EXPRESSION AND STRUCTURAL STUDIES OF MULTIDOMAIN PROTEINS AND COMPLEXES

Thesis Presented for the Degree of
Doctor of Philosophy
By
Dean Chamberlain

Department of Biochemistry and Molecular Biology
Royal Free Hospital School of Medicine
University of London

May 1998

1999

ACKNOWLEDGEMENTS.

I would like to thank my supervisor Dr S. J. Perkins for his help and encouragement over the past three and a half years with the work presented in this thesis.

I would also like to thank Dr C. G. Ullman for his help with the factor I work (in particular the biochemical characterization) in sections 6.2.1-6.2.6 and 7.2.1. The experimental procedures detailed in sections 6.2.3, 6.2.5, 6.2.6, and the free cysteine assay in section 7.2.1 were performed by Dr C. G. Ullman. I also thank Mr A. W. Ashton and Mr M. K. Boehm for their help with some of the script programming used in this work.

I would also like to thank the following people from the Department of Biochemistry and Molecular Biology, University College London, in particular Drs L. H. Pearl, B. P. O'Hara, and S. A. Wilson for the collaboration with the AmiC work presented in Chapter 5, and Drs I. Tsaneva, M. Oram and A. Keeley for the collaboration with the RuvA work presented in Chapter 8, performing the work presented in sections 8.2.1 and 8.2.2 and assistance with the absorbance coefficient determination presented in section 8.2.3. I also thank Dr R. B. Sim at the MRC Immunochemistry Unit, Oxford for his support for the Factor I studies.

I also thank the Biotechnology and Biological Sciences Research Council for an Earmarked studentship, and the Engineering and Physical Sciences Research Council for access to X-ray and neutron facilities. I would also like to thank 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.

I would also like to thank my friends and family for their encouragement and support.

ABSTRACT.

It is generally accepted that there is a level of organization in proteins that overlaps the classical definitions of tertiary and quaternary structure, i.e. sequentially consecutive residues in polypeptide chains fold into distinct compact regions called domains. Many multidomain proteins are flexible and are not amenable to X-ray crystallography or are too big for multi-dimensional nuclear magnetic resonance techniques, while other proteins form oligomeric structures from subunits. It is possible using small-angle X-ray and neutron scattering, coupled with molecular modelling techniques, to locate the relative positions of these domains or subunits relative to each other within the full protein structure.

This PhD thesis has looked at a variety of native and recombinant oligomeric proteins and domains and attempts have been made to produce low resolution structures of their oligomerisation or their multidomain structures. Expression systems used include a *Pseudomonas aeruginosa* overexpression system and the baculovirus expression system.

One multidomain protein was studied, namely factor I of the complement system. Two forms of factor I were studied, a native form purified from human plasma, and a recombinant form produced in insect cells. Scattering modelling was used to elucidate a bilobal domain arrangement in factor I, in which the different types of carbohydrate present on the two different forms could be modelled.

The quaternary structures of two complexes were determined, namely the homo-oligomeric complexes of the *Ps. aeruginosa* amidase regulatory protein, AmiC, and the *Mycobacterium leprae* Holliday junction protein, RuvA. It was determined that in solution AmiC exists as a monomer-trimer equilibrium, and that RuvA adopts an octameric structure, both when free and when complexed with DNA, within which the Holliday junction is buried in the RuvA-DNA complex.

CONTENTS PAGE

CHAPTER 3

CHAPTER 4

CONTENTS: FIGURES PAGE

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CONTENTS: TABLES PAGE

CHAPTER 8

CHAPTER 9

LIST OF ABBREVIATIONS

2D-NMR	two dimensional nuclear magnetic resonance
APS	ammonium persulphate
BSA	bovine serum albumin
CCP	complement control protein
CD	circular dichroism
DEAE	diethylaminoethyl
DNA	deoxyribonucleic acid
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
EGF	epidermal growth factor
FB	factor B
FI	factor I
FIM	factor I module
FIMAC	factor I module/membrane attack complex
FPLC	fast performance liquid chromatography
FTIR	Fourier transform infra red
GAL	galactose
GST	glutathione S-transferase
HPLC	high performance liquid chromatography
HTH	helix-turn-helix
LB	Luria Bertani
LDLr	low density lipoprotein receptor
MHC	major histocompatibility complex
mRNA	messenger RNA
OD	optical density
ORF	open reading frame
PBS	Dulbecco's 'A' phosphate buffered saline
PCR	polymerase chain reaction
PMSF	phenylmethylsulphonylfluoride
rFI	recombinant factor I
RNA	ribonucleic acid
SAS	small angle scattering
SANS	small angle neutron scattering
SAXS	small angle X-ray scattering
SCR	short consensus repeat (synonym of CCP)
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SDW	sterile deionized and distilled water
sFI	serum factor I
SH2, SH3	src-homology 2, (and 3)
SP	serine protease
SV	simian virus
UV	ultraviolet
vWF	von Willebrand factor

AMINO ACID ABBREVIATIONS

Amino acid	3 letter format	1 letter format
alanine	Ala	A
arginine	Arg	R
aspartic acid	Asp	D
asparagine	Asn	N
cysteine	Cys	C
glutamic acid	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

NUCLEOTIDE ABBREVIATIONS

adenine	A
guanine	G
thymidine	T
cytosine	C

CHAPTER 1

THE DOMAIN NATURE OF PROTEINS

(1.1) Introduction.

It is generally accepted that there is a level of organization in proteins that overlaps the classical definitions of tertiary and quaternary structure, i.e. sequentially consecutive residues in polypeptide chains tend to fold into distinct compact regions called domains (Wetlaufer, 1973). In this respect, a domain can be defined as that part of a protein that can fold up independently of neighbouring sequences (Doolittle, 1995).

There have been varying estimates of what the total number of different domains will be. Cyrus Chothia has predicted from the number of families that are currently known and the present rate at which "superfamilies" are being found, that there may only be ~1000 different types of protein fold or domain, and that examples of all of them will be found by 2015 (Chothia, 1992). It is thought that the currently known domains originated from 20 major ancestral types. These can be further reduced to two or three all-alpha domains, a few all-beta domains, and a small number of alpha-beta domains (Doolittle, 1995).

In nearly all proteins, the local folding of the polypeptide chain leads to the formation of α-helices or β-sheets, and these assemble to give the molecules their globular three-dimensional structures. The protein fold describes three major aspects of its three-dimensional structure: the secondary structures of which it is composed, their relative arrangement, and the path taken through the structure by the polypeptide chain (Chothia and Finkelstein 1990). The basis of domain organization is the cohesion between side chains that stabilises unique structures. There is a need for a critical number of residues in a sequence, typically 50-100 (some domains can be much larger), before a domain can be realised, although smaller segments (e.g. Gla domains and short consensus repeat (SCR) domains) can be further stabilised by folding around metal centres or by the formation of disulphide bonds, or in the case of short repetitive units, packing against each other. The average polypeptide chain length for most proteins or their consistent units is ~350 residues, and most proteins contain two or more domains, although many proteins consist of only one domain (Doolittle, 1995).

(1.2) **Identifying Domains.**

The domains in a protein may be packed together tightly or loosely. If the packing is loose, domains may be separated out by limited proteolysis if a suitable protease site is present in the link region. Scanning calorimetry has been used to distinguish domains because of their differences in thermostability. A differential scanning calorimeter is used to measure the comparative difference in the absorption of heat as the temperature is raised between a solution containing a protein sample and an identical solution which lacks the protein. Two cells, sample and reference, contain precisely matched coils that introduce identical quantities of heat into each cell and establish a constant rate of temperature increase. The sample cell has an auxiliary coil that provides additional heat to keep its temperature exactly the same as the reference cell. The power supplied to the auxiliary coil is a measure of the excess heat absorbed by the sample. As the temperature rises a protein will unfold and this transition proceeds with the absorption of heat. This heat absorption is a convenient way to follow the progress of the protein unfolding (Kyte, 1995). Donavan and Milhalyi (1974) used calorimetry to show that fibrinogen had two transitions occurring at different temperatures. These transitions were assigned to domains D and E.

The most accurate way to distinguish domains is by direct observation of the three-dimensional atomic structures obtained by NMR or X-ray crystallography. The number of solved structures is small compared to the number of available sequences. Family relationships between sequences are being discovered repeatedly in database searches. The level of relationships can often be identified at the amino acid sequence similarity, e.g. the immunoglobulin family (Williams and Barclay, 1988; Jones, 1993), but sometimes is only obvious at the three-dimensional structure level, e.g. the ß-trefoil family (Murzin *et. al.*, 1992).

It has been established that many proteins in multicellular organisms are made from combinations of several clearly identifiable, autonomously folding domains. The databases of protein sequences and protein structures are growing at a remarkable rate. The various genome sequencing projects around the world are also turning up possible exon sequences which when assigned a reading frame can be assigned amino acid sequences (subject to the correct codon usage tables). There are numerous computer programs available for searching an amino acid sequence against databases in an effort to find related sequences. The most popular program is BLAST (Altschul *et al.*, 1990), which replaced FastA (Pearson and Lipman, 1988) as the most frequently used routine. BLAST can identify matching segments and is able to perform 4000 comparisons per second (depending on the hardware platform).

It is expected that there will be limits to the effectiveness of relationships detected using amino acid sequences alone. Sequences change more or less stochastically, and eventually a point will be reached where not enough information remains to distinguish genuine homology from chance similarity. This is especially true in the case of smaller domains composed of fewer than a hundred residues, which includes many of the most common domains. Family memberships will in general be undercounted, and distant relationships may have to be established exclusively by three-dimensional structures (Doolittle, 1995).

