scattering curve is to be rationalised in terms of atomic structures for the six domains within IgE-Fc. Planar or linear IgE-Fc domain structures do not fit the scattering data. The significance of this Fc structure is that it confirmed the bent structure previously proposed for intact human IgE by fluorescent labelling studies and showed how this bent structure can be formed. It also clarified how the domain structure of IgE-Fc can interact with its FcεRI receptor and opens the way for the scattering modelling of intact IgE which is in progress.

3.3. Carcinoembryonic antigen

Carcinoembryonic antigen (CEA) is one of the most widely-used cell-surface markers for tumour monitoring and for targeting by antibodies in cancer therapy [32]. It belongs to the same immunoglobulin superfamily as IgG and IgE, but is different in that it exists as a monomer of one V-type and six C2-type Ig domains, in contrast to the dimeric IgG structure that contains four V-type and eight C1-type Ig domains. Unlike IgG,

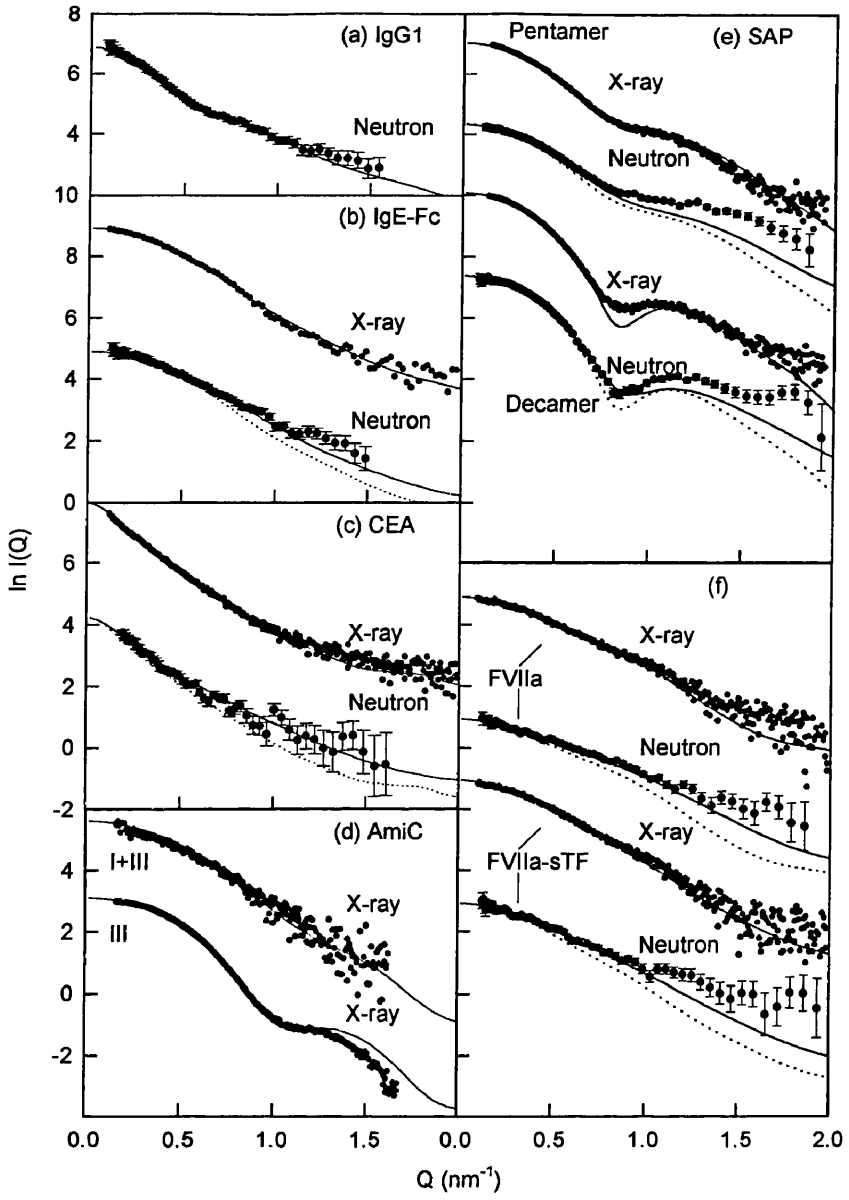


Fig. 4. Final *X*-ray and neutron curve fits based on the best-fit models from Fig. 3. The *X*-ray data were obtained from Stations 8.2 for IgE-Fc and CEA, and from Station 2.1 for AmiC, SAP, FVIIa and the FVIIa-sTF complex. Neutron data using 100% ^2H_2O buffer systems were obtained from LOQ. The continuous lines correspond to the curve calculated from the best-fit model in each case. Neutron beam smearing corrections were applied to the calculated curve prior to comparisons with the data. The dashed lines attached to the neutron curves indicate how the *X*-ray curve is different as the result of hydration and smearing corrections. Statistical error bars are shown when these are large enough to be seen.

CEA is heavily glycosylated with 28 oligosaccharide chains that comprise 50% carbohydrate by weight of CEA. An atomic structure for CEA would clarify its functional role and the optimal design of antibodies that will react with CEA. For reason of its glycosylation and interdomain flexibility, it is most unlikely that CEA could be crystallised intact. As CEA is readily cleaved from membranes and as two-domain crystal structures for two homologous cell surface proteins CD2 and CD4 were available for modelling, this opened the way for a detailed scattering study of CEA [8]. The scattering data collection showed from Guinier analyses that the X-ray R_G of CEA was 8.0 ± 0.6 nm. The X-ray R_{XS} was high at 2.1 ± 0.2 nm and is consistent with carbohydrate structures in CEA that are extended away from the protein surface. Combination of the R_G and R_{XS} values showed that CEA is of length 27–33 nm. As each domain in CD2 and CD4 is about 4 nm long, CEA is seen to possess an extended arrangement of seven domains in solution. The neutron $I(0)/c$ value from Guinier analysis resulted in a molecular weight of 150 000. In combination with a value

of 152 500 calculated from its composition, this showed that CEA was monomeric.

The creation of a starting model for the automated curve fit analysis of CEA was based on the two-domain CD2 crystal structure. CD2 showed greater sequence similarity with the CEA domains than CD4 and the linker peptide connecting the V- and C2-type CD2 domains was similar in length to those in CEA. Accordingly the CD2 domains were separated, the C2-type domain was duplicated five times and the seven domains were arranged in a straight line. Given the known carbohydrate composition of CEA, over 50 oligosaccharide structures present in the Brookhaven database were analysed to show that a consensus oligosaccharide structure could be created using that found in the Fc fragment of IgG (chain A in 1fc1). A total of 28 oligosaccharide chains in extended conformations were positioned at the glycosylation sites in CEA.

The objective of the automated search was to identify a general CEA structure that best represented its solution structure. A computationally-prohibitive number of structures would be generated if all six interdomain interfaces were independently varied, which is not justified by the structural resolution of solution scattering. Accordingly the search was simplified by setting all six X-, Y- and Z-axis rotation angles between the CEA domains to be the same in each model and also the interdomain separation was fixed to be that in CD2 (Fig. 2c). A three-parameter search based on 15° rotational steps about the X-, Y- and Z-axes generated 4056 models, which could be grouped into four families of structures, namely linear, curved, zig-zag and helical. A curve-fit search based on a single electron density model for CEA showed that the zig-zag family gave reasonable curve fits to the X-ray data, but worsened ones to the neutron data. Two-density CEA models were therefore used in a second search showed that the zig-zag model fitted well to both the X-ray and neutron curves. In the two-density models, the protein and carbohydrate spheres were assigned weights of two and three respectively, for X-ray fits and one and one for neutron fits, based on calculation of the electron and nuclear scattering densities [29]. From this search, the 100 best-fit CEA models had a mean X-axis rotation of $160° \pm 25°$, Y-axis rotation of $10° \pm 30°$ and Z-axis rotation of $-5° \pm 35°$ and gave a mean R_G of 7.8 ± 0.2 nm and R_{XS} of 2.02 ± 0.06 nm that were within error of the experimental values. The two-density zig-zag CEA model in Fig. 3c and Fig. 4c ($X = 165°$, $Y = 30°$, $Z = 15°$) gave an $R_{2.0}$ of 4.7%. Note that the X-rotation is close to 180° and corresponds to the reversal in orientation of neighbouring domain faces along the long axis of CEA, while the other two rotations are close to 0° and correspond to slight bends along the long axis of CEA. It was also noteworthy that the independent use of the two-domain CD2 crystal structure to generate a seven-domain CEA model by successive domain superimpositions to retain the orientation between the two CD2 domains also resulted in a good curve fit. Interestingly this CD2-like CEA model had X-, Y- and Z-rotations that were similar to those of the 100 best-fit search models that were filtered from 4056 models.

The biological significance of the CEA model was best determined by molecular graphics, since the scattering fits only show that CEA is extended and monomeric and that it can be modelled from known structures. The C2-type immunoglobulin fold is a simple β-sandwich structure which is formed from two β-sheets EBA and GFCC′ that form two opposite sides of the fold. Since X is close to 180° in the CEA models, this implies that the EBA and GFCC′ β-sheets of adjacent domains alternate with each other along one side of the long CEA structure. Further inspection of the model showed that the GFCC′ β-sheets contain little or no carbohydrate, which is suggestive that they are possible protein ligand sites. This clarifies how highly extended CEA molecules on adjacent cell surfaces might reach out and form adhesive interactions with each other through the matching of opposing GFCC′ faces, as well as suggesting how anti-CEA antibodies might be rationally targeted to bind to exposed protein surfaces on CEA at its GFCC′ faces. The joint study of CEA by scattering and molecular graphics is a good example of how function can be understood by this approach.

3.4. Monomeric and trimeric forms of AmiC

AmiC plays a key role in amide metabolism in the cytosol of *Pseudomonas aeruginosa*, a pathogenic bacterium involved in opportunistic infections [33]. Despite its occurrence in the cytosol, AmiC is a member of a large superfamily of two-domain periplasmic binding proteins [34]. In accordance with this, the crystal structure of

AmiC-acetamide shows that acetamide is bound at the bottom of a closed cleft formed between the two domains. Other crystal structures in this superfamily show that this cleft is significantly closed in the liganded form, but is opened when the ligand is removed and this conformational change is detectable by scattering [9]. X-ray and neutron scattering was performed to investigate this change for AmiC. Unexpectedly AmiC was found to exists as a monomer–trimer equilibrium at concentrations between 0.4 and 16.4 mg AmiC/ml. The R_G and M_r varied with the AmiC concentration and the position of the equilibrium depended on whether acetamide or the anti-inducer butyramide was present. The R_G data for trimeric AmiC were the same for AmiC bound to acetamide or butyramide, i.e. no conformational changes were seen. These results were surprising because other members of this superfamily are monomeric in solution and because AmiC-acetamide formed an antiparallel dimer in its crystal structure that might have existed in solution. It would appear that the dimer is an artefact of crystallisation.

Using the crystal structure of monomeric AmiC-acetamide, modelling searches were performed to validate the interpretation of the AmiC scattering curves in terms of oligomer formation [9]. To simplify these, advantage was taken of the constraint that a trimeric structure would possess a three-fold axis of symmetry (Fig. 2d). Trimers were formed by arranging the long axes of three monomers parallel to each other and positioning the monomers about a three-fold axis of symmetry with their ligand-binding clefts arbitrarily set to face outwards (Fig. 4d). The centres of the three monomers in the starting model were coincident on the central three-fold axis of symmetry, so were sterically overlapped. Translations generated 21 homotrimer models by moving the monomers outwards from this central axis in 0.2 nm steps for 4 nm. The best fit from this search using data for trimeric AmiC-butyramide at high concentration had an R-factor of 4.7% (curve III in Fig. 4d). The weighted sum of the scattering curves for 40% monomer and 60% trimer gave good curve fits to AmiC-butyramide at low concentration (curves I + III in Fig. 4d). From this fit, an association constant of 2×10^{10} M^{-2} could be estimated from the ratio of monomer and trimer. The success of these fits confirmed the presence of a monomer–trimer equilibrium.

Other automated curve fits were performed to assess alternative models. For example, it might be that the experimental curves arise from a mixture of the crystallographic monomer, dimer, trimer and tetramer in solution. Calculation from these four individual structures gave poor curve fits with $R_{2.0}$ between 9.7 and 39.3%. A search of 176 851 combinations of these four scattering curves showed that a mixture of 51% dimer and 49% tetramer was optimal, but this gave a high $R_{2.0}$ value of 6.3% and the curve fit deviated at Q values above 0.8 nm^{-1}, thus ruling out this model. Another example was based on the premise that the trimer might be formed from an asymmetric combination of the crystallographic monomer and dimer. The translation of the monomer relative to the dimer in 0.2 nm steps created 39 041 trimer models. After filtering for overlap and R_G values, the best-fit X-ray $R_{2.0}$ value was 3.9–4.1%, which is better than that of the symmetric AmiC trimer model. Even though this model is ruled out on symmetry grounds, it was interesting that a better fit was obtained starting from incorrect assumptions, as this showed the importance of using a correctly defined starting model in automated searches.

3.5. Pentamer and decamer formation in the serum amyloid P component

The serum amyloid P component (SAP) is a plasma glycoprotein composed of identical subunits that are non-covalently associated as a flat disc-like pentamer with 5-fold cyclic symmetry [35]. SAP binds to all forms of amyloid fibril in vitro and protects them from proteolysis and is universally present in amyloid deposits. SAP also binds to sulphated glycosaminoglycans, DNA and chromatin and is a calcium-dependent lectin. The crystal structure of the pentamer shows that each subunit contains two antiparallel β-sheets and two α-helices. An N-linked oligosaccharide site is located at the outer edge of the α-helix A-face, which is on the opposite side to the calcium binding B-face. SAP forms very stable decamers in the absence of calcium. Since the decamer has maximal calcium-dependent ligand binding and is susceptible to proteolysis in the absence of calcium, the decamer is probably formed by the association of two A-faces. Solution scattering was performed in order to determine a structure for the SAP decamer which has not yet been crystallised, as

well as that for the oligosaccharides that were not visible in the pentamer crystal structure (Fig. 3e). Since the SAP ring is rigid, the SAP pentamer also provided a good opportunity to test the procedure for calculating scattering curves from a crystal structure.

X-ray and neutron data analysis on SAP pentamers and decamers showed that the decamer was formed by the association of the pentameric A-faces [10]. This result was obtained from molecular weight calculations based on the X-ray and neutron $I(0)/c$ values from Guinier analyses (Section 2). These consistently showed that the ratio of $I(0)/c$ values for the decamer and pentamer was not 2.0 as expected but was closer to 1.7. This was deduced to be the result of an altered absorption coefficient for the decamer compared to the pentamer which affected the determination of c. Inspection of the SAP crystal structure showed that four Trp residues per protomer were close to the A-face and this would bring into proximity a total of 40 Trp residues if the A-faces associated to form the decamer. This interpretation was confirmed by difference absorbance and fluorescence spectroscopy which showed that the Trp residues in SAP were significantly perturbed upon dissociation of the decamers into pentamers.

The aim of the automated curve modelling for SAP was to distinguish between the possible A–A and B–B structures for the decamer. Firstly, curve modelling from the SAP pentamer coordinates confirmed the fit procedure for the X-ray and neutron data (Section 2) and showed that a good fit was obtained with extended oligosaccharide structures of the type used above for CEA (Fig. 3e and Fig. 4e). Next, based on the coordinates of this pentamer model, the decamer was modelled using symmetry constraints to reduce the number of models to be tested. Two pentamers were superimposed on a common central 5-fold axis of symmetry, then one was turned by 180° to reverse the orientation of its A- and B-faces. To generate both symmetric forms of SAP, the pentamers were either directly aligned with each other, or one was rotated by 36° relative to the other about the 5-fold axis of symmetry. The search was performed by separating the pentamer centres by 4 nm, then translating one pentamer completely through the other pentamer without regard for steric overlap in 0.1 nm steps by 8 nm along the central axis (Fig. 2e). The 80 models included the two possible A–A and B–B structures and the degree of steric overlap, R_G values and R-factors were assessed for all 80 models. As expected, two minima were found that corresponded to the A–A and B–B structures, both of which had very similar R-factor values. At the A–A minimum, the R_G value was 4.23 nm which agreed with the experimental value of 4.23 ± 0.12 nm and gave a satisfactory X-ray curve fit in Fig. 4e, while the B–B minimum corresponded to a slightly larger R_G value of 4.32 nm. The separation between the two SAP pentamers was 3.3 nm which is consistent with the 3.6 nm thickness of the SAP disk if the two pentamers were rotated by 36° relative to each other to improve the steric contacts between them. While the difference between the two structures is not large, the A–A structure was favoured over the B–B structure.

Solution scattering provides an unambiguous means of distinguishing between SAP pentamers and decamers (Fig. 4e), which is not straightforward by other methods [35]. The $I(0)/c$ values permitted the decamer to be identified as an A–A structure. The curve modelling indicated extended oligosaccharide structures. The most favoured model for the decamer in which the two pentamers are rotated by 36° relative to each other is interesting in that the oligosaccharides from opposite pentamers are in proximity to each other and may interact with each other. While the contribution of SAP glycosylation to the stabilisation of A–A decamers is not clear at present, the scattering study has provided key insights that complement the atomic detail revealed by the crystal structure, as well as providing a stimulus for further experiments to explore SAP function.

3.6. The heterocomplex between tissue factor and factor VIIa

Exposure of the membrane-bound receptor, tissue factor, to plasma initiates the blood coagulation pathways in which tissue factor forms a very stable catalytic enzyme–cofactor complex with the serine protease factor VIIa (FVIIa) [36,37]. Soluble tissue factor (sTF) contains two fibronectin type III domains, which are similar to Ig folds. FVIIa contains four domains, namely a Gla domain, two epidermal growth factor domains and one serine protease domain (Fig. 2e). In the absence of a crystal structure for the complex, Guinier analyses showed how the complex was formed. The mean X-ray and neutron scattering

R_G values were 3.25, 2.13 and 3.14 nm (±0.13 nm) for FVIIa, sTF and their complex, in that order. The mean R_{XS} values were 1.33, 0.56 and 1.42 nm (±0.13 nm), in that order. The mean lengths L from $P(r)$ analyses were 10.3, 7.7 and 10.2 nm, in that order. In combination with the dimensions of domains that are known homologues to those in FVIIa and the crystal structure of sTF, it was readily inferred from these data that in solution both unbound proteins have extended domain structures and that the complex is formed by the compact side-by-side alignment of the two proteins along their long axes [11]. The high binding affinity of sTF for FVIIa could therefore be explained by the occurrence of many intermolecular contacts in the complex. This analysis was confirmed by the subsequent crystal structure of the complex between active-site inhibited FVIIa and proteolytically-cleaved sTF [36,37].

The FVIIa–sTF crystal structure raised further questions about the structure of free FVIIa. In the complex, FVIIa was observed as a extended linear conformation and this differed significantly from the shorter bent four-domain arrangement seen in the structural homologue factor IXa. Calculations based on the scattering curve for free FVIIa in solution showed that the crystal structure of FVIIa in the complex was essentially unchanged in conformation in the absence of sTF. Good curve fits were obtained with an oligosaccharide conformation that was less extended into solution than those in CEA and SAP (compare Fig. 3c and Fig. 3e with Fig. 3f). Poorer curve fits were obtained for the crystal structure of factor IXa (not shown). An automated search for domain conformations of the Gla and EGF-1 domains relative to the EGF-2/SP domain pair was also performed to assess these results more generally. In the starting FVIIa model, the Gla and EGF-1 domains were arranged with their N- and C-terminal α-carbon atoms on the same linear axis as that of the EGF-2 domain and separated by 0.5 nm. Two extended N-linked oligosaccharides were added to the SP domain. A six-parameter search could be performed based on two sets of X-, Y- and Z-axis rotations in steps of 72° (Fig. 2f). The successive filtering of 15 625 models for steric overlap and R_G values greater than 3.21 nm left only 317 models [12]. The search showed that only the most extended FVIIa structures gave good curve fits. The importance of this result is to show that free FVIIa exists as a preformed template that is ideal for rapid strong interaction with tissue factor at the onset of coagulation.

Curve calculations also showed that the crystal structure of the sTF-FVIIa complex was consistent with its solution scattering curve (Fig. 3f and Fig. 4f). This reassurance is useful, given the hypothetical possibility of domain rearrangements between the crystal and solution states. In the absence of a crystal structure for a multidomain heterodimeric protein complex, an automated curve-fit search would have been applied. The feasibility of this was examined for the FVIIa–sTF complex. Two major differences from the analyses of oligomeric AmiC and SAP complexes are the need to assume the absence of major conformational change in either component on complex formation and the absence of symmetry constraints to simplify the searches. While little can be done in relation to the former, it is possible to simplify the searches by the use of known biochemical constraints. Such constraints were available from the known alignment of sTF relative to the Gla and EGF domains in FVIIa, the identification from mutants of sTF residues known to interact with FVIIa and the location of the three N-linked oligosaccharide sites in sTF which are known not to interact with FVIIa. After modelling trials based on three translational and three rotational axes (Fig. 2e), full searches were based on translating four orientations of sTF in 0.5 nm steps along three axes to generate 4×9261 models. Interestingly, after filtering based on steric overlap and R_G values and applying the biochemical constraints, it was possible to locate sTF within the large interface between the SP domain and the Gla/EGF domains, much as observed in the crystal structure of the complex. While the use of these constraints improved the analyses, this still left a range of compact structures for the complex. Nonetheless the calculations showed that these calculations are potentially of value for the study of heterodimeric complexes.

4. Conclusions

The diverse range of applications of solution scattering for the study of molecular structures indicate the power of this method, once the availability of relevant crystal structures is exploited in full. Examples in this review are summarised in Table 1, together with the key

parameters defining each example and include three single multidomain proteins as well as three different complexes. In the specific case that a crystal structure is available, quantitative comparisons can be made to verify its overall structure in solution and to deal with other questions such as the conformation of oligosaccharides on the protein surface. More generally, the curve fit analyses for the multidomain proteins IgG, IgE-Fc and CEA illustrate how an automated method for constrained modelling based on known homologous structures and the fixed connections between these structures can be easily set up. A limited family of good curve fits is filtered from a large number of possible models and indicate molecular structures that correspond to the scattering data. The biological significance of these studies corresponds to low resolution structural questions in relation to the location of known active sites or key residues in the individual domains. Thus the IgG and IgE-Fc studies indicated how accessible their domain structures were for interactions with receptors, while the CEA study showed how its structure could form homodimeric adhesive complexes between cells.

The biological significance of these structures depends on the precision of the modelling. It should be remembered that a good curve fit is only a test of consistency and will not constitute a unique structure determination, although the use of strong constraints will limit the inherent ambiguity of scattering. The advantage of automation is to remove the tedium of hand-fitted modelling fits and enables a comprehensive assessment of the constrained structures that fit a given curve to be made. The precision of the best fit models is readily estimated from the mean of the structures that gives curve fits within experimental error. We are often asked about the effect of macromolecular flexibility on this structural modelling. The curve fits necessarily produces a family of similar structures that may well be related by flexibility, but the analyses do limit what is allowed by flexibility. While the IgE-Fc structure was shown to be bent, this cannot be linear even if this was flexible. Likewise, while CEA and FVIIa exhibit highly extended structures, significantly bent structures are ruled out by the curve fits.

The extension of curve fit analyses to analyse protein–protein complexes is summarised for AmiC trimers, SAP decamers and the FVIIa–sTF complex. The modelling of protein–protein complexes was less straightforward for reason of the absence of covalent links between the different subunits to constrain the models. Nonetheless the AmiC and SAP analyses were successfully constrained by symmetry considerations based on their known crystal structures and this simplified the automated searches. The heterodimeric FVIIa–sTF complex was more difficult to analyse, however success is possible from the use of biochemical constraints during the curve fit modelling. All three modelling studies provided biologically useful information on the ligand-dependent trimer formation of AmiC, the orientation of SAP pentamers in the decamer and the mode of association of sTF with the FVIIa light chain in their complex.