(1.3) <u>Domains at the DNA Level.</u>

Not all exons encode domains. The belief that "one exon = one domain" has been encouraged by the notion that primitive proteins were encoded by "minigenes" that were spliced together, the genome eventually maturing into a state where coding regions were separated by introns (Seidel *et al.*, 1992). The "introns early" scenario (Gilbert and Glynias, 1993) assumes that the prokaryotic lineage has lost all (coding region) introns over the course of time as a secondary adaption in line with streamlining the genome or quickening replication times (Doolittle, 1978). The "introns late" scenario assumes that the class of introns that interrupt protein coding (or structural gene) regions are a late occurrence which made their impact after the occasion of the major endosymbiotic events that led to the appearance of mitochondria (Doolittle, 1995).

Studies have been undertaken to try and deduce intron positions in contemporary protein coding regions that may correlate with rudimentary structures (Doolittle, 1995). Early studies in the mid 1980s were vague about what exactly constituted basic structural elements. Two later efforts used a more rigorous approach in which observed intron distributions were measured against what would be expected by chance (Weber and Kabsch, 1994; Stolzfus *et al*, 1994). Both groups came to the conclusion that in ancient proteins there is little or no correlation of present-day intron positions with recognized elements of protein structure. This does not mean that introns do not "help" to move defined domains throughout the genome. It has been seen in animal proteins which have arisen in the last billion years that introns may have been involved (Gilbert 1978; Doolittle 1985). It may be the case that the enormous success of the metazoan radiation (the evolution and spread of multicellular organisms) is the direct result of the introduction of introns into coding regions. It still must be emphasized that most exons do not encode domains, and that many moveable domains are in fact themselves interrupted by introns. This means that for structural integrity the full complement of exons need to be shuffled as a unit (Doolittle, 1995).

The genes of some animal proteins seem to fit the concept of exon shuffling where as others do not (see Table 1.1 for examples). It is a general rule that proteins with intron-delimited domains have arisen more recently. Beyond that the discordance in other instances may be attributed to two scenarios (Doolittle, 1995):-

1. Protein may have been assembled by exon shuffling, but subsequent intron loss and intron gain has obscured the relationship. Several of the proteins in Table 1.1 may fall into this category.

2. Protein may have been assembled in a modular fashion without the involvement of introns. Trypsinogen has two well-defined and similarly folded domains but none of its four introns occur at the obvious joining point (Craik *et al.*, 1984). It is possible that a simple contiguous duplication gave rise to the structure and at a later point the introns

were introduced.

DNA replication is subject to a number of iterative errors and chance duplications. Duplications can range from a few base pairs to entire chromosomes and even genomes. It was reported recently that there is evidence of an entire genome duplication in *Saccharomyces cerevisiae* (Wolfe and Shields, 1997). Various kinds of recombination, homologous (legitimate) or illegitimate, occur as the result of random or non-random breakage and reunion of DNA. Homologous recombination is the result of mistaken mismatching of similar DNA sequences, and as a result, there is a tendency for duplication to cause more duplication (Doolittle, 1995). The more recombination occurs at a given locus, the more opportunities there will be for mismatch. In compensation, intron sequences drift away much more rapidly than coding sequences, thereby diminishing opportunities for homologous recombination. Introns, however, can provide safe havens for transposable elements or highly repetitive sequences like the Alu family. These in turn can become involved in mismatches that give arise to rearrangements (Stoppa-Lyonnet *et al.*, 1990). Additionally, there is a good deal of rearrangement by more obscure mechanisms that are described as "illegitimate recombination" (Doolittle, 1995). There is an additional constraint imposed by the necessity for shuffled exons to have compatible introns so that neighbouring domains would not be put out of register (Patthy, 1987).

(1.4) Distribution of Domains.

It is assumed that there will be some proteins common to all organisms ("ancient proteins") and that these proteins were in existence before the last common ancestor

Table 1.1: Correlations or lack of correlations of animal protein domains with exons (based on Doolittle 1995).

Good correlation with exons	Reference
tissue plasminogen activator	Ny *et al.*, 1984
LDL-receptor	Sudhof *et al.*, 1985
von Willebrand factor	Mancuso *et al.*, 1989
protein S (vitronectin)	Jenne and Stanley, 1987
link protein	Kiss *et al.*, 1987
fibronectin	Patel *et al.*, 1987
factor XIII b chain	Bottenus *et al.*, 1990
homing receptor	Dowbenko *et al.*, 1991
lipoprotein-associated coagulation inhibitor	van der Logt *et al.*, 1991

Poor or no correlation with exons	Reference
trypsinogen	Craik *et al.*, 1984
serpins	Wright, 1993
complement C6	Hobart *et al.*, 1993
complement C3-C5	Vik *et al.*, 1991
α spectrin	Kotula *et al.*, 1991
preadipocyte EGF-like protein-1	Smas *et al.*, 1994
notch	Wharton *et al*, 1985
lin 12	Greenwald 1985

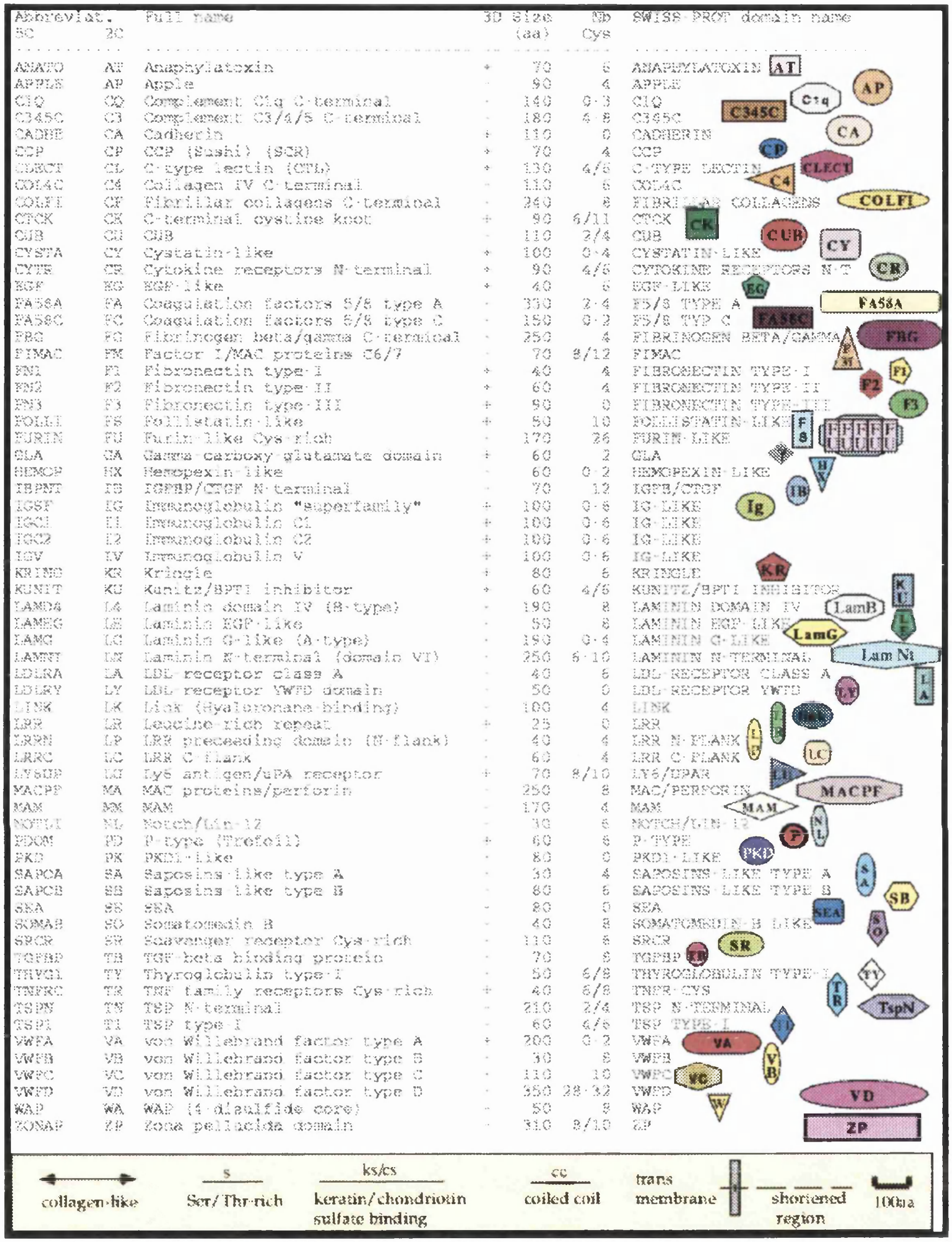
(Doolittle, 1995). Such domains include the nucleotide-binding fold (Rossmann *et al.*, 1974), flavin binding fold (Correll *et al.*, 1993), and the haem-binding domains (Vasudevan *et al.*, 1993). Most of these domains are central to metabolic processes, and the evidence of their ancient mobility is only apparent when three-dimensional structures are compared (Smith *et al.*, 1994). In contrast, many non-catalytic domains still move around in various positions peripheral to catalytic domains (Figure 1.6) (Doolittle, 1995).

Biological processes consisting of proteins made of various domains (Figure 1.1) include the complement system (Figure 1.2, Figure 1.3, Table 1.2: note the differences in nomenclature for the same domains between Figures 1.2 and 1.3), the blood clotting system and the fibrinolysis system (Figure 1.4). Figures 1.5 and 1.6 show other examples.