All the modelling studies described here depend on the reliability of a procedure to calculate scattering curves from atomic coordinate models. That described in Section 2 is essentially based on a survey of electron and nuclear densities published in 1986 [29] and has worked well in all the calibration and modelling analyses since that time. The two major corrections for coordinate models before curves can be calculated are the need to add a hydration shell for the modelling of X-ray curves (and sedimentation coefficients) and to allow for possible large internal scattering density fluctuations in X-ray and neutron curve modelling. The hydration shell is relatively straightforward to add (Section 2) and corresponds to a monolayer of water molecules surrounding the macromolecule. Internal density fluctuations are more difficult to compute, where the electron and nuclear densities of carbohydrate are notably higher than those for protein. They also vary strongly between the 20 hydrophilic and hydrophobic amino acids, where hydrophilic residues have a higher scattering density than hydrophobic ones. The principle advantage of the joint neutron/X-ray approach is that the macromolecule is visualised in high negative and positive solute–solvent contrasts, respectively. This provides a simple experimental test to show whether internal density fluctuations are significant by comparisons of the X-ray and neutron curve fits. In this context, SAP was unique in that the distribution of hydrophilic and hydrophobic residues in its hollow ring structure removed the contrast dependence of the scattering curve from the experimental data. The opposite extreme was encountered with CEA, where the curve fits showed that a two-density modelling strategy was

unavoidable in order to take proper account of the 50% carbohydrate content in CEA [8]. By the same token, the occurrence of differential $^{1}H-^{2}H$ exchange at amide and hydroxyl groups within the protein may be thought to affect neutron modelling analyses[2]. Curve simulations based directly on atomic coordinates show that these effects are negligible.

The calculation of scattering curves from coordinates also requires allowance for the instrumental geometry. This is unimportant for synchrotron X-ray cameras. The magnitude of these corrections for neutron cameras is illustrated in Fig. 4. It is reasonably well characterised for Instruments D11 and D17 at the ILL [10], but may require reinvestigation for the new Instrument D22 at the ILL. It requires further development for LOQ at ISIS for reason of the very different time-of-flight method used to achieve monochromisation. The logarithmic plots of Fig. 4 show that most of the neutron curve fits deviate upward by small amounts at large Q. This may be the result of a small uniform residual background due to incoherent scatter from the proton content in the protein samples.

Acknowledgements

We thank the Biotechnology and Biological Sciences Research Council, the Engineering and Physical Sciences Research Council, the Wellcome Trust and the Clement Wheeler-Bennett Trust for grant support. We also thank our biochemical collaborators for their generous provision of samples, the instrument scientists at the SRS, ISIS and ILL for their support and A.S. Nealis, Dr K.F. Smith, Dr A.J. Beavil and M.O. Mayans for invaluable contributions.

References

[1] Perkins SJ. Biochem J 1988;254:313–27.
[2] Perkins SJ. New Comp Biochem 1988;11B:143–264.
[3] Perkins SJ. In: Jones C, Mulloy B, Thomas AH, editors. Physical Methods of Analysis in Methods in Molecular Biology. New Jersey: Humana Press, 1994;22:39–60.
[4] Glatter O, Kratky O, editors. Small-angle X-ray Scattering. New York: Academic Press, 1982.
[5] Perkins SJ, Nealis AS, Sutton BJ, Feinstein AJ. Mol Biol 1991;221:1345–66.
[6] Mayans MO, Coadwell WJ, Beale D, Symons DBA, Perkins SJ. Biochem J 1995;311:283–91.
[7] Beavil AJ, Young RJ, Sutton BJ, Perkins SJ. Biochemistry 1995;34:14449–61.
[8] Boehm MK, Mayans MO, Thornton JD, Begent RHJ, Keep PA, Perkins SJ. J Mol Biol 1996;259:718–36.
[9] Chamberlain D, O'Hara BP, Wilson SA, Pearl LH, Perkins SJ. Biochemistry 1997;36:8020–9.
[10] Ashton AW, Boehm MK, Gallimore JR, Pepys MB, Perkins SJ. J Mol Biol 1997;272:408–22.
[11] Ashton AW, Kemball-Cook G, Johnson DJD, Martin DMA, O'Brien DP, Tuddenham EDG, Perkins SJ. FEBS Lett 1995;374:141–6.
[12] Ashton AW, Boehm MK, Johnson DJD, Kemball-Cook G, Perkins SJ. 1997, unpublished results.
[13] Boulin C, Kempf R, Koch MHJ, McLaughlin SM. Nucl Instrum Meth 1986;A249:399–407.
[14] Towns-Andrews E, Berry A, Bordas J, Mant GR, Murray PK, Roberts K, Sumner I, Worgan JS, Lewis R, Gabriel A. Rev Sci Instrum 1989;60:2346–9.
[15] Worgan JS, Lewis R, Fore NS, Sumner IL, Berry A, Parker B, D'Annunzio F, Martin-Fernandez ML, Towns-Andrews E, Harries JE, Mant GR, Diakun GP, Bordas J. Nucl Instrum Methods Phys Res 1990;A291:447–54.
[16] Heenan RK, King SM. Proceedings of an International Seminar on Structural Investigations at Pulsed Neutron Sources, Dubna, 1–4 September, 1992. Report E3-93-65, 1993, Joint Institute for Nuclear Research, Dubna.
[17] Lindley P, May RP, Timmins PA. Physica B 1992;180:967–72.
[18] Kratky O. Progr Biophys Chem 1963;13:105–73.
[19] Jacrot B, Zaccai G. Biopolymers 1981;20:2413–26.
[20] Wignall GD, Bates FS. J Appl Crystallogr 1987;20:28–40.
[21] Pilz I. In: Leach SJ, editor. Physical Principles and Techniques of Protein Chemistry, Part C. New York: Academic Press, 1973:141–243.
[22] Hjelm RJ. J Appl Crystallogr 1985;18:452–60.
[23] Svergun DI, Semenyuk AV, Feigin LA. Acta Crystallogr 1988;A44:244–50.
[24] Semenyuk AV, Svergun DI. J Appl Crystallogr 1991;24:537–40.
[25] Svergun DI. J Appl Crystallogr 1992;25:495–503.
[26] Perkins SJ, Weiss H. J Mol Biol 1983;168:847–66.
[27] Smith KF, Harrison RA, Perkins SJ. Biochem J 1990;267:203–12.
[28] Perkins SJ, Smith KF, Kilpatrick JM, Volanakis JE, Sim RB. Biochem J 1993;295:87–99.
[29] Perkins SJ. Eur J Biochem 1986;157:169–80.
[30] Burton DR, Woof J. Adv Immunol 1992;51:1–84.
[31] Sutton BJ, Gould HJ. Nature (London) 1993;366:421–8.
[32] Thompson JA, Grunert F, Zimmerman W. J Clin Lab Anal 1991;5:344–66.
[33] Tam R, Saier MH Jr. Microbiol Rev 1993;57:320–46.
[34] Drew R, O'Hara B, Williams R. In: Nakazawa T, Furukawa K, Haas D, editors. Molecular Biology of Pseudomonads. Washington: ASM Press, 1996:331–41.
[35] Pepys MB, Booth DR, Hutchinson WL, Gallimore JR, Collins PM, Hohenester E. Amyloid Int J Exp Clin Invest, 1997;4:274–95.
[36] Banner DW, D'Arcy A, Chene C, Winkler FD, Guha A, Konigsverg WH, Nemerson Y, Kirchofer D. Nature (London) 1996;380:41–6.
[37] Bazan JF. Nature (London) 1996;380:21–1.

Stephen J. Perkins
Christopher G. Ullman
Nigel C. Brissett
Dean Chamberlain
Mark K. Boehm

Analogy and solution scattering modelling: new structural strategies for the multidomain proteins of complement, cartilage and the immunoglobulin superfamily

Authors' address
Stephen J. Perkins, Christopher G. Ullman, Nigel C. Brissett, Dean Chamberlain, Mark K. Boehm,
Protein Structure Group, Department of Biochemistry and Molecular Biology, Royal Free Hospital School of Medicine, London, UK.

Correspondence to:
Stephen J. Perkins
Protein Structure Group
Department of Biochemistry and Molecular Biology
Royal Free Hospital School of Medicine
Rowland Hill Street
London NW3 2PF
UK
Fax: 44 171 794 9645
e-mail: steve@rfhsm.ac.uk

Acknowledgements
This research was supported by the Wellcome Trust, the Biotechnology and Biological Sciences Research Council, the Engineering and Physical Sciences Research Council and the Clement Wheeler-Bennett Trust. We thank our biochemical collaborators for their generous provision of samples, and the instrument scientists at SRS, ISIS and ILL for invaluable support.

Immunological Reviews 1998
Vol. 163: 237–250
Printed in Denmark.
Immunological Reviews
ISSN 0105-2896

Summary: Many immunologically relevant proteins possess multidomain structures. Molecular structures both at the level of the individual domain and that of the intact protein are required for a full appreciation of function and control. Two recently developed structural approaches are reviewed here. Analogy modelling methods are based on the current understanding of many protein structures, and make possible the identification of folds for superfamilies of unknown structures. An integrated multidisciplinary predictive approach has been successfully applied to the von Willebrand factor type A, proteoglycan tandem repeat and factor I/membrane attack complex domains. The available experimental and predictive evidence is assembled in order to identify a known three-dimensional structure related to the unknown one of interest. Neutron and X-ray scattering curve modelling provides information on the full multidomain structure in solution. As scattering curves can be calculated from known atomic structures, the present availability of structures for many domains in conjunction with tight constraints based on these structures and the covalent connections between them results in a small family of allowed best-fit structures for a given scattering curve. The curve-fit procedure can be automated, and whole multidomain structures can be determined to a positional precision of the order of 0.2–1 nm. Such models are informative on the steric accessibility of each domain and their functional activity, and this is illustrated for antibody, cell-surface and complement proteins.

Introduction

Immunologically relevant proteins frequently occur as many independently folded subunits or domains that are covalently linked to form large multidomain protein structures. The elucidation of both atomic structures for single domains and the arrangement of domains within these intact proteins is important for the visualisation of these proteins and the determination of a molecular explanation of functional activities. Both protein crystallography and multidimensional NMR have yielded over 7,000 unique or related atomic structures now available in the Protein Database at Brookhaven (February 1998). A number of immunologically relevant atomic structures containing between 1 to 4 domains are now known. This still leaves many unknown structures to be determined, both at the level of individual domains, as well as that of the full structure, with as many as 20–70 domains. Crystallography is limited by the difficulty of

obtaining sufficiently well-ordered crystals of what are often significantly glycosylated and flexible structures, while NMR has size limitations in that it is ordinarily applicable only to single domain or domain pairs, and both approaches require large amounts of pure material. These factors have encouraged the development of new strategies to determine structures. One involves the identification of unknown protein folds by analogy modelling methods, based on correlations of experimental data and predictions with known molecular structures (1, 2). Another involves the determination of the domain arrangement in large uncrystallisable multidomain proteins using the constrained comparison of multidomain structural models with neutron and X-ray scattering curves (3–7). Both these strategies are discussed in this article.

Analogy modelling is based on the detection of a protein fold with low sequence similarity to the one of interest but nonetheless closely similar in structure. Homology modelling based directly on a crystal structure is usually applicable when the pairwise sequence identity between two sequences of at least 80 residues in length, one of which is that of the crystal structure, is greater than 25%. Analogy modelling becomes relevant when this sequence identity is less than 10%. That analogy modelling is possible is the consequence of the better conservation of protein structures in evolution than their sequences. In 1994, 40% of the new structures with little sequence similarity to a known fold demonstrated an evolutionary relationship with a previously known structure (8). The number of ways in which α-helix and β-strand structures within a globular protein interact with each other is necessarily limited. The simplest illustration of this is to note that, if an averaged-sized protein of 200–250 residues contains 10 α-helices or β-strands, there are 2^{10} or 1,024 possible linear permutations of these. If various compact three-dimensional packing schemes are taken into account, an upper limit to the number of arrangements may be close to 4,000 (9). Consideration of the number of genes in a genome and the proportion of these that correspond to a known structure suggests that there are only about 1,000 structural families (10). A number of unknown structures found in complement, cell-surface and cartilage proteins have been identified by these means and these are reviewed here.