It is important to stress that the evolutionary power of domain shuffling should not be exaggerated. The combinatorial advantage is significant but it is not unlimited (Doolittle, 1995). There are lots of examples of domains and proteins evolving independently on more than one occasion, these include (Doolittle, 1994):-

Superoxide dismutases
Sugar kinases
Alcohol dehydrogenases
Ribonucleotide reductases
PEP carboxykinases

Aldolases
Serine proteases
Aminoacyl tRNA synthases
Topoisomerases
Malate dehydrogenases



Abbreviat. 5C	2C	Full name	3D	Size (aa)	Nb Cys	SWISS-PROT domain name
ANATO	AT	Anaphylatoxin	+	70	6	ANAPHYLATOXIN
APPLE	AP	Apple	-	90	4	APPLE
C1Q	CQ	Complement C1q C-terminal	-	140	0-3	C1Q
C345C	C3	Complement C3/4/5 C-terminal	-	180	4-8	C345C
CADHE	CA	Cadherin	+	110	0	CADHERIN
CCP	CP	CCP (Sushi) (SCR)	+	70	4	CCP
CLECT	CL	C-type lectin (CTL)	+	130	4/6	C-TYPE LECTIN
COL4C	C4	Collagen IV C-terminal	-	110	6	COL4C
COLFI	CF	Fibrillar collagens C-terminal	-	240	8	FIBRILLAR COLLAGENS
CTCK	CK	C-terminal cystine knot	+	90	6/11	CTCK
CUB	CU	CUB	-	110	2/4	CUB
CYSTA	CY	Cystatin-like	+	100	0-4	CYSTATIN-LIKE
CYTR	CR	Cytokine receptors N-terminal	+	90	4/6	CYTOKINE RECEPTORS N-T
EGF	EG	EGF-like	+	40	6	EGF-LIKE
FA58A	FA	Coagulation factors 5/8 type A	-	330	2-4	F5/8 TYPE A
FA58C	FC	Coagulation factors 5/8 type C	-	150	0-2	F5/8 TYP C
FBG	FG	Fibrinogen beta/gamma C-terminal	-	250	4	FIBRINOGEN BETA/GAMMA
FIMAC	FM	Factor I/MAC proteins C6/7	-	70	8/12	FIMAC
FN1	F1	Fibronectin type-I	+	40	4	FIBRONECTIN TYPE-I
FN2	F2	Fibronectin type-II	+	60	4	FIBRONECTIN TYPE-II
FN3	F3	Fibronectin type-III	+	90	0	FIBRONECTIN TYPE-III
FOLLI	FS	Follistatin-like	+	50	10	FOLLISTATIN-LIKE
FURIN	FU	Furin-like Cys-rich	-	170	26	FURIN-LIKE
GLA	GA	Gamma-carboxy glutamate domain	+	60	2	GLA
HEMOP	HX	Hemopexin-like	-	60	0-2	HEMOPEXIN-LIKE
IBPNT	IB	IGFBP/CTGF N-terminal	-	70	12	IGFB/CTGF
IGSF	IG	Immunoglobulin "superfamily"	+	100	0-6	IG-LIKE
IGC1	I1	Immunoglobulin C1	+	100	0-6	IG-LIKE
IGC2	I2	Immunoglobulin C2	+	100	0-6	IG-LIKE
IGV	IV	Immunoglobulin V	+	100	0-6	IG-LIKE
KRING	KR	Kringle	+	80	6	KRINGLE
KUNIT	KU	Kunitz/BPTI inhibitor	+	60	4/6	KUNITZ/BPTI INHIBITOR
LAMD4	L4	Laminin domain IV (B-type)	-	190	8	LAMININ DOMAIN IV
LAMEG	LE	Laminin EGF-like	-	50	8	LAMININ EGF-LIKE
LAMG	LG	Laminin G-like (A-type)	-	190	0-4	LAMININ G-LIKE
LAMNT	LN	Laminin N-terminal (domain VI)	-	250	6-10	LAMININ N-TERMINAL
LDLRA	LA	LDL-receptor class A	-	40	6	LDL-RECEPTOR CLASS A
LDLRY	LY	LDL-receptor YWTD domain	-	50	0	LDL-RECEPTOR YWTD
LINK	LK	Link (Hyaluronane-binding)	-	100	4	LINK
LRR	LR	Leucine-rich repeat	+	25	0	LRR
LRRN	LP	LRR preceeding domain (N-flank)	-	40	4	LRR N-FLANK
LRRC	LC	LRR C-flank	-	60	4	LRR C-FLANK
LY6UP	LU	Ly6 antigen/uPA receptor	+	70	8/10	LY6/UPAR
MACPF	MA	MAC proteins/perforin	-	250	8	MAC/PERFORIN
MAM	MM	MAM	-	170	4	MAM
NOTLI	NL	Notch/Lin-12	-	30	6	NOTCH/LIN-12
PDOM	PD	P-type (Trefoil)	+	60	6	P-TYPE
PKD	PK	PKD1-like	-	80	0	PKD1-LIKE
SAPOA	SA	Saposins-like type A	-	30	4	SAPOSINS-LIKE TYPE A
SAPOB	SB	Saposins-like type B	-	80	6	SAPOSINS-LIKE TYPE B
SEA	SE	SEA	-	80	0	SEA
SOMAB	SO	Somatomedin B	-	40	8	SOMATOMEDIN-B LIKE
SRCR	SR	Scavenger receptor Cys-rich	-	110	6	SRCR
TGFBP	TB	TGF-beta binding protein	-	70	8	TGFBP
THYG1	TY	Thyroglobulin type-I	-	50	6/8	THYROGLOBULIN TYPE-I
TNFRC	TR	TNF family receptors Cys-rich	+	40	6/8	TNFR-CYS
TSPN	TN	TSP N-terminal	-	210	2/4	TSP N-TERMINAL
TSP1	T1	TSP type-I	-	60	4/6	TSP TYPE-I
VWFA	VA	von Willebrand factor type A	+	200	0-2	VWFA
VWFB	VB	von Willebrand factor type B	-	30	8	VWFB
VWFC	VC	von Willebrand factor type C	-	110	10	VWFC
VWFD	VD	von Willebrand factor type D	-	350	28-32	VWFD
WAP	WA	WAP (4-disulfide core)	-	50	8	WAP
ZONAP	ZP	Zona pellucida domain	-	310	8/10	ZP

Figure 1.1: Nomenclature and captions of Bork and Bairoch (1995) domain cartoons. (Taken from the EMBL Heidelberg World Wide Web Pages www.embl-heidelberg.de).

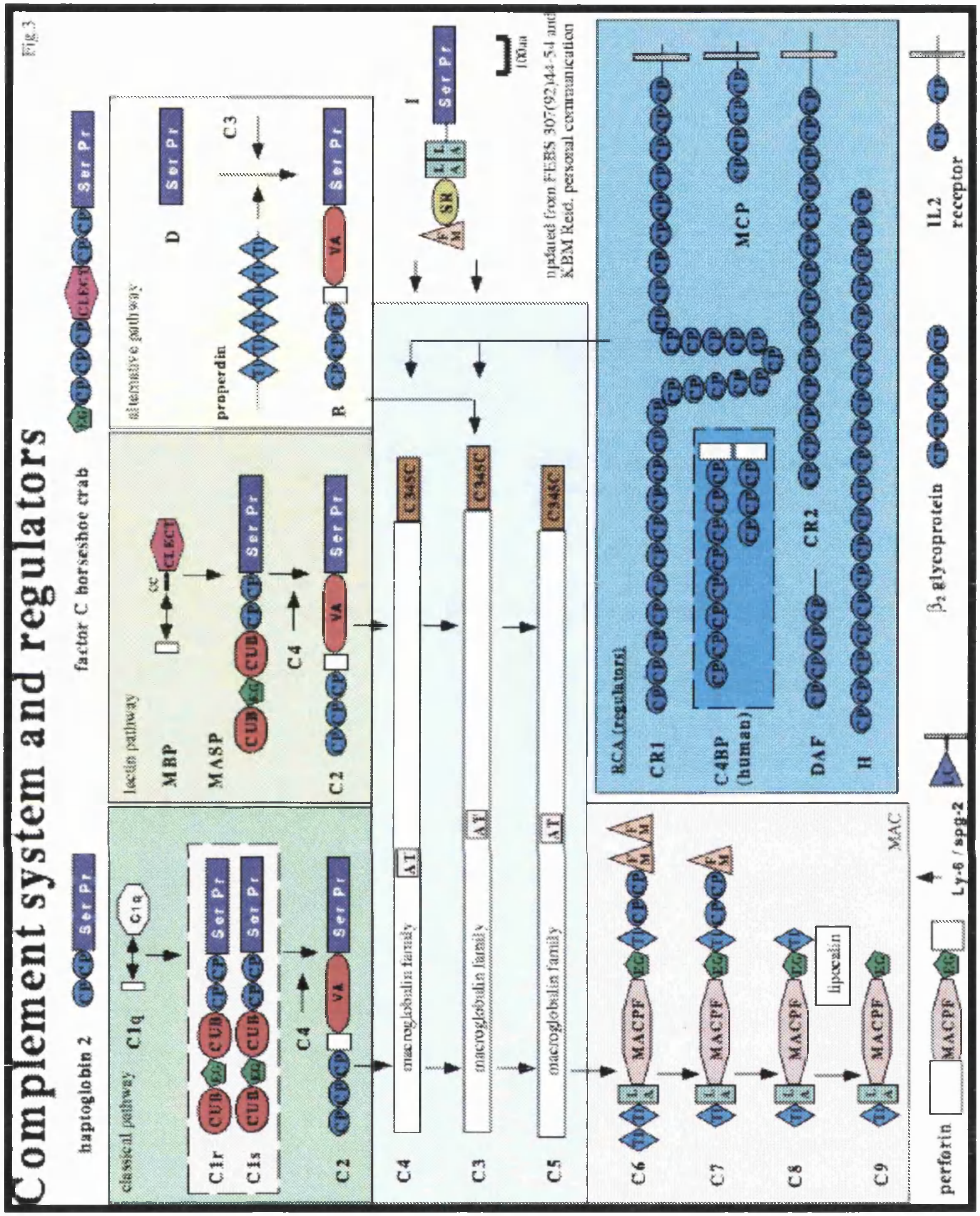


Figure 1.2: Bork and Bairoch, (1995), domain cartoons for the complement system. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

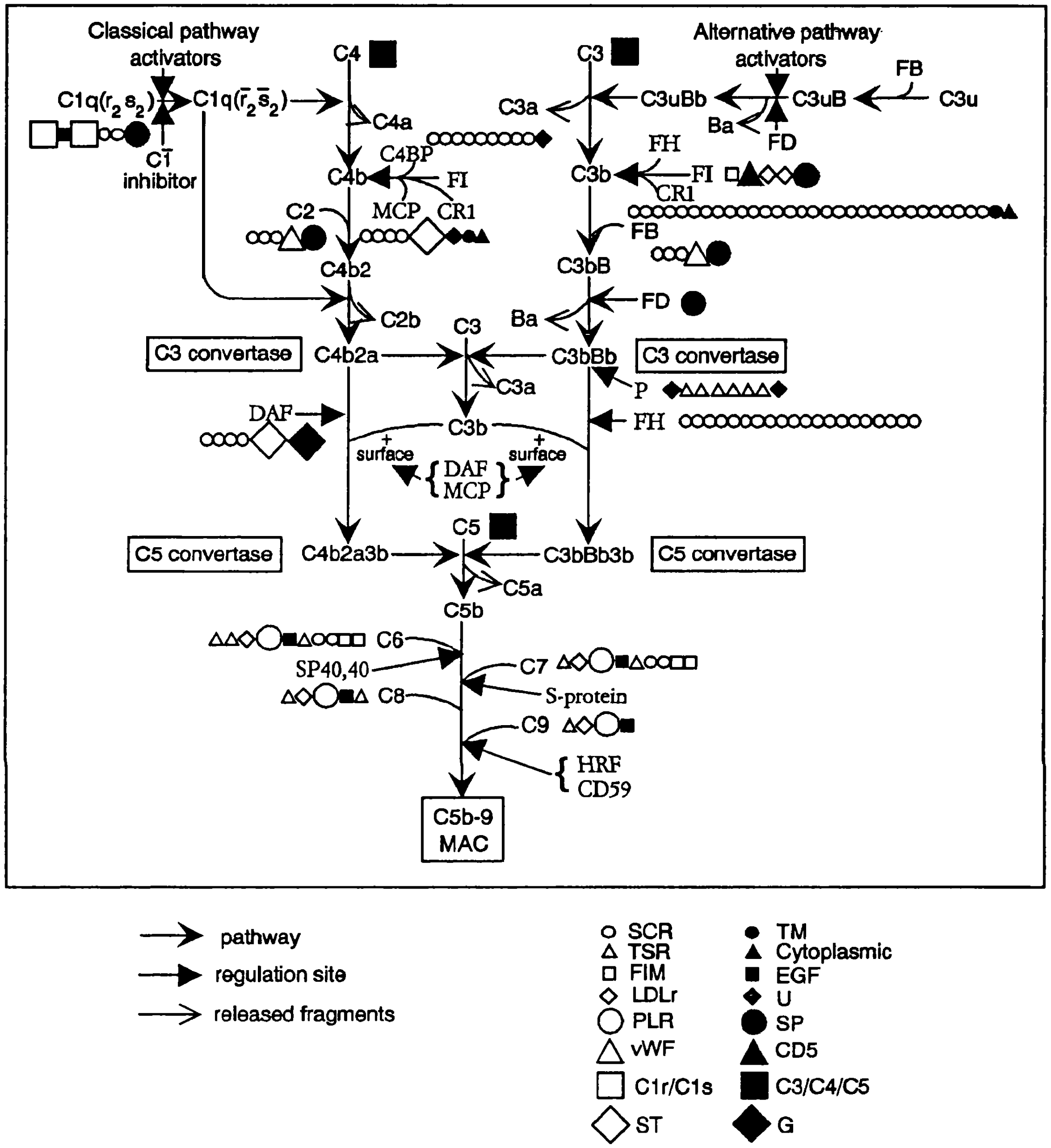


Figure 1.3: The complement cascade of proteins, showing the modular nature of the complement components, (reproduced from Smith, 1992). Abbreviations are shown in Table 1.2.

Table 1.2: Summary of the Physiological Concentrations and Domain Structures of the Complement Components. Table reproduced from Smith, 1992.

	M$_r$ (kDa)	Approx. serum conc. (mg l^{-1})	Domains
Classical Pathway			
C1q	457	80	stalks, head
C1r	172	50	2 RS, 2 SCR, EGF, SP
C1s	158	50	2 RS, 2 SCR, EGF, SP
C4	197	600	C3/C4/C5
C2	102	20	vWF, SP, 3 SCR
C3	187	1300	C3/C4/C5
Alternative Pathway			
Factor D	24	1	SP
Factor B	89	210	3 SCR, vWF, SP
Terminal Pathway			
C5	194	70	C3/C4/C5
C6	107	64	3 TSR, LDLr, PLR, EGF, 2 SCR, 2 FIM
C7	95	56	2 TSR, LDLr, PLR, EGF, 2 SCR, 2 FIM
C8	154	55	2 TSR, LDLr, PLR, EGF
C9	66	59	TSR, LDLr, PLR, EGF
Control proteins			
Plasma			
C$\overline{1}$ inhibitor	71	200	N-terminus, serpin
Factor J	20	5	
Factor I	74	35	FIM, CD5, 2 LDLr, SP
Properdin (trimer)	162	20	3 x 6 TSR
C4BP	491	250	7 x 8 SCR + 3 SCR
Factor H	150	480	20 SCR
S-Protein	83	505	
SP-40,40	70	100	coiled-coil
Carboxy-peptidase N	310	35	
Membrane bound			
MCP	45-70		4 SCR, ST, U, TM, CYT
DAF	70		4 SCR, ST, G
HRF	65		
CD59	18-20		murine LY-6 Antigen
Receptors			
CR1	160, 190 220, 250		30 SCR, TM, CYT
CR2	140		16 SCR, TM, CYT
CR3	265		vWF, 3 MB, TM, CYT
CR4	245		vWF, 3 MB, TM, CYT
C5a receptor	39		7 TM
C1q receptor	56		

Abbreviations

CYT	cytoplasmic domain	SCR	short consensus repeat
EGF	epidermal growth factor	SP	serine protease domain
G	glycolipid anchor	ST	serine/threonine-enriched area
FIM	factor I module	TM	transmembrane domain
LDLr	LDL receptor	TSR	thrombospondin repeat
MB	metal binding domain	U	unknown functional significance
PLR	perforin like region	vWF	von Willebrand Factor
RS	C1r/C1s domain		