Constrained solution scattering modelling is based on the use of known structures for domains within the multidomain protein in question (6). It is often forgotten that crystallography and NMR are themselves based on constraints. Many protein crystal structures are reported at 0.2–0.3 nm resolution, at which atomic detail in the electron density map is not resolved. Consequently, crystal structures are derived from standard atomic structures of amino acid residues, in which the rotations between adjacent chemical groups are determined by map fitting on the basis of the amino acid sequence. By the same token, protein NMR spectral interpretation depends initially on signal assignments, after which the structure determination is achieved by the use of distance constraints between pairs of nuclei that are applied using known atomic structures for individual amino acid residues and the known protein sequence. In application to solution scattering, although unique structures cannot be determined by this method, it turns out that relatively few multidomain structural arrangements offer good curve fits. This forms the basis of the new methodology. In constrained scattering modelling of multidomain proteins, the three major constraints are knowledge of the protein and domain volumes from the sequence, the known atomic structure of each domain (either directly, or indirectly by homology or analogy modelling), and the known covalent connections between each domain. These simple constraints are powerful enough to permit a limited number of allowed structures to be extracted to a positional precision of 1 nm from the full range of stereochemically allowed structures. In favourable cases, a single small family of related good-fit structures to a precision of 0.2 nm can be obtained. These constitute structure determinations which provide useful insights on domain accessibilities and functions. We present several examples to show the extent to which this new method has been applied for immunologically relevant multidomain proteins.

Protocol for protein fold predictions by an integrated approach

For a given protein superfamily, it is now relatively straightforward to obtain experimental and predictive data in the absence of a structure determination. For the non-specialist, the key steps are now summarised. The purpose is to provide data to correlate the superfamily with a known protein fold if a related one exists.

Secondary structure determinations are performed experimentally by means of circular dichroism (CD) and Fourier transform infrared (FT-IR) spectroscopy (11, 12). These are necessary to assign and quantify the secondary structure of the superfamily to an all-α , an all-β or an α/β or α+β class. The quantification of the CD spectrum (up to 0.5 mg/ml protein) is usually a good measure of the α-helix content as this is readily identified from a pronounced negative ellipticity between 210–220 nm and a positive ellipticity at 190 nm. Quantification of the FT-IR spectrum is best achieved using H_2O buffer (10 mg/ml protein) rather than ^2H_2O buffer (1

Table 1. Analogy modelling of three unknown protein folds

Protein fold	Evidence used in predictive analyses				Related crystal structure	Outcome of predictive analyses		
	FT-IR and CD spectroscopy	Secondary structure prediction	THREADER fold recognition	Location of key residues		Functional similarity	Sequence similarity	Gene similarity
vWF-A	Yes	Yes	Yes	Yes	Nucleotide-binding fold (ras-p21, flavodoxin)	No	Distant	No
PTR	No	Yes	Yes	Yes	C-type lectin fold (mannose-binding protein)	Yes	Distant	Yes
FIMAC	Yes (indirect)	Yes	No	Yes	Follistatin fold (SPARC, osteonectin or BM-10 protein)	Partial	Distant	Yes

mg/ml protein) in order to avoid perturbations due to ^2H_2O solvent effects. FT-IR bands at 1,620–1,640 cm^{-1} arising from β-sheets are well identified. FT-IR bands at 1,650–1,655 cm^{-1} arising from α-helices can also be identified, but need to be distinguished from loop structures that may occur at similar frequencies. The extent of band shifts observed between H_2O and ^2H_2O buffers provides indications of the degree of solvent exposure of the secondary structure.

A multiple sequence alignment requires at least 20 sequences to become useful, although 50 are preferable. Database searches that lead to the assembly of such an alignment may detect weak sequence similarities with proteins of known atomic structure in distant superfamilies (1). The alignment is useful to identify conserved Cys residues that define the disulphide bridge connectivity within the superfamily. It can also identify the conservation of functionally important residues such as those in metal-binding sites. Most usefully from the alignment, the consensus secondary structure can be predicted. That from each sequence can be averaged using the alignment to minimise the variability encountered with one-by-one predictions. Prediction accuracies of secondary structures on a residue-by-residue basis are of the order of 72 ± 9% (13). We have used averages based on the classical Chou-Fasman and GOR I/III methods as well as the more recent PHD and SAPIENS methods which incorporate averaging schemes within their algorithms (14–16). Their predictive outcomes are of comparable accuracies. The consensus prediction should identify the majority of the α-helices and β-strands that are present, and this is validated by comparison with quantitative CD and FT-IR data. Solvent accessibilities can also be predicted from the multiple sequence alignment to show whether the predicted secondary structure elements are solvent exposed or buried, and these can be compared with FT-IR band shifts measured in H_2O and ^2H_2O buffers.

The third step of analysis involves the direct detection of a related protein structure. An amino acid sequence can be scored for its compatibility with a known protein structure by the "threading" of the sequence through the three-dimensional path taken by the polypeptide chain within the structure using the THREADER program (2, 17). The pairwise interaction energy terms between all the residues in the threaded sequence are computed for the fold in question. By comparing these calculations one-by-one for a total of 254 known protein folds, it is possible to see whether the sequence is energetically compatible with a known structure in this library (8, 18). The outcome can be statistically variable, but repeating this analysis for all members of a multiple sequence alignment and taking the mean can improve matters (16). At best, a good fold correlation is discovered. If this is not possible, the unknown structure may be assigned to a general type of protein fold.

As none of these three methods can provide complete answers (1, 2, 13), their outputs are combined to strengthen the overall analysis. A positive identification of a protein fold requires that the consensus secondary structure prediction has identified most of the α-helices or β-strands present in the related fold, and that the prediction has behaved in the same way for both the known and unknown structures (19). For such a candidate fold, the construction of a molecular model based on this constitutes the fourth step of analysis to check the disulphide bridge topology and show that functionally important residues are properly located in the protein fold (*Table* 1). This is facilitated by the multiple sequence alignment. The outcome of a positive identification can be supported further by noting whether the two superfamilies are functionally similar. The exon structures of the two superfamilies in question can be examined to see whether these are also similar. While the fold predictions are a step in the direction of extracting three-dimensional information, and can provide guidance for experimental planning to test any proposed three-dimensional models, atomic structure determinations will of course be ultimately required for a proper verification of an analogy model.

The above integrated strategy successfully identified related structures for three protein superfamilies of the immune system or the extracellular matrix. The progression of the analogy modelling was different in each case (*Table 1*).

The von Willebrand factor type A (vWF-A) domain

The prediction of the vWF-A domain structure (206 residues in length) in integrins and complement proteins by analogy modelling was accomplished in several stages (14–16). FT-IR spectroscopy of a recombinant vWF-A domain from complement factor B as well as on factor B itself and its fragments showed that the vWF-A domain contained both α-helix and β-sheet structures, of which the β-sheet content was shown to be solvent-inaccessible (14). The prediction from 75 aligned sequences showed that an alternation of α-helix and β-strand structures was present (14). As both sets of data favoured an open or closed β-sheet structure flanked by α-helices, analogous known structures for these were investigated in more detail (15). THREADER sequence analyses showed that nucleotide-binding folds such as those represented by ras-p21 and flavodoxin scored highly (*Fig. 1*). The correlation was completed when it was realised that the two conserved Asp residues in the vWF-A superfamily that bound Mg^{2+} were positioned correctly at the switch point between the two halves of the central doubly-wound β-sheet, which constituted the active site in folds of this type. This prediction was performed blind. It was confirmed by the subsequent crystal structure of the vWF-A domain from the complement receptor type 3 (20). A post-mortem analysis of the prediction showed that most of the details of the protein fold prediction had been correctly identified, where the secondary structure had been predicted with 77% accuracy, and only one minor difference was found in that a β-hairpin found at one end of the structure was reversed in its direction (16). This success showed that analogy modelling is able to yield a meaningful outcome.

Since that time, the availability of several crystal structures for the vWF-A superfamily has meant that standard homology modelling methods have supplanted fold prediction methods (21). For example, the combination of homology modelling with a multiple sequence alignment of the A1 and A2 domains of 28 mammalian sequences for von Willebrand factor showed that a heparin-binding site in the A1 domain could be predicted. The different functional consequences involved in type 2B and 2M von Willebrand disease (a common bleeding disorder) could be explained in terms of this homology model and two sets of different locations of the mutation sites on opposite

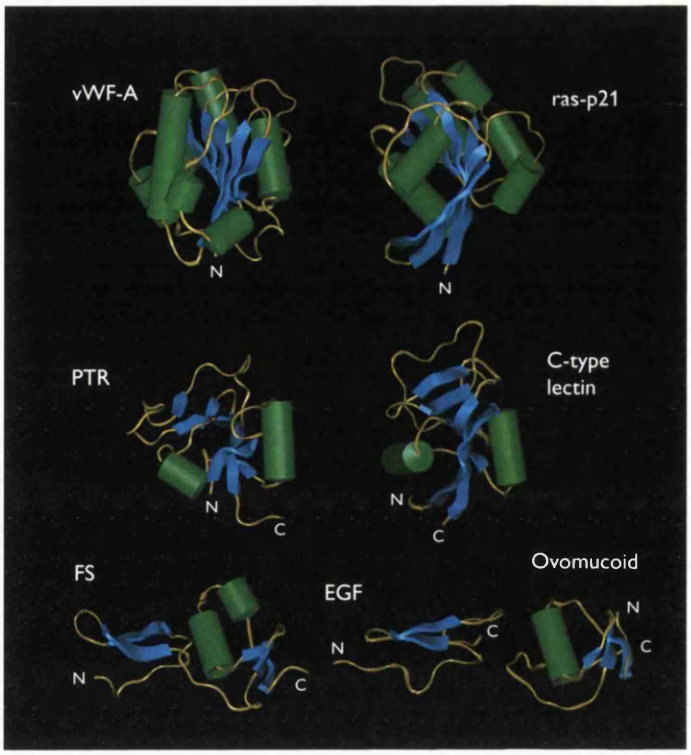


Fig. 1. Comparison between three predicted protein folds and the related crystal structures for which analogy relationships had been described. The vWF-A structure of complement and integrin proteins was correlated best with ras-p21 and flavodoxin. The PTR structure of aggrecan was correlated with the C-type lectin fold of mannose-binding protein of the complement lectin pathway. The follistatin structure (FS) of SPARC protein is a hybrid of an EGF domain and an ovomucoid domain, and shows similarities to the FIMAC domain of several complement proteins although this identity is yet to be made definitive. Green, α-helix; blue, β-strand; N, N-terminus; C, C-terminus. This figure and others were prepared using INSIGHT II 95.0 (Biosym/MSI).

sides of the A1 domain, one of which involved the predicted heparin site, and the other a Gp1b receptor site. These observations provided a molecular explanation in allosteric terms of the upregulation or downregulation of the binding of von Willebrand factor to its Gp1b receptor on platelets in the type 2B and 2M disease conditions.

The proteoglycan tandem repeat (PTR) domain

The prediction of the unknown PTR structure found in aggrecan and link protein of the extracellular matrix followed a related path to that of the vWF-A prediction (19). Aggrecan contains an immunoglobulin (Ig) fold and four PTR domains in two globular regions, G1 and G2, at its N-terminus, and a C-type lectin domain and variably spliced amounts of two other domains in a globular region, G3, at its C-terminus (22). The PTR domains in G1 interact with the anionic polysaccharide hyaluronan. Analogy modelling of the PTR fold on the basis of

59 aligned sequences was carried out side-by-side with the homology modelling of the C-type lectin domain in G3 of aggrecan using an alignment of 129 sequences. The happy coincidence of our work with both these superfamilies (23, 24) resulted in the recognition that the unknown PTR structure was closely related to the C-type lectin structure found in mannose-binding protein of the complement lectin pathway and other diverse families (Fig. 1). Both multiple sequence alignments gave the same consensus secondary structure prediction of a βαβα-β structure with an extended loop and β-sheet region at the centre (-). Both alignments also contained the same conserved disulphide bridge within this βαβα-β motif, both scored highly with a crystal structure for the C-type lectin fold in THREADER analyses, and both had similar exon structures that correlated with a distinct βαβα-β motif structure of a three-stranded β-sheet core flanked by α-helices on each side. Since C-type lectins generally interact with carbohydrate, a functional similarity with the PTR is present. These correlations resulted in the prediction that the PTR fold was similar to that of the C-type lectin fold (19). The construction of a molecular model was consistent with this, and demonstrated a surface patch of conserved basic residues that may interact with hyaluronan. The subsequent publication of an NMR structure for the PTR showed that its secondary structure had been predicted with 85% accuracy, and that the PTR fold had been identified as predicted, with only one minor exception in that the second α-helix of the βαβα-β motif was reorientated compared to that in the C-type lectin fold (Fig. 1) (22, 25). Interestingly, in an extension of the prediction, the homology modelling of the group of C-type lectins found in G3 of aggrecan and other proteoglycans also demonstrated a set of conserved basic residues in a surface patch (22). One of these residues is also conserved in the PTR sequences and has been implicated in interactions with hyaluronan. This indicated an even closer relationship between both folds.