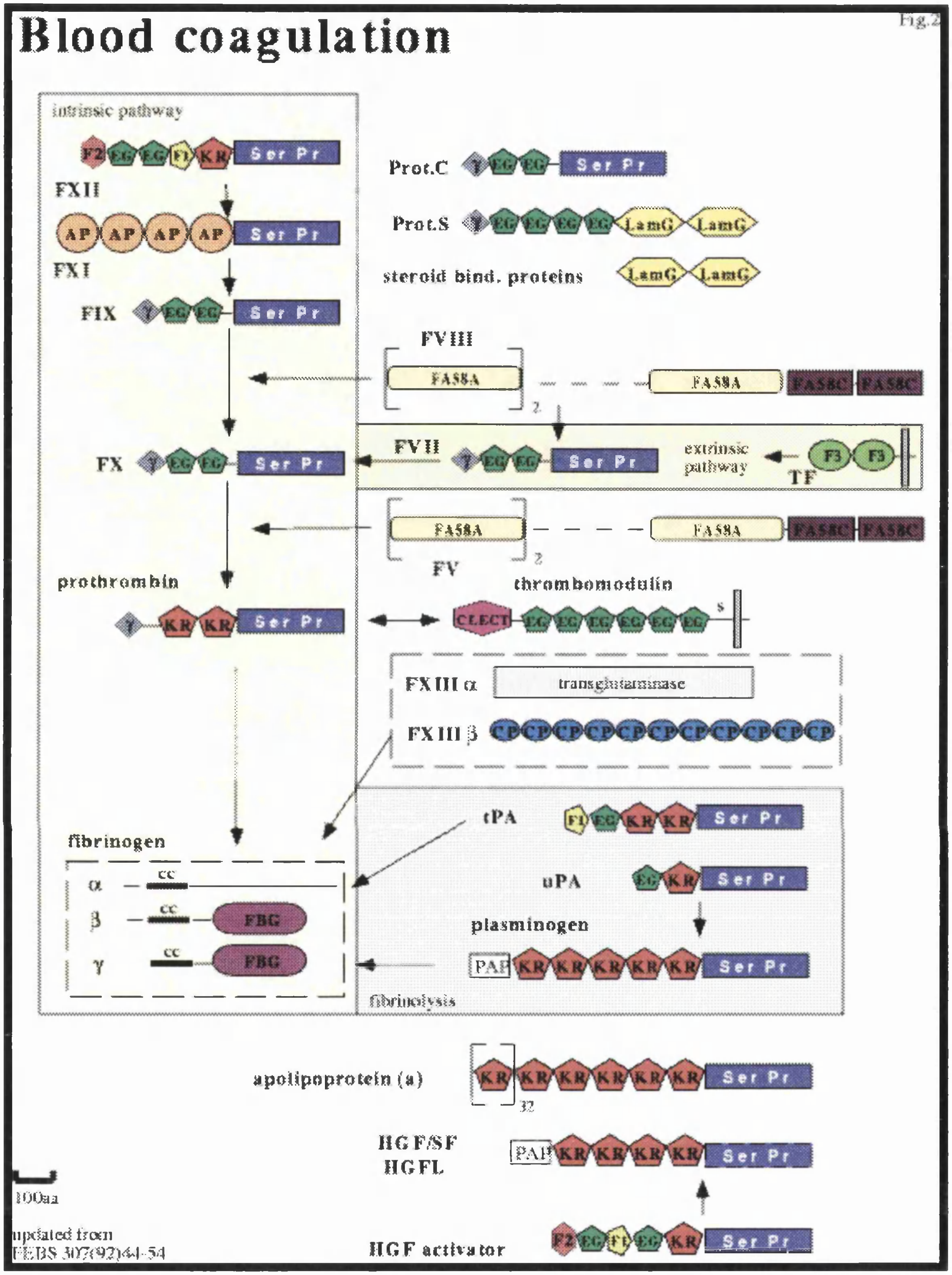


Figure 1.4: Bork and Bairoch, (1995), domain cartoons for the blood coagulation system. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

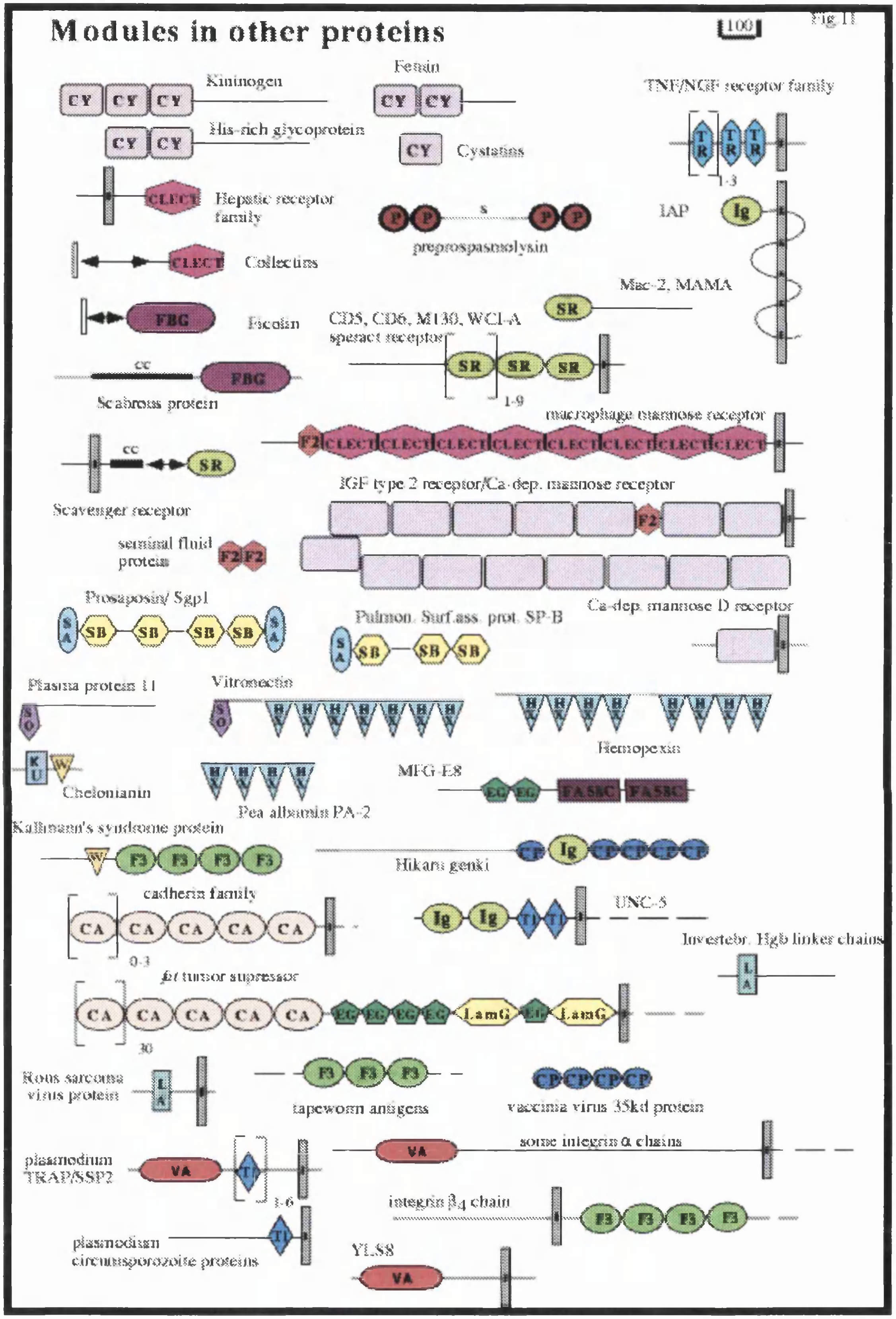


Figure 1.5: Bork and Bairoch, (1995), cartoon representations for various multidomain proteins. (Taken form the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

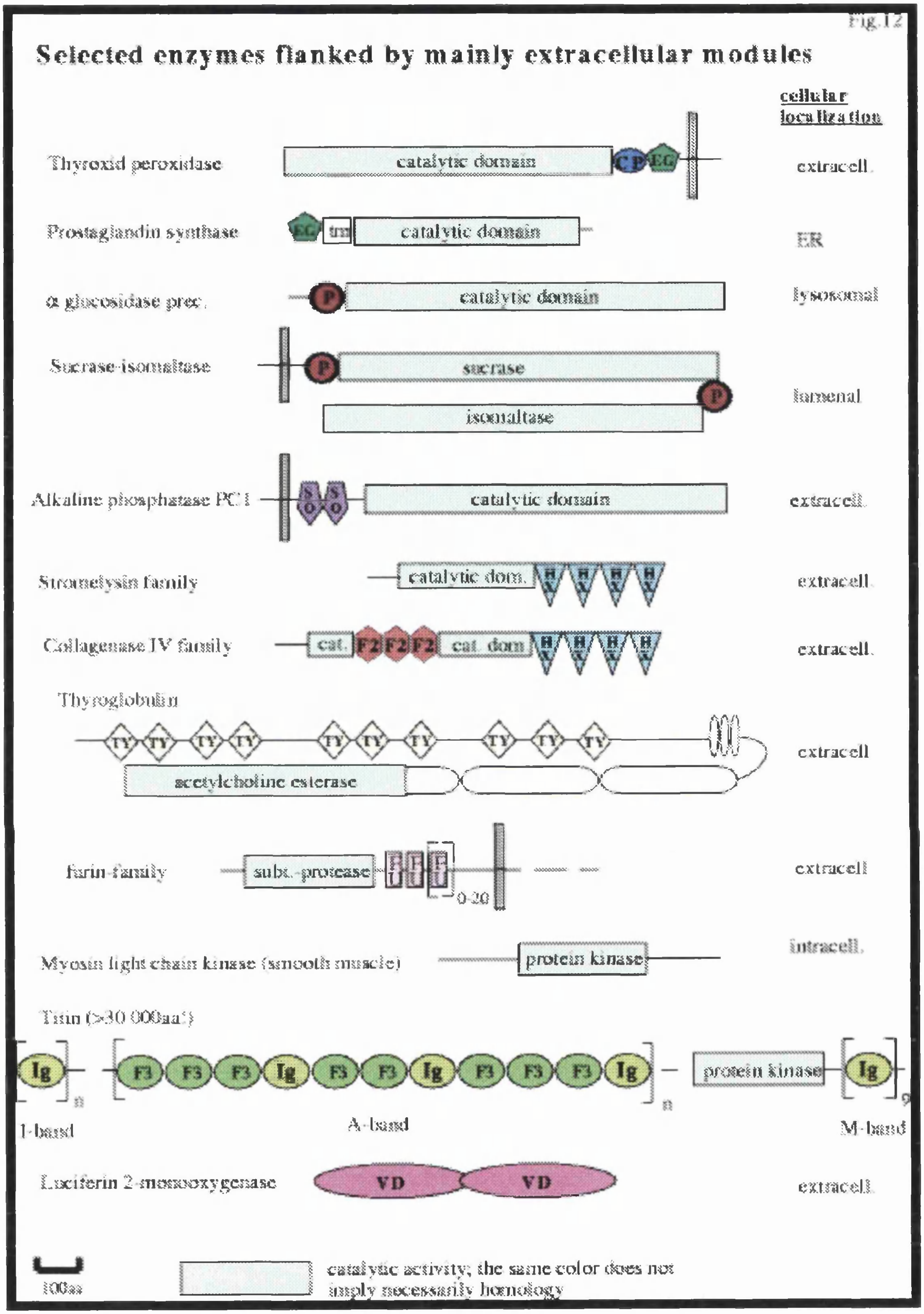


Figure 1.6: Bork and Bairoch, (1995), cartoon representations of selected enzymes flanked by defined domains. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

As well as there being domains common to prokaryotes and eukaryotes, there are also examples of domains found in prokaryotes but not eukaryotes and *vice-versa*. Each may have major lineages that have domains that were fashioned along the way. A number of hypotheses have been put forward to explain this, including frame-shifts (Keese and Gibbs, 1992) and intron capture (Golding *et al.,* 1994). Fungi, plants and animals have many sequences that appear not to have counterparts elsewhere (see below). It is possible that new domains have been devised, old ones have been selectively lost by some lineages, or in many cases relationships have been blurred by excessive sequences. If the relationships are blurred, the only way of identifying relationships would be at the three dimensional structure level as tertiary structures are better conserved than primary structures during evolution, assuming similar structures have been determined (Doolittle, 1995).