The factor I/membrane attack complex (FIMAC) domain

The prediction of the fold for the complement FIMAC domain followed a third path (26). As the FIMAC domain with 10 Cys residues had only been identified in three complement proteins, it was postulated that analogues might exist in other sequence superfamilies. A sequence database search was performed which detected distant similarities with the follistatin superfamily of the extracellular matrix and the endocrine system. Members of this superfamily also contained 10 Cys residues at positions similar to those found in the FIMAC domain. This similarity was strengthened by noting that similar consensus βββββαβ secondary structure predictions were calculated from sequence alignments of each of the FIMAC and follistatin superfamilies. The results of these were consistent with CD and FT-IR data for complement factor I that showed that this is mostly β-sheet in its structure (27). At this point, the crystal structure of the follistatin domain in the SPARC protein of the extracellular matrix became available (28). A molecular model of the FIMAC domain could be readily constructed from this by three small deletions in surface loop regions to provide further support for the premise that the two superfamilies were indeed related (29). The follistatin structure is in fact a hybrid of an epidermal growth factor (EGF) domain (ββ) with an ovomucoid domain (ββαβ), and neither of these was detected by THREADER fold recognition searches (Fig. 1). A definitive verification of the structural identity between the FIMAC and follistatin superfamilies is currently under investigation.

Background to X-ray and neutron scattering

X-ray and neutron scattering is summarised for the non-specialist (5). The protein in solution (1–10 mg/ml) is irradiated by a collimated beam of known wavelength. This results in a two-dimensional radially symmetric diffraction pattern. Its intensity I(Q) as a function of the scattering vector Q is measured on an area detector positioned behind the sample (where $Q = 4\pi \sin\theta/\lambda$; scattering angle = 2θ; wavelength =λ). Guinier analyses of I(Q) at low Q values give the radius of gyration R_G (a measure of structural elongation) and the molecular weight. At larger angles, further structural details are resolved. Scattering is complementary to electron microscopy, in which a static macromolecular structure is visualised when flattened or stained on a grid, often in a vacuum, although these conditions of measurements can potentially perturb the structure of interest. It is also complementary to analytical ultracentrifugation, which provides less detailed information on macromolecular elongation from sedimentation co-efficients and on molecular weights from sedimentation equilibrium. More detailed accounts of scattering and the complementary nature of X-ray and neutron scattering are provided elsewhere (3–7).

X-rays observe structures in which the hydration shell at the protein surface is detectable, while neutrons do not see this shell (30). Different atoms have different scattering properties, which are defined by their atomic scattering factors. Those for X-rays depend on the number of electrons, while those for neutrons depend on the nucleus, being similar and positive for ^{2}H, C, N and O, but negative for ^{1}H. The X-ray scattering factor for O compared to C and N turns out to be the highest for these three atoms, while that for neutrons is the lowest of the three. In application to the hydration shell, water molecules occupy a

smaller volume in this than in bulk water which contains looser hydrogen-bonding arrangements. Consequently, the electron density of the hydration shell is higher than that of bulk water, and this is detectable by X-ray scattering. In neutron scattering, the joint effect of the occurrence of rapid ^1H-^2H exchange in the hydration shell and the lower neutron scattering factor of O causes this to become invisible. This has been confirmed in structural calibration tests with proteins of known crystal structure in the molecular weight range 23,000–127,000 (31–33). This means that the scattering curve modelling of protein structures can be directly based on atomic structures for neutron data, while these structures have to be adapted by addition of a hydration shell for X-ray modelling.

X-ray and neutron scattering monitors different solute-solvent contrasts (4). Scattering occurs when the scattering density of the protein is different from that of the buffer. X-rays visualise proteins in a high positive contrast because the protein is more electron dense than the buffer. By neutrons, this contrast can be manipulated when H_2O and ^2H_2O buffers are used because these have scattering densities that are respectively below or above that of the protein. Lipids, proteins, carbohydrates and RNA/DNA each have distinct nuclear scattering densities. To analyse immunologically relevant glycoproteins, it has to be shown that the different scattering densities between protein and carbohydrate (or even between hydrophilic and hydrophobic amino acids) do not significantly affect scattering curve modelling based on atomic structures. The necessary evidence is provided by comparative measurements by X-rays in positive contrasts and by neutrons in negative contrasts using ^2H_2O buffers, and this is why both measurements are carried out.

Constrained scattering curve modelling of multidomain proteins

What structural resolutions can be achieved by scattering modelling analyses? For a scattering curve I(Q) measured between Q of 0.05 to 2.00 nm^{-1}, the theoretical resolution is given by $2\pi/Q$ and is about 3 nm. Domains are usually about 3–4 nm in length (e.g. an Ig fold with 100 residues is 4.4 nm × 3.4 nm). So scattering is easily able to discriminate between extended or compact domain arrangements in multidomain proteins (*Fig.* 2). Scattering becomes more useful if it can be used to derive molecular structures. In such cases, the scattering curve modelling itself must be strictly constrained by the known volume, atomic structure and steric connectivity between the domains in the multidomain structure. The correctness of the assumptions used to initiate a modelling analysis is therefore important. From a suitable starting model, a full range of stereochemically allowed structures is generated to test these against the experimental scattering curve. The co-ordinate models are converted to small-sphere models in order to calculate the scattering curves (*Fig.* 2). The curve fits are automatically assessed using filters based on the observed scattering parameters, and significantly only a low proportion of a large number of structures usually give good curve fits. Even though scattering analyses are not able to identify a unique structure, if only a single family of related good-fit structures is found to fit the data, the domain arrangement in question is then identified. Since domain translations of as little as ± 0.2 nm can result in visibly worsened curve fits, this appears to be an attainable level of structural precision when the modelling constraints are strict enough. The quality of the scattering curve fits is monitored by the R-factor which is defined in the same way as that used by protein crystallographers. Typically for good quality I(Q) scattering data extending to a Q value of 2 nm^{-1} for which the value of I(0) is normalised to 1,000, the R-factors for X-ray and neutron curve fits should range between 2–10%. The R_G value of a good structural model should be within the experimental error of measurement (± 0.1 nm or 5%).

If subunits or domains can be located at best to within ± 0.2 nm in position, how biologically useful are scattering models of multidomain proteins? It cannot be assumed that all multidomain proteins will possess highly extended domain arrangements. Scattering models are thus important when it is not possible to crystallise a multidomain protein for reason of interdomain flexibility or high glycosylation. That these domain arrangements are studied in solution enhances the usefulness of this method. It will show whether a specific domain arrangement can exist that is related to its function. It is possible that the surface accessibility of specific domains may be important for function, for example, if they are masked in a precursor form and become unmasked upon activation. The proximity relationship between two domains may be important, where one domain may directly control the activation or inhibition of the other. The spatial separation of domains may be important for the generation of multiple binding sites within a given macromolecular structure. The carbohydrate conformation in glycoproteins can be assessed.

More precise scattering analyses are possible if the constraints are directly based on a newly determined crystal structure. The crystal lattice packing does not always reveal whether such an observed structure corresponds to a monomer, a dimer or some higher oligomer in solution. Alternatively, a monomeric structure may have been crystallised, even though a dimeric structure is known to exist in solution. Such investiga-

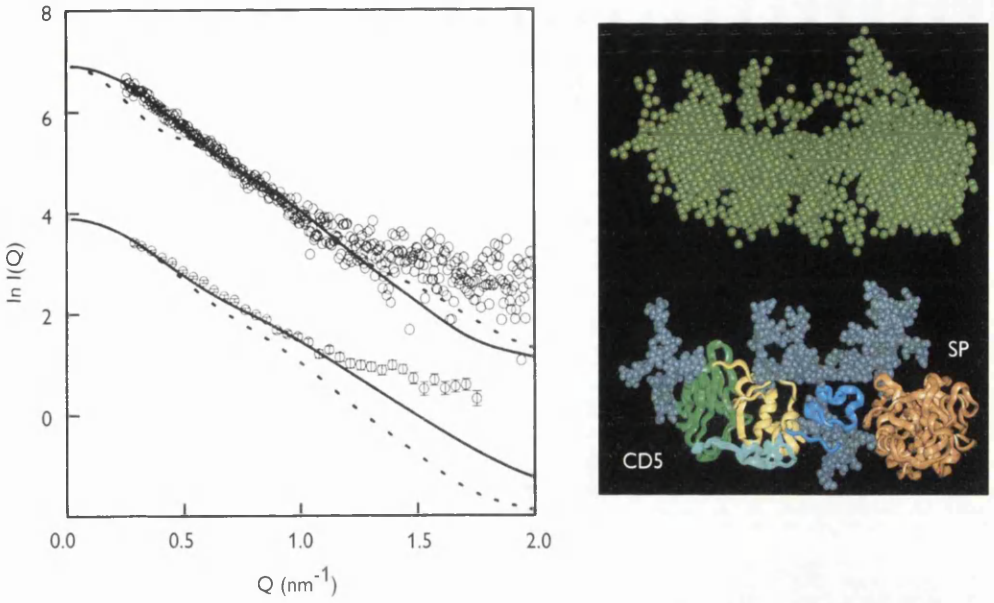


Fig. 2. X-ray and neutron best-fit modelled curves for the bilobal five-domain model for factor I. The FIMAC, CD5, LDLr-1/2 and SP domains are denoted by yellow, green, two shades of blue and orange in that order. The six oligosaccharide structures are shown in grey. The co-ordinates of the model are used to derive the sphere model shown in olive green, from which the scattering curve is calculated. The X-ray (upper) and neutron (lower) curve fits are shown as circles for the experimental data and continuous lines for the best-fit modelled curves. The dashed line attached to the X-ray curve indicates how the calculated curve is different if all five domains form a linear arrangement. The dashed line attached to the neutron curve corresponds to the best-fit X-ray curve to show how this is different as the result of X-ray hydration and neutron instrumental corrections.

tions are essential for a full molecular description of the crystal structure. Since the modelling of the experimental curve is now more tightly constrained, the analyses involve only a choice between a restricted number of alternative structures, and the precision is improved.

Domain arrangements in multidomain antibody structures

For monomeric antibodies such as IgG and IgE, the locations of the antigen-binding loops in the Fab fragments relative to both the hinge at the centre of the antibody structure and the steric accessibility of the receptor sites on the Fc fragment are important for function. In the case of IgG and pentameric IgM, the Fc interactions with C1q of complement are also important for function.

Extended structures of the bovine IgG1 and IgG2 isotypes

The bovine IgG isotypes IgG1 and IgG2 differ in that only IgG1 is transported into milk by specific cell receptors. Both isotypes contain two four-domain Fab fragments that recognise antigen while the four-domain Fc fragment contains the receptor-binding site. The number of residues in the two polypeptide hinges joining the Fab and Fc fragments differs in IgG1 and IgG2, where IgG2 has a seven-residue deletion in the hinge sequence. The receptor specificity for IgG1 and not for IgG2 could variously result from a specific binding site at the hinge region-Fc junction in IgG1 that is absent in IgG2, different hinge conformations in IgG1 and IgG2, or steric obstruction of the receptor site by the Fab fragments in IgG2. Accordingly, there was much interest in determining the relative positions of the Fab and Fc fragments to distinguish between these alternatives. Very few antibodies have been crystallised intact for reason of flexibility at the central hinge (34). Molecular modelling done prior to the scattering analysis suggested that short-, medium- or long-length hinge regions were possible, and steric hindrance of effector function was considered to be likely. Constrained scattering modelling was carried out in order to clarify the solution structures of the two isotypes (35).