(1.4.1) Intracellular Domains Apparently Unique to Eukaryotes.

One domain with a characteristic sequence motif that has so far only been found in eukaryotes is the WD-40 domain. This domain was first found in beta-transducins from animals and cell-division proteins from yeast (Fong *et al.*, 1986), but is now known to occur in over 50 different proteins (Neer *et al.,* 1994). This small domain of 40 amino acids has a definable sequence pattern (X..GH..X..WD) that has been conserved over the billion years since fungi, plants, and animals had a common ancestor. However, its evolutionary origin remains a mystery. It has not been found in bacteria (Doolittle, 1995). The domain is usually found as a set of 4-8 repeats and is associated with a large number of functions and interactions with other proteins. Its proposed structure, modelled from similar sequences from parts of known structures (Neer *et al.,* 1994), is a small five-stranded antiparallel β-sheet.

There are other apparently critical domains that are known to be common to fungi, plants and animals, but which have not yet been traced to a prokaryotic origin. For example, the dimerisation domain that became known as the leucine zipper (Landschulz *et al.*, 1988) was first identified by sequence searching when it was found next to the carboxyl-terminal segment of the yeast GCN4 protein and vertebrate oncogenes like *jun* and *myc* (Vogt *et al.*, 1987). The sequence databases contain scores of related sequences, but they are all from eukaryotes (Doolittle, 1995).

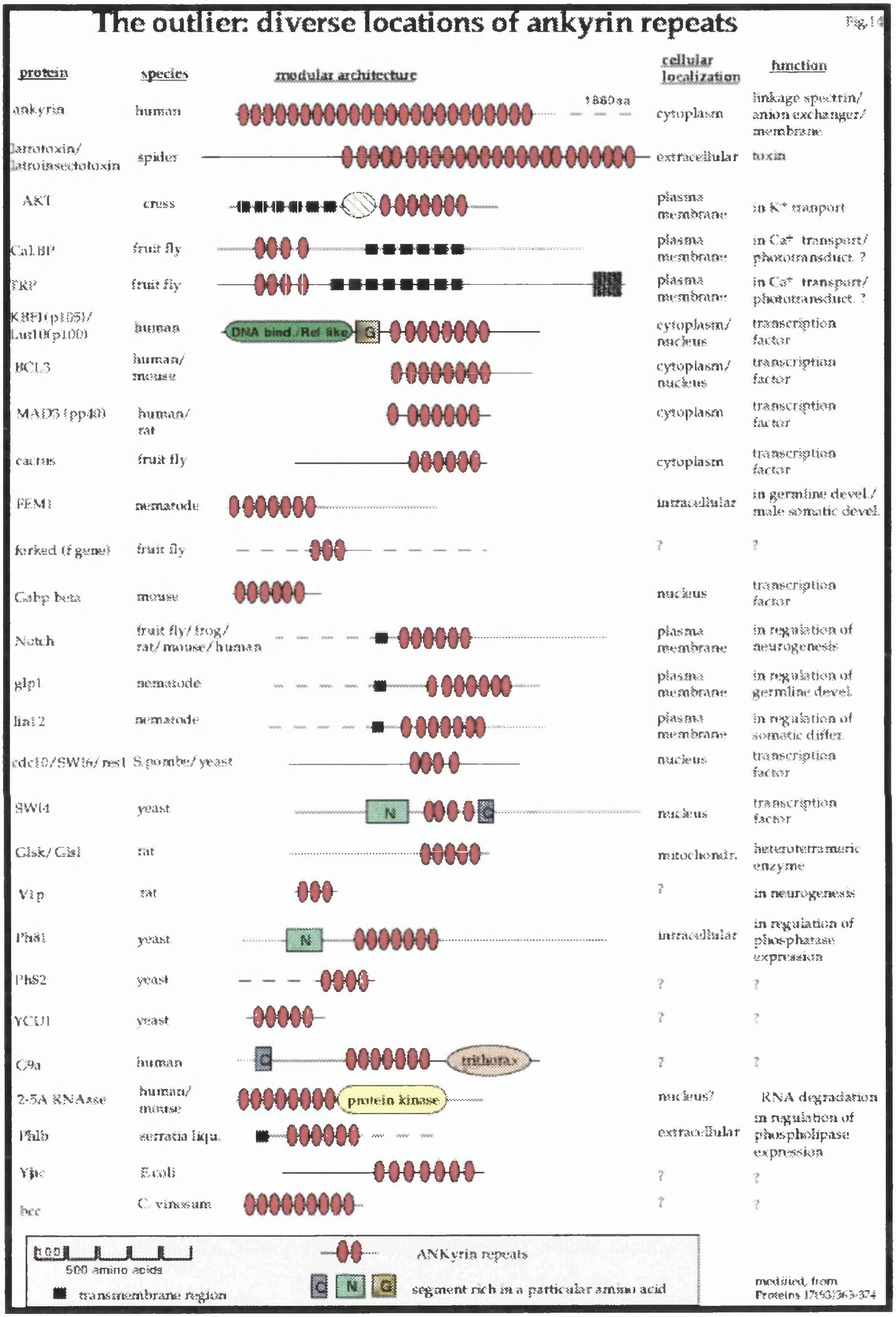
Another example of a set of widely distributed domains restricted to eukaryotes is involved in recognizing and binding phosphorylated proteins (Doolittle, 1995). The src homology region 2 (SH2) is known to bind phosphorylated tyrosines in proteins (Koch *et al.*, 1991; Waksman, 1992). SH2 is frequently found in association with another well defined domain, the src homology region 3 (SH3), whose function is not yet clear (Koch *et al.*, 1991). It is interesting to note that its three-dimensional structure (Netter *et al.*, 1993) has been shown to resemble a photosystem protein from a cyanobacterium (Falzone *et al.*, 1994).

A commonly shuffled eukaryotic domain is the PH domain (Pleckstrin Homology), which is thought to be involved in the binding and recognition of phosphorylated serines and threonines (Musacchio *et al.*,1993). There is also a 42-amino acid repeat that occurs in a wide variety of fungal and animal cells that apparently participates in multiple interactions within the nucleus. In mammals it was called plakoglobin (Franke *et al.,* 1989); in Drosophila the homologous protein (63% amino acid conservation) corresponds to a mutant called *armadillo* (Peifer and Wieschaus, 1990). The first reported use of this domain in yeast was shown to be involved in a suppressor of temperature-sensitive mutations (Yano *et al.,* 1994). PH has also been found in the animal adhesion junction protein beta-catein (Peifer *et al.*, 1992). The number of repeats within any protein containing this domain can be as many as eight (Doolittle, 1995).

Another small domain is the 33-residue repeat known as the ankyrin repeat. Database searching revealed that the domain is spread through a wide variety of organisms (Figure, 1.7; Bork, 1993). The distribution is unusual enough that the possibility of horizontal gene transfer has been raised. As in the case of other relatively short domains, it is thought that protein stability is achieved by the ankyrin repeats packing closely together. A result of this is that this domain cannot exist as a single unit (Doolittle, 1995).

(1.4.2) <u>Extracellular Multidomain Proteins in Animals.</u>

The large increase in the number of genetically movable domains that occurred with the evolution of multicellular organisms gives a rich demonstration of the construction of new proteins by shuffling domains (Doolittle, 1995). The various genome sequencing projects have and will provide an enormous amount of data from which relationships can be explored between different types of organisms. Any shuffling events occurred much more recently and is therefore more evident. Amino acid sequences alone are usually sufficient for following the dispersal of domains. It also



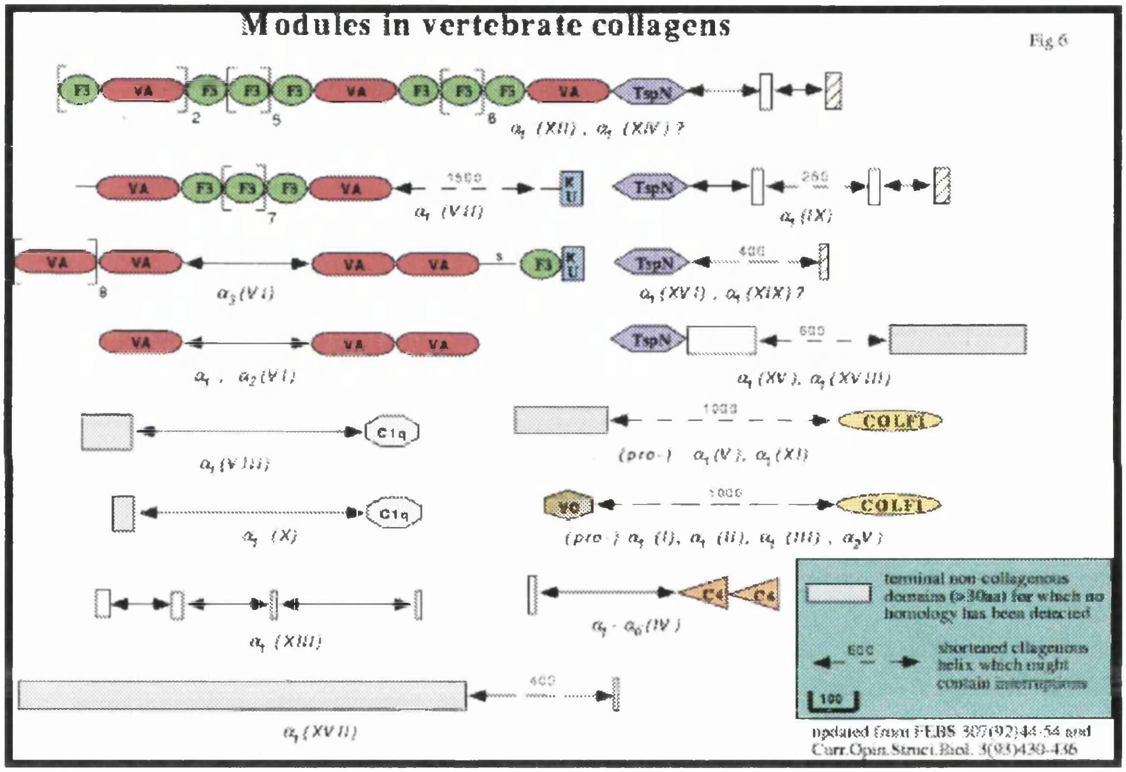
The outlier: diverse locations of ankyrin repeats

Fig.14

protein	species	modular architecture	cellular localization	function
ankyrin	human	1880aa	cytoplasm	linkage spectrin/ anion exchanger/ membrane
latrotoxin/ latroinsectotoxin	spider		extracellular	toxin
AKT	cress		plasma membrane	in K⁺ transport
CalBP	fruit fly		plasma membrane	in Ca⁺ transport/ phototransduct. ?
TRP	fruit fly		plasma membrane	in Ca⁺ transport/ phototransduct. ?
KBF1(p105)/ Lyt10(p100)	human	DNA bind./Rel like G	cytoplasm/ nucleus	transcription factor
BCL3	human/ mouse		cytoplasm/ nucleus	transcription factor
MAD3 (pp40)	human/ rat		cytoplasm	transcription factor
cactus	fruit fly		cytoplasm	transcription factor
FEM1	nematode		intracellular	in germline devel./ male somatic devel.
forked (f gene)	fruit fly		?	?
Gabp beta	mouse		nucleus	transcription factor
Notch	fruit fly/ frog/ rat/ mouse/ human		plasma membrane	in regulation of neurogenesis
glp1	nematode		plasma membrane	in regulation of germline devel.
lin12	nematode		plasma membrane	in regulation of somatic differ.
cdc10/SWI6/ res1	S.pombe/ yeast		nucleus	transcription factor
SWI4	yeast	N C	nucleus	transcription factor
Glsk/ Glsl	rat		mitochondr.	heterotetrameric enzyme
Vlp	rat		?	in neurogenesis
Ph81	yeast	N	intracellular	in regulation of phosphatase expression
PhS2	yeast		?	?
YCU1	yeast		?	?
G9a	human	C trithorax	?	?
2-5A RNAase	human/ mouse	protein kinase	nucleus?	RNA degradation
PhlB	serratia liqu.		extracellular	in regulation of phospholipase expression
Yjhc	E.coli		?	?
bcc	C. vinosum		?	?

modified from Proteins 17(4):363-374

Figure 1.7: Bork and Bairoch, (1995), domain cartoons for proteins containing the ankyrin repeat. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).


Modules in vertebrate collagens
Fig 6
F3
VA
TspN
2
5
6
α_1 (XII) , α_1 (XIV) ?
1500
K
U
α_1 (VII)
7
250
α_1 (IX)
8
α_3 (VI)
5
400
α_1 (XVI) , α_1 (XIX) ?
α_1 , α_2 (VI)
600
α_1 (XV), α_1 (XVIII)
C1q
α_1 (VIII)
1000
COLFI
(pro-) α_1 (V), α_1 (XI)
α_1 (X)
VC
(pro-) α_1 (I), α_1 (II), α_1 (III) , α_2V)
C4
α_1 (XIII)
α_1 - α_6 (IV)
α_1 (XVII)
terminal non-collagenous domains (>30aa) for which no homology has been detected
600
shortened cllagenous helix which might contain interruptions
100

updated from FEBS 307(92)44-54 and Curr.Opin.Struct.Biol. 3(93)430-436

Figure 1.8: Bork and Bairoch, (1995), domain cartoons showing the components of selected vertebrate collagens. The collagen types are denoted by Roman numerals. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

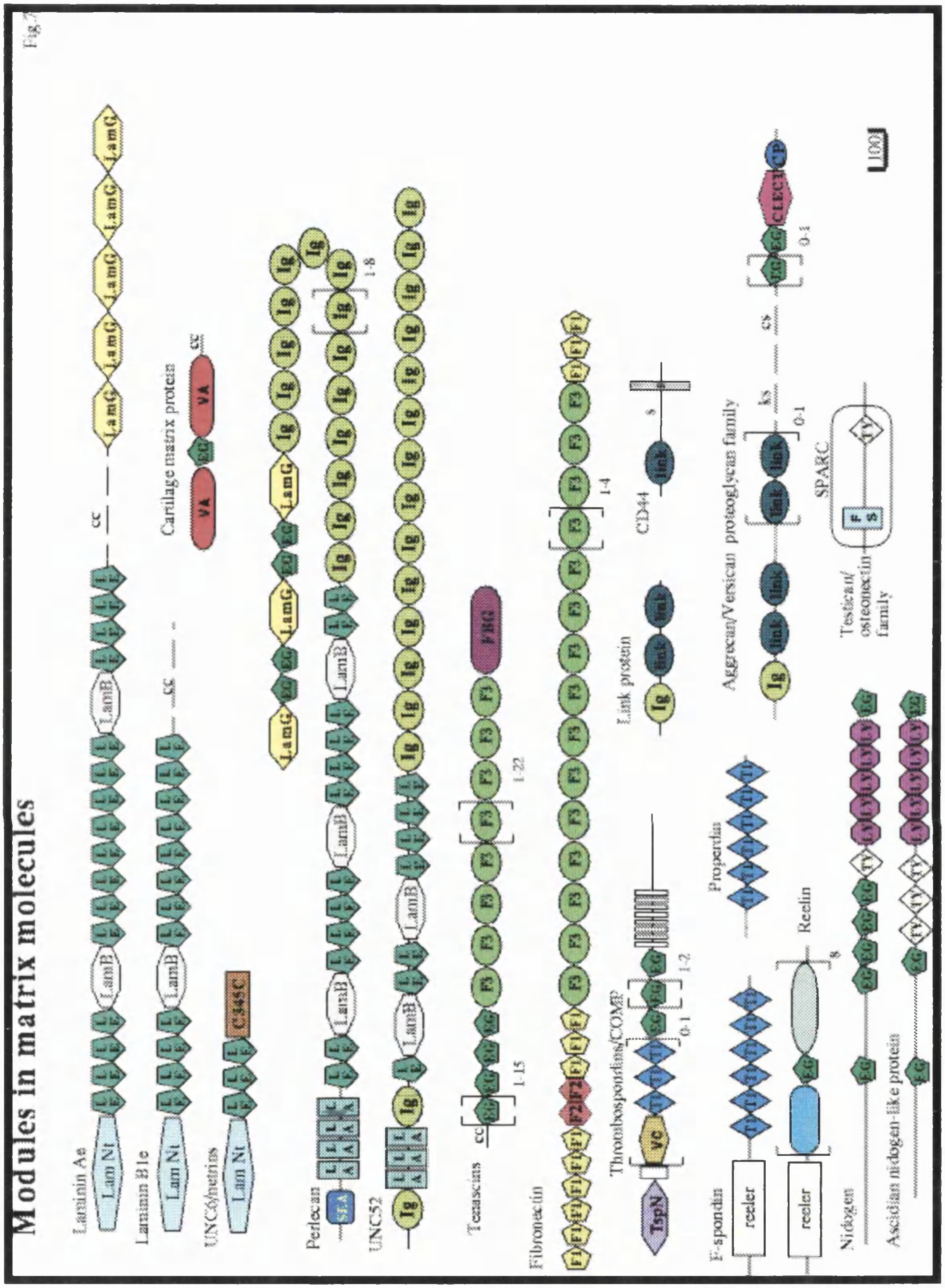


Figure 1.9: Bork and Bairoch, (1995), domain cartoons showing the components of selected extracellular matrix molecules. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