Interestingly, neutron scattering showed that IgG1 and IgG2 had similar radii of gyration R_G of 5.64–5.71 nm in ^2H_2O buffers, and gave similar distance distribution functions P(r). As the cross-sectional radii of gyration R_{XS} were also similar, it was concluded that bovine IgG1 and IgG2 possessed similar domain arrangements in solution, despite their different hinge

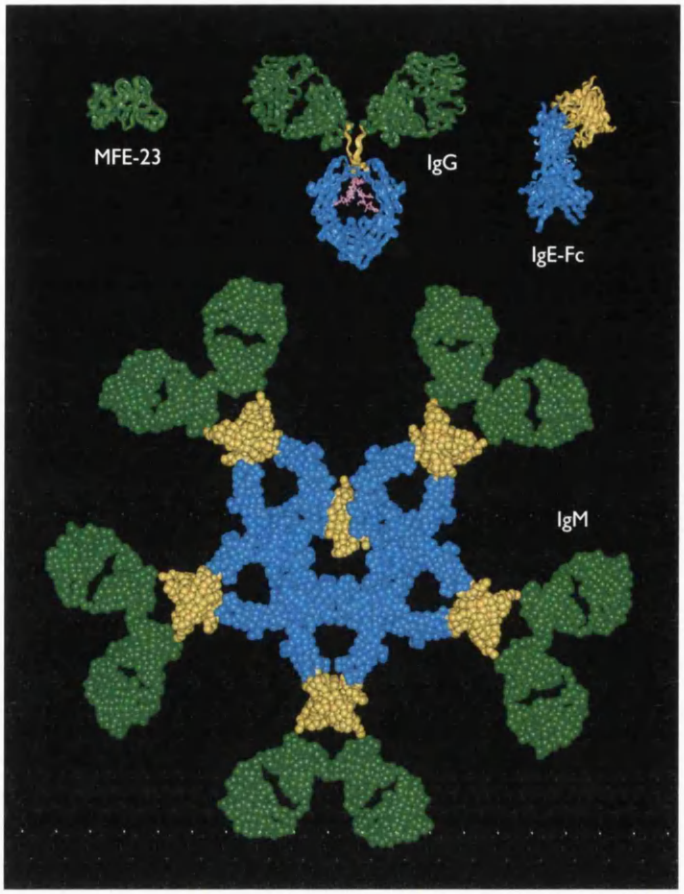


Fig. 3. Outline structures of four antibody structures, MFE-23, IgG, IgE-Fc and IgM, determined by solution scattering. All four models correspond to the best-fit scattering curve structures. The Fab fragments are denoted by green, and the Fc fragments are denoted by blue. The hinge polypeptide in IgG, the $(C\varepsilon 2)_2$ domains in IgE and the five $(C\mu 2)_2$ domains and the central J chain in IgM are denoted by yellow. Carbohydrate is shown only for IgG (magenta).

sequences. Scattering modelling was carried out in order to locate the relative positions of homology models for the Fab and Fc fragments. A translational search was performed in which the two Fab fragments were displaced in two directions in a plane that corresponded to the major plane of the Fc fragment. The hinge was omitted from these searches at this stage as this is small in size. Good curve fits were obtained in models where the separation of the Fab C-terminus and Fc N-terminus α-carbon atoms ranged between 2.9–3.6 nm. Having optimised the location of the three fragments, the hinge sequences were now added to the models, and energy refinements showed that the 2.9–3.6 nm separation between the Fab and Fc fragments was stereochemically consistent with the two different hinge sequences (*Fig.* 3). From this work (35), the sequence deletion in the IgG2 hinge was seen to be a likely cause of the exclusion of this isotype from the transport process into milk, as gross structural differences between the two forms had been ruled out from the scattering analyses.

The bent IgE-Fc fragment of IgE

The interaction between IgE and its high-affinity receptor FcεRI is central in the allergic response (36). IgE contains 14 domains, in which an additional pair of domains $(C\varepsilon 2)_2$ between the Fab and Fc fragments replaces the hinge peptides in IgG (*Fig.* 3). The location of the $(C\varepsilon 2)_2$ domain pair relative to those of the 4 domains in IgE-Fc is relevant for understanding IgE-receptor interactions, and prompted its study by X-ray and neutron scattering (37). The upper limit on its R_G value was determined to be 3.53 nm and that on its maximum length was 13 nm.

The scattering modelling of IgE-Fc was initiated with two available homology models for IgE-Fc in the Brookhaven database (codes 1ige and 2ige). Even though calibration studies showed that good scattering curve fits can be obtained using known crystal structures, neither of the two models gave good curve fits. This was attributed to the unrefined position of the $(C\varepsilon 2)_2$ domains as it was believed that the $(C\varepsilon 3)_2$ and $(C\varepsilon 4)_2$ domains could be represented by the crystal structure of the corresponding four domains in IgG-Fc. Accordingly, the IgE-Fc model with the correct disulphide bridging (2ige) was subjected to automated constrained modelling which optimised the relative positions of the $(C\varepsilon 2)_2$ domain pair, the two Cε3 domains, and the $(C\varepsilon 4)_2$ domain pair by translations and rotations. The testing of thousands of possible IgE-Fc models showed that a bent IgE-Fc structure with a Cε2 rotation of 70° out of the Fc plane and an unchanged Cε4 rotation of 0° gave an excellent X-ray curve fit (*Fig.* 3). Contour maps showed that this corresponded to a single best-fit minimum. The precision of this model was estimated to be between 40° and 90° for the Cε2-Cε3 bend angle and ± 50° for the Cε3-Cε4 bend angle. Planar or linear IgE-Fc domain structures did not fit the scattering data. This outcome confirmed the bent structure previously proposed for intact human IgE by fluorescent labelling studies (38), and yielded an optimised set of dimensions for the six domains in this Fc fragment. Such a study provides a basis for understanding the molecular interaction between IgE-Fc and its FcεRI receptor.

The pentameric IgM structure and its interaction with complement

IgM is a planar pentameric antibody in which each monomer contains 14 domains in two Fab fragments and one Fc fragment, and a J chain with 1 domain is also present in the pentamer. Even when its binding to single determinants is weak, its pentameric structure enhances its binding to antigens with multiple epitopes. Once IgM is bound polyvalently through its Fab fragments to an antigenic surface, it is stabilised into a five-

legged table-like conformation that is able to bind complement C1q, a hexameric molecule with six globular heads joined by collagenous stalks. The purpose of the scattering modelling of pentameric IgM was to define the relative position of the five $F(ab')_2$ arms relative to a central disulphide-linked disk of five Fc fragments (39). This permitted the flexibility of IgM to be assessed, as well as the location and accessibility of the Cμ3 domains in the Fc disk that contain the C1q-binding site.

The complexity of modelling this 71-domain structure meant that X-ray scattering curves were recorded not only for intact human and mouse IgM but also for the IgM monomer (14 domains), the Fc_5 central disk (21 domains: blue in Fig. 3), the $F(ab')_2$ fragment (10 domains) and a Fab fragment (4 domains: green in Fig. 3). Each multidomain structure was modelled individually using X-ray curve fits. In particular, the Fc_5 central disk was modelled by the constrained docking of five homologous Fc crystal structures in order to create the correct disulphide bridges in a ring, and such a structure gave a good scattering curve fit. The intact IgM structure was assembled from these fragments for testing against its scattering curve. Surprisingly, a rigid planar five-fold symmetric pentameric IgM structure (Fig. 3) did not give a good curve fit. This was investigated to show that good curve fits were obtained if it was postulated that the $F(ab')_2$ fragments had conformational mobility in the plane of the Fc_5 disk and could swing side-to-side in this plane. If the planes of the $F(ab')_2$ fragments were orientated perpendicular to the plane of the Fc_5 disk, much worsened curve fits were obtained and this perpendicular model was clearly ruled out.

The dimensions of this pentameric IgM scattering model permitted the assessment of the C1q-binding site on IgM that would activate complement. That in IgG is found at residues Glu318-Lys320-Lys322 on its Fc fragment. As the C1q-IgM interaction showed a strong ionic-strength dependence, this implicated likewise charged residues in the Fc fragment of IgM. By the comparison of IgM sequences from nine species, Asp/Gly432 and His430 were identified as suitably conserved charged residues in the Fc disk. These residues were located on the outer periphery of the Fc_5 homology model at separations that were consistent with an interaction with C1q heads. These locations were at sites that are sterically inaccessible to the C1q heads in free planar IgM for reason of the proximity of the $F(ab')_2$ arms (between the blue and yellow domains in Fig. 3). The IgM model showed that these sites become exposed to C1q binding after IgM has bound to an antigen in the table-like conformation with the $F(ab')_2$ bent out of the plane. Comparison of the dimensions of the IgM model and a scattering model for C1q showed that C1q binding to IgM would cause conformational perturbation of the C1q stalks. Thus, a molecular explanation could be proposed to explain how IgM-antigen binding can transmit a conformational change to C1q to activate complement. Experimental support for this mechanism has since been provided by site-specific mutation and functional studies of recombinant IgM (40).

Domain arrangements in cell-surface proteins and complement proteins

In cell-surface proteins, cellular adhesion will be governed by the location of domains at set distances from the membrane surface. The scattering analysis of the seven-domain carcinoembryonic antigen (CEA) clarified its domain arrangement, as well as the role of its 50% carbohydrate content. Domain structures are often poorly understood for the complement proteins, and the non-extended arrangement of the five domains in factor I determined by scattering was of particular interest.

Extended structure of CEA

CEA is a homophilic cell adhesion molecule of the Ig superfamily that is widely used as a cell-surface marker for tumour monitoring and antibody targeting (41). A molecular structure for CEA would clarify its adhesive function as well as the rational design of anti-CEA antibodies. CEA differs from IgG, IgE and IgM in that it contains one V-type, three I-type and three C2-type Ig domains in place of V-type and C1-type domains. Each CEA domain is constructed as a β-sandwich structure from two β-sheets denoted EBA and GFC. CEA also contains 28 oligosaccharide chains. For reason of interdomain flexibility and glycosylation, it is most unlikely that CEA could be crystallised. Accordingly, CEA was cleaved from membranes and subjected to X-ray and neutron scattering analyses (42). Guinier plots gave an R_G value of 8.0 ± 0.6 nm that showed that CEA contained an extended domain arrangement. A high cross-sectional R_G value was measured to show that the oligosaccharide chains were extended outwards from the protein surface. The molecular weight determination of 150,000 showed that CEA was monomeric.

The scattering modelling of CEA was performed using both domains in the two-domain crystal structure of the CD2 cell-surface protein, which were duplicated to give a linear seven-domain structure. The 28 oligosaccharide chains were added at glycosylation sites in extended conformations. Rotational searches about three axes at each domain interface were performed, in which for simplicity all six sets of interdomain angles were fixed to be the same. The resulting 4,056 models could be classified into four structural families, namely linear,

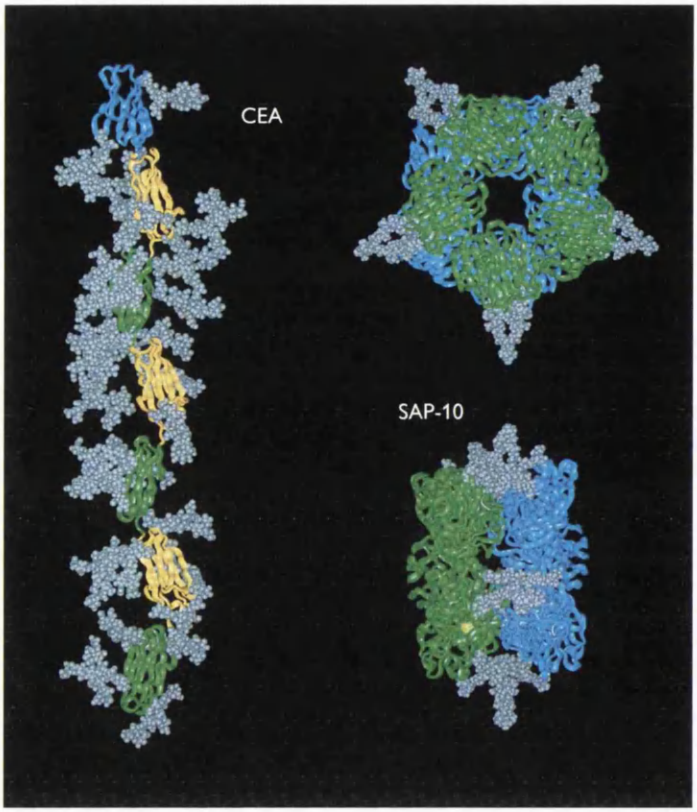


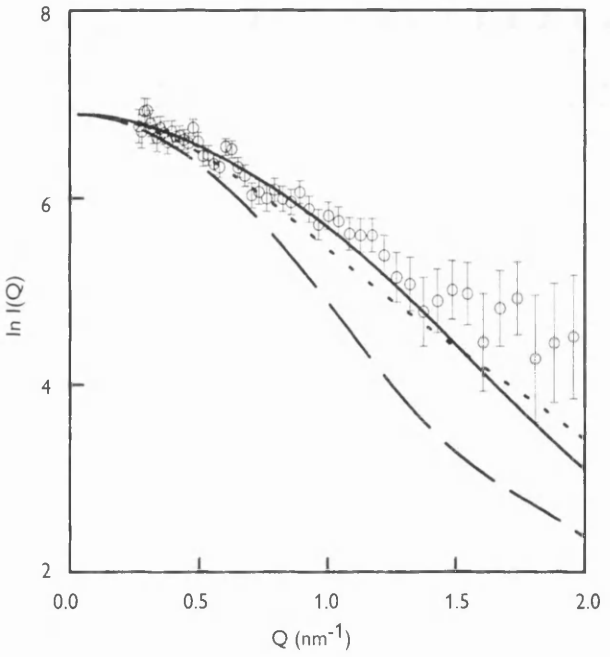
Fig. 4. Structures of the heavily glycosylated structures for CEA and decameric SAP studied by scattering analyses. The best curve-fit zig-zag CEA model (left) is shown as a V-type domain (blue) together with three I-type domains (yellow) and three C2-type domains (green). The 28 extended oligosaccharide chains are shown in grey. Note that the N-terminal V-type domain is comparatively carbohydrate-free at its GFC face. The best curve-fit SAP decamer model (right) is shown in face-on and side-on views, with the two pentamers in green and blue. The 10 extended oligosaccharide chains are shown in grey, and the scattering modelling suggests that the chains may be in proximity to each other in opposing pentameric subunits.

curved, zig-zag and helical, of which only the zig-zag family gave good curve fits. Interestingly, the best-fit CEA model had an interdomain orientation similar to that seen in the CD2 crystal structure (*Fig.* 4). The best-fit CEA model suggested that the EBA and GFC faces of adjacent domains alternate with each other along the long axis. As the GFC β-sheets were seen to contain little or no carbohydrate, especially so for the V-type domain (*Fig* 4), the modelling also suggested that the alternation of these protein surfaces along the axis of CEA may be optimal for the formation of homophilic ligand sites for CEA molecules between different adjacent cell surfaces.

Unusual compact domain structure in factor I

Factor I of complement is a five-domain two-chain serine protease (SP) with 26% carbohydrate that regulates the classical and alternative pathways of activation (43). It cleaves the complement components C3b and C4b into smaller fragments in the presence of soluble or membrane-bound protein co-factors. Deficiencies of factor I lead to the excessive consumption of C3 and recurrent pyogenic infections. The heavy chain contains four small domains, namely the FIMAC domain, the CD5-type domain (also known as the scavenger receptor cysteine-rich domain) and two low-density lipoprotein receptor (LDLr-1/2) domains. The heavy chain is disulphide-linked to the light chain which contains a single large SP domain. The domain arrangement in factor I was shown by scattering to be of maximum length 14 nm, which is too short to be explained by an extended arrangement of these five domains (29).

A molecular model for the domains in factor I was constructed using homologous crystal structures for the FIMAC, LDLr-1/2 and SP domains and electron microscopy dimensions for the CD5 domain (*Fig.* 2). Of the 40 Cys residues present in factor I, 38 could be assigned to predicted buried disulphide bridges in the five domains, while 2 were predicted to be surface-exposed on the FIMAC and LDLr-1 domains. As no free Cys residues could be detected in assays, it was postulated that these surface Cys residues were disulphide-bridged to form a large compact triangular arrangement of the FIMAC, CD5 and LDLr-1 domains. This was constructed using molecular graphics. Translational and rotational searches were used to test 9,600 arrangements of the triangular structure relative to the large SP domain, of which only a limited subset gave a small family of good curve-fit structures. Interestingly, all the searches positioned the LDLr-2 and SP domains close together (*Fig.* 2). This same result was obtained from four independent curve-fit analyses based on X-ray and neutron data for two dif-



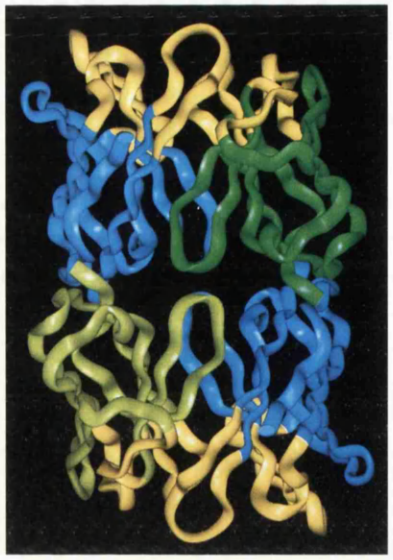


Fig. 5. Neutron curve fits for three alternative models of the scFv molecule MFE-23 seen in its crystal structure. The V_L and V_H domains are shown in blue and green, respectively. The antigen-binding loops are shown in yellow. The neutron data are shown as circles. The calculated curves are as follows: the compact monomer (upper or lower pair of blue and green domains; continuous line), the elongated monomer (left or right pair of blue and green domains; dotted line) and the full dimer (four domains; dashed line).

ferent forms of factor I. It was deduced that this non-extended domain arrangement in factor I exposed the SP domain and the CD5/LDLr-1 domain pair at opposite ends of factor I for possible independent interactions with C3b/C4b and co-factor proteins, respectively, while the FIMAC domain was inaccessible at the centre. This outcome is most useful to guide the rational design of further experiments to test this possible bilobed model.

Direct usage of crystal structures in scattering curve modelling

Unambiguous results were obtained in scattering analyses to distinguish the monomeric or dimeric forms of a single-chain Fv (scFv) antibody fragment (44) and identify the association of two pentamers of serum amyloid P component (SAP) to form a decamer (33).

Monomeric and dimeric structure of scFv molecules

Antibodies have been used in a wide variety of clinical applications (45). MFE-23 is an anti-CEA scFv antibody molecule that was genetically engineered for good tumour targeting. A two-domain scFv molecule contains a V_H domain joined to a V_L domain by a peptide linker to prevent the loss of antigen-binding function by separation of the two domains. This is the smallest antibody fragment to retain antigen-binding functions (Fig. 3), and should show better tumour penetration and reduced immunogenicity than larger four-domain Fab fragments. The MFE-23 crystal structure showed well-defined four-domain dimers in the crystal lattice packing (44). Since the existence of such a dimer would negate the size advantage of an scFv molecule compared to a Fab fragment, and since many scFv molecules are known to be dimeric in solution, the domain structure of MFE-23 was identified by neutron scattering (32). Its molecular weight was determined to be 27,300 at 1 mg/ml or less and its R_G value was 1.88 nm. Three alternative arrangements of the four MFE-23 domains were distinguished by modelling calculations (Fig. 5). Only the compact V_H-V_L two-domain arrangement agreed with the observed R_G value and neutron scattering curve. Thus, MFE-23 was determined to be monomeric at the concentrations used therapeutically, and this validates its clinical usage.

Pentamer and decamer structures of SAP

SAP is a plasma glycoprotein that contains five subunits in a flat disk with five-fold cyclic symmetry (46). It binds to amyloid

Table 2. Domain arrangements in immunologically relevant proteins by constrained scattering modelling

Multidomain or multisubunit protein	Available crystal structures	Significance of multidomain arrangement	Molecular weight	Experimental R_G (nm)	Modelled R_G (nm)	R-factor of curve fit
IgG	2 × Fab and 1 × Fc from IgG	Semi-extended hinge region, permitting access to receptors	144,000	5.64–5.71	5.31	1.2
IgE-Fc	1 × Fc from IgG, and a domain pair from Fab	Fc structure is bent, influencing its interactions with its receptor	75,300	3.52–3.53	3.22	3.4–6.3
IgM	5 × Fab and 5 × Fc from IgG and a domain pair from Fab	Planar structure, ability to bend to expose complement C1q sites	976,000	12.3	12.3	1.9
CEA	CD2 and CD4 (7 × Ig domains)	Extended zig-zag structure, forming homophilic cell adhesion	152,500	8.0–8.8	6.9–8.0	4.7–8.7
Factor I of complement	1 × SPARC; 1 × CD5 (electron microscopy); 2 × LDLr; 1 × urokinase-type plasminogen activator	Bent-back domain arrangement, forming a bilobal structure that interacts with its substrate and co-factor	85,300	4.00–4.04	4.10–4.17	10.2
ScFv from IgG (MFE-23)	Dimeric MFE-23	Monomeric when used therapeutically	27,200	1.88	1.80–2.00	9.5
SAP	Pentameric SAP	Decamer is formed by the interaction between the Trp-rich face of two pentamers	127,000 (254,000)[a]	3.69–3.99 (4.09–4.23)	3.80–3.97 (4.13–4.23)	3.7–4.0 (3.4–4.7)

[a] Bracketted values correspond to the SAP decamer

fibrils, sulphated glycosaminoglycans, DNA and chromatin, and is universally present in amyloid deposits. Its crystal structure showed that each subunit is predominantly a β-sheet sandwich. There is a single N-linked oligosaccharide site and an α-helix on its A-face and a calcium-binding site on its B-face. SAP forms stable decamers in the absence of calcium. Since the decamer had resisted all efforts to crystallise it, X-ray and neutron scattering was performed to determine how the decamer is formed (33). The scattering curves of SAP pentamers and decamers are clearly different, and this provides an unambiguous way to discriminate these which is not easy to do by gel filtration (46). Guinier molecular weight calculations curiously reported the ratio of the decamer:pentamer molecular weights to be 1.7 instead of 2.0 as expected. This was attributed to the proximity of 4 Trp residues near the A-face of each subunit, which would bring a total of 40 Trp residues close to each other if the ten A-faces associated to form the decamer, and would affect the 280 nm absorption co-efficient of SAP used to determine its molecular weight by scattering. This proved that the two A-faces were associated in the SAP decamer.

Curve modelling identified the oligosaccharide conformation for pentameric SAP which was not visible in the crystal structure (33). Good fits were only obtained with extended oligosaccharide structures. Curve modelling for SAP decamers was carried out to assess the possible A-A and B-B structures (33). The modelling was much simplified by the use of the five-fold symmetry about a common central axis. All that was required was to place two opposing pentamers on this common axis and to translate one completely through the other in 0.1 nm steps to create a full range of models. This favoured the A-A structure, for which the translational precision could be determined to ±0.2 nm (*Fig. 4*). In this decamer structure, it was likely that the oligosaccharide chains from opposite pentamers were close to each other, and prompted the hypothesis that these might assist stabilisation of the decamer.

Conclusions

The importance of three-dimensional information on the domains in immunologically important multidomain proteins has prompted the development of the two new strategies described in this review.

The ability to predict unknown protein folds by combining CD and FT-IR spectroscopy data, averaged secondary structure predictions, protein fold recognition searches and trial model building has proved most useful in our analyses of the vWF-A, PTR and FIMAC superfamilies (*Table 1*). Not only could the major features of these three protein folds be identified, but the experience from these three studies has clarified functional aspects of these proteins. This facilitates the rational planning of molecular biology and biochemical strategies in the absence of crystal structures. The extent of success in these investigations depends on whether or not a related fold exists. For example, in our predictive analysis of the LDLr superfamily found in the complement proteins, it was only possible to show that this is a small β-sheet protein similar in nature to that of the EGF but structurally distinct from it (27). This limited outcome was nonetheless confirmed from the subsequent NMR and crystallographic structures (29). Much interest has been generated in this strategy (47). Another study based on sequence conservation, secondary structure predictions, fold recognition, and model building of disulphide bridges resulted in the proposal

that the seven sequence repeats found at the N-terminus of the integrin α subunit are structurally related to a seven-fold β-propeller structure found in the trimeric G-protein and several other proteins (48). Some necessary caution is shown by the different prediction outcomes for the integrin β subunit by two groups, in which attempts were made to fit the 200 residues of a vWF-A structure to an alternating pattern of predicted α-helices and β-strands found in the full 800-residue sequence (49, 50). Caution is also indicated by minor differences seen between the analogy modelled structures of the vWF-A and PTR domains with their subsequent structure determinations. In general, analogy modelling makes useful contributions toward structural analyses, but crystal structures or other experimentation will always be needed to verify the outcome of these analyses.