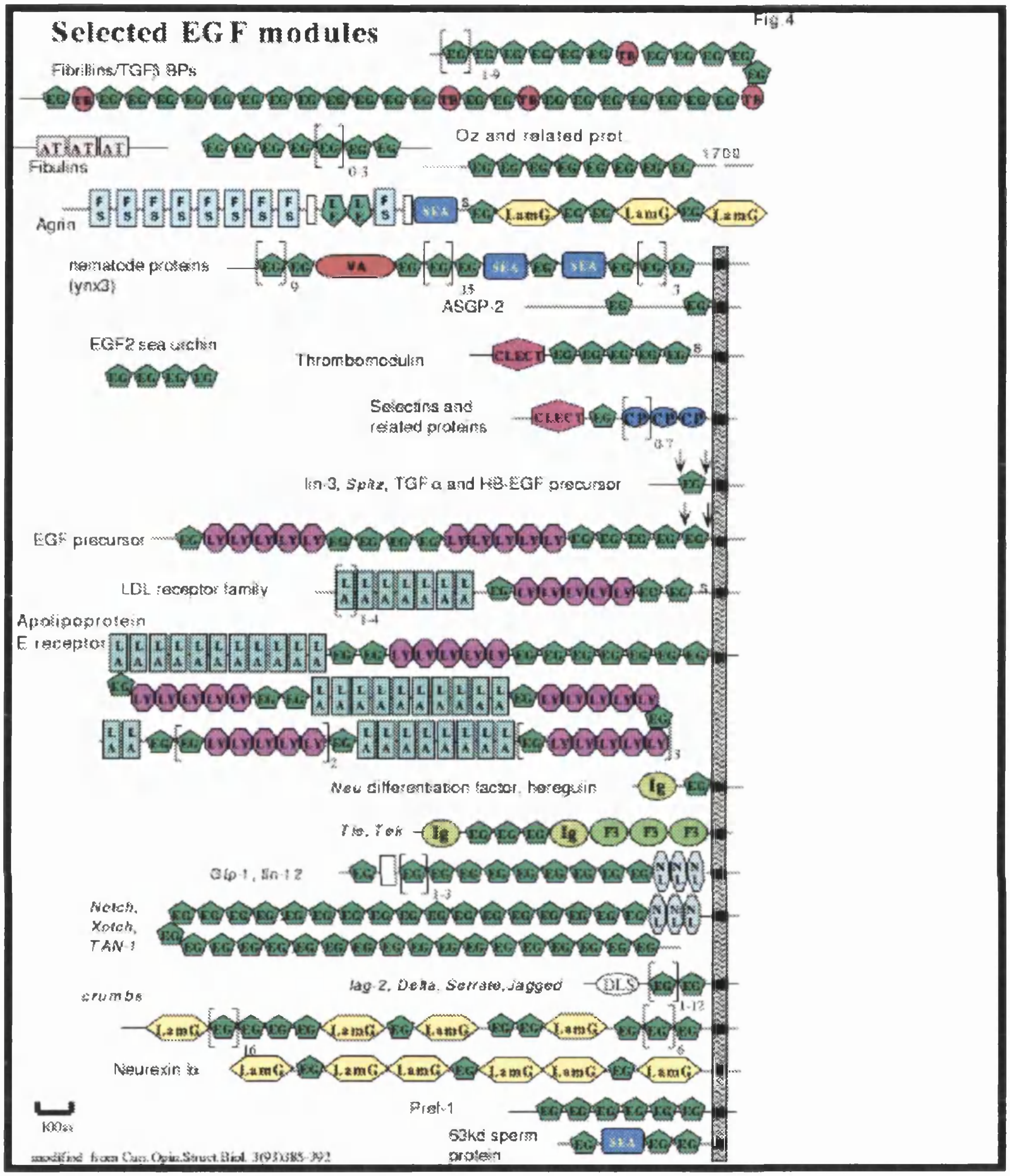


Figure 1.10: Bork and Bairoch, (1995), domain cartoons showing the components of selected proteins containing EGF domains. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

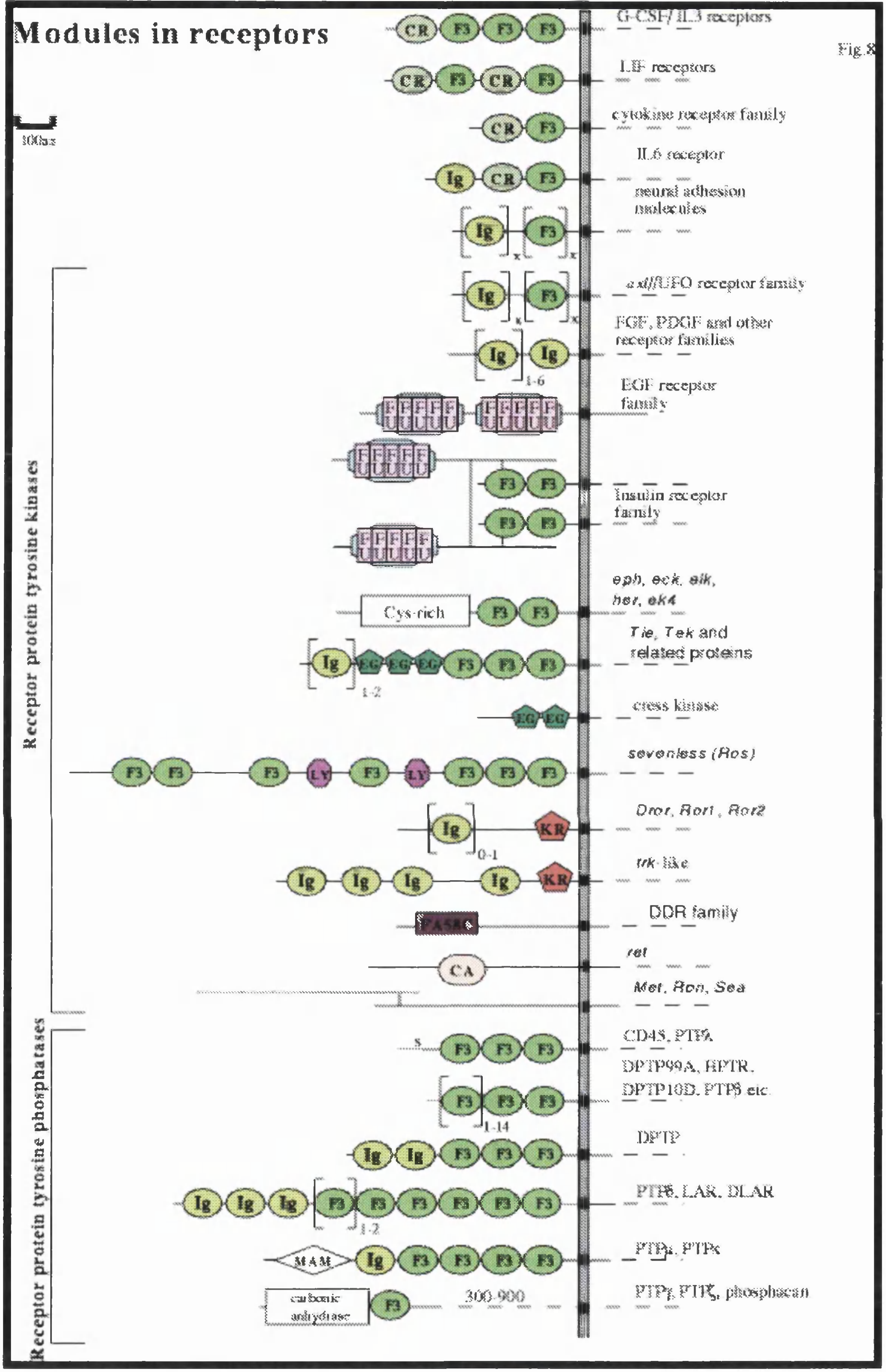


Figure 1.11: Bork and Bairoch, (1995), domain cartoons showing the components of selected receptors. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

appears that domain shuffling has been more prevalent in animals and has been part of multicellular organism development, giving rise to many unique proteins needed in cell-to-cell signalling and to the systems of defence and repair that are so essential to a multicellular existence (for examples, see Figures 1.2 and 1.4; Doolittle, 1994).

Many different animal cells exist in an extracellular matrix (ECM) that is composed of a huge number of macromolecules, with many of the components appearing to be present in all animals but virtually nowhere else (Doolittle, 1994). The most abundant of these proteins is collagen. This protein is structurally diverse, in which different collagen types form cables in some instances and sheets in others. Such diversity is attributable to its being associated with a host of different non-collagen domains (Figures 1.2, 1.4 and 1.8). In addition to collagen there are a mixture of other proteins such as syndecans, perlecans and laminins, (examples Figure 1.9) which make up part of the proteoglycan portion of which is known to be rich in sulphated mucopolysaccharides (Doolittle, 1994).

About 50 different domains have been identified that are shuffled between various animal extracellular proteins (Bork, 1992; Bork, 1993). Of these domains a few occur much more abundantly than all the others combined. These include the EGF-like domain (Figure 1.10) and the immunoglobulin (Ig) and fibronectin type III (Fn3) domains (Figure 1.11). It should be noted that Ig and Fn3 domains are not restricted to animals or to an extracellular existence.

Other frequently shuffled domains include the calcium-dependant carbohydrate-recognition domain (C-type lectins), SCR's , vWF type A domains, chaderins, collagen segments, some domains originally observed in the LDL receptor (Chapters 6 and 7), serine proteases (Chapters 6 and 7). These are shown in Figure 1.1 as well as other examples (Doolittle, 1995).

(1.4.3) Domains Unique to Plant or Fungal Proteins.

About a billion years ago, plants, animals and fungi shared a common ancestor, and as noted in earlier sections they share common domains that are involved in gene expression and intracellular regulation. However the three kingdoms have adapted different strategies for many of their extracellular involvements. Accordingly, it is perhaps not so surprising that most of the commonly shuffled domains that occur in extracellular proteins in animals have not yet been found in plants or fungi (Doolittle, 1995). An exception to this was the discovery of a somatomedin domain and four hemopexin-type sequence repeats in the cytosolic plant protein PA2, a major storage protein component in pea seeds, leading to the suggestion that these two domains existed in the common ancestor of plants and animals (Jenne