Neutron and X-ray solution scattering are powerful complementary techniques that enable the structures of multidomain proteins to be determined in solution. Its reliability has been tested by calibration studies based on known atomic coordinates. It can be applied to a diverse range of problems in immunology that range from the determination of the unknown domain arrangement in a multidomain protein to the completion of a newly determined crystal structure (*Table* 2). Based on constraints from known atomic structures, an automated curve-fit search will identify a best-fit structure for a given scattering curve from several thousand possible structures. Strictly, a good scattering curve fit is only a test of consistency and will not constitute a unique structure determination. However, if only a single family of related best-fit structures is obtained after a comprehensive constrained structural search, it is not unreasonable to conclude that this can represent a unique structure determination with a positional precision as high as ± 0.2 nm. It is of course possible that the multidomain protein possesses interdomain flexibility which will affect the outcome of scattering modelling, but in such cases the scattering analysis simply places limits on the extremes of conformations allowed by flexibility. In application to antibody structures, examples were presented to show how the elucidation of domain arrangements in IgG, IgE and IgM clarified the molecular basis of their interactions with receptors or complement proteins, in particular the steric accessibility of specific domains. In application to the linear domain arrangements found in CEA and factor I, it was interesting to find that the seven domains in CEA comprise an extended zig-zag arrangement in solution which could nonetheless mediate cellular adhesion despite its high carbohydrate content, while the five domains in factor I formed a more compact arrangement with various types of domain accessibility within this structure. In application to recent crystal structures, it was possible to identify a monomeric MFE-23 structure in solution in the conditions when it binds to CEA, while pentamers and decamers of SAP could be readily distinguished and the structure of the SAP decamer could be determined. Solution scattering is sensitive to oligosaccharide conformations on protein surfaces, and these could be identified as extended structures in CEA and SAP, even though these are not ordinarily visible in crystal structures. The exploitation of tight molecular constraints in scattering curve modelling has resulted in significant advances in this field.

References

1. Sternberg MJE, ed. Protein structure prediction. Oxford: IRL Press; 1996.
2. Jones DT. Progress in protein structure prediction.
Curr Opin Struct Biol 1997;**7**:377–387.
3. Perkins SJ. Structural studies of proteins by high-flux X-ray and neutron solution scattering.
Biochem J 1988;**254**:313–327.
4. Perkins SJ. X-ray and neutron solution scattering.
New Comp Biochem1988;**11B**:143–264.
5. Perkins SJ. High-flux X-ray and neutron solution scattering. In: Jones C, Mulloy B, Thomas AH, eds. Methods in molecular biology. New Jersey: Humana Press; 1994. p. 39–60.
6. Perkins SJ, Ashton AW, Boehm MK, Chamberlain D. Molecular structures from low angle X-ray and neutron scattering studies.
Int J Biol Macromol 1998;**22**:1–16.
7. Glatter O, Kratky O, eds. Small-angle X-ray scattering. New York: Academic Press; 1982.
8. Chothia C, Hubbard T, Brenner S, Barns H, Murzin A. Protein folds in the all-β and all-α classes.
Annu Rev Biophys Biomol Struct 1997;**26**: 597–627.
9. Lindgård P-A, Bohr H. How many protein fold classes are to be found? In: Bohr H, Brunak S, eds. Protein folds: A distance based approach. Boca Raton: CRC Press; 1996. p. 98–102.
10. Chothia C. One thousand families for the molecular biologist.
Nature 1992;**357**:543–544.
11. Haris PI, Chapman D. Analysis of polypeptide and protein structures using Fourier transform infrared spectroscopy. In: Jones C, Mulloy B, Thomas AH, eds. Methods in molecular biology. New Jersey: Humana Press; 1994. p. 183–202.
12. Drake AF. Circular dichroism. In: Jones C, Mulloy B, Thomas AH, eds. Methods in molecular biology. New Jersey: Humana Press; 1994. p. 219–244.
13. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions.
Annu Rev Biophys Biomol Struct 1996;**25**: 113–136.
14. Perkins SJ, Smith KF, Williams SC, Haris PI, Chapman D, Sim RB. The secondary structure of the von Willebrand domain in factor B of human complement by Fourier transform infrared spectroscopy: its occurrence in collagen types VI, VII, XII and XIV, the integrins and other proteins by averaged secondary structure predictions.
J Mol Biol 1994;**238**:104–119.

15. Edwards YJK, Perkins SJ. The protein fold of the von Willebrand Factor type A domain is predicted to be similar to the open twisted β-sheet flanked by α-helices found in human ras-p21.
FEBS Lett 1995;**358**:283–286.
16. Edwards YJK, Perkins SJ. Assessment of protein fold predictions from sequence information: the predicted α/β doubly wound fold of the von Willebrand Factor Type A domain is similar to its crystal structure.
J Mol Biol 1996;**260**:277–285.
17. Jones D, Taylor WR, Thornton JM. A new approach to protein fold recognition.
Nature 1992;**358**:86–89.
18. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families.
Protein Eng 1993;**6**:485–500.
19. Brissett NC, Perkins SJ. The protein fold of the hyaluronate-binding proteoglycan tandem repeat domain of link protein, aggrecan and CD44 is similar to that of the C-type lectin superfamily.
FEBS Lett 1996;**388**:211–216.
20. Lee JO, Rieu P, Arnaout MA, Liddington R. Crystal structure of the A domain from the α subunit of integrin CR3 (CD11b/CD18).
Cell 1995;**80**:631–638.
21. Jenkins PV, Pasi KJ, Perkins SJ. Molecular modelling of ligand and mutation sites of the Type A domains of human von Willebrand factor and their relevance to von Willebrand's disease.
Blood 1998;**91**:2032–2044.
22. Brissett NC, Perkins SJ. Conserved basic residues in the C-type lectin and short complement repeat domains of the G3 region of proteoglycans.
Biochem J 1998;**329**:415–424.
23. Brissett NC, Perkins SJ. Molecular modelling analyses of the C-type lectin domain in human aggrecan.
Biochem Soc Trans 1996;**24**:99S.
24. Perkins SJ, Nealis AS, Dudhia J, Hardingham TE. Immunoglobin fold and tandem repeat structures in proteoglycan N-terminal domains and link protein.
J Mol Biol 1989;**206**:737–754.
25. Kohda D, et al. Solution structure of the link module: a hyaluronan-binding domain involved in extracellular matrix stability and cell migration.
Cell 1996;**86**:767–775.
26. Ullman CG, Perkins SJ. The factor I and follistatin domain families: the return of a prodigal son.
Biochem J 1997;**326**:939–941.
27. Ullman CG, Haris PI, Smith KF, Sim RB, Emery VC, Perkins SJ. β-Sheet secondary structure of an LDL receptor domain from complement factor I by consensus structure predictions and spectroscopy.
FEBS Lett 1995;**371**:199–203.
28. Hohenester E, Maurer P, Timpl R. Crystal structure of a pair of follistatin-like and EF-hand calcium-binding domains in BM-40.
EMBO J 1997;**16**:3778–3786.
29. Chamberlain D, Ullman CG, Perkins SJ. Possible arrangement of the five domains in human complement factor I by a combination of X-ray and neutron scattering and homology modelling. (Submitted).
30. Perkins SJ. Protein volumes and hydration effects: the calculation of partial specific volumes, neutron scattering matchpoints and 280 nm absorption coefficients for proteins and glycoproteins from amino acid sequences.
Eur J Biochem 1986;**157**:169–180.
31. Smith KF, Harrison RA, Perkins SJ. Structural comparisons of the native and reaction centre cleaved forms of α_1-antitrypsin by neutron and X-ray solution scattering.
Biochem J 1990;**267**:203–212.
32. Perkins SJ, Smith KF, Kilpatrick JM, Volanakis JE, Sim RB. Modelling of the serine protease fold by X-ray and neutron scattering and sedimentation analyses: its occurrence in factor D of the complement system.
Biochem J 1993;**295**:87–99.
33. Ashton AW, Boehm MK, Gallimore JR, Pepys MB, Perkins SJ. Pentameric and decameric structures in solution of the serum amyloid P component by X-ray and neutron scattering and molecular modelling analyses.
J Mol Biol 1997;**272**:408–422.
34. Harris LJ, Larson SB, Hasel KW, McPherson A. Refined structure of an intact IgG2a monoclonal antibody.
Biochemistry 1997;**36**:1581–1597.
35. Mayans MO, Coadwell WJ, Beale D, Symons DBA, Perkins SJ. Demonstration by pulsed neutron scattering that the arrangement of the Fab and Fc fragments in the overall structures of bovine IgG1 and IgG2 in solution is similar.
Biochem J 1995;**311**:283–291.
36. Sutton BJ, Gould HJ. The human IgE network.
Nature 1993;**366**:421–428.
37. Beavil AJ, Young RJ, Sutton BJ, Perkins SJ. Bent domain structure of recombinant human IgE-Fc in solution by X-ray and neutron scattering in conjunction with an automated curve fitting procedure.
Biochemistry 1995;**34**:14449–14461.
38. Zheng Y, Shopes B, Holowka D, Baird B. Conformations of IgE bound to its receptor FcεRI and in solution.
Biochemistry 1991;**30**:9125–9132.
39. Perkins SJ, Nealis AS, Sutton BJ, Feinstein A. The solution structure of human and mouse immunoglobulin IgM by synchrotron X-ray scattering and molecular graphics modelling: a possible mechanism for complement activation.
J Mol Biol 1991;**221**:1345–1366.
40. Arya S, Chen F, Spycher S, Isenman DE, Shulman MJ, Painter RH. Mapping of amino acid residues in the Cμ3 domain of mouse IgM important in macromolecular assembly and complement-dependent cytolysis.
J Immunol 1994;**152**:1206–1212.
41. Thompson JA, Grunert F, Zimmerman WJ. Carcinoembryonic antigen gene family: molecular biology and clinical perspectives.
Clin Lab Anal 1991;**5**:344–366.
42. Boehm MK, Mayans MO, Thornton JD, Begent RHJ, Keep PA, Perkins SJ. Extended glycoprotein structure of the seven domains in human carcinoembryonic antigen by X-ray and neutron solution scattering and an automated curve fitting procedure: implications for cellular adhesion.
J Mol Biol 1996;**259**:718–736.
43. Law SKA, Reid KBM. Complement. 2nd ed. Oxford: IRL Press; 1995.
44. Boehm MK, et al. Crystal and solution structure of an anti-carcinoembryonic antigen single-chain Fv MFE-23 by X-ray crystallography and neutron scattering. (In preparation).
45. Raag R, Whitlow M. Single-chain Fvs.
FASEB J 1995;**9**:73–80.
46. Pepys MB, Booth DR, Hutchinson WL, Gallimore JR, Collins PM, Hohenester E. Amyloid P component: a critical review.
Amyloid Int J Exp Clin Invest 1997;**4**:274–295.
47. Russell RB, Sternberg MJE. How good are we?
Curr Biol 1995;**5**:488–490.
48. Springer TA. Folding of the N-terminal, ligand-binding region of integrin α-subunits into a β-propeller domain.
Proc Natl Acad Sci USA 1997;**94**:65–72.
49. Tozer EC, Liddington RC, Sutcliffe MJ, Smeeton AH, Loftus JC. Ligand binding to integrin αIIbβ3 is dependent on a MIDAS-like domain in the β3 subunit.
J Biol Chem 1996;**271**:21978–21984.
50. Tuckwell DS, Humphries MJ. A structure prediction for the ligand-binding region of the integrin β subunit: evidence for the presence of a von Willebrand factor A domain.
FEBS Lett 1997;**400**:297–303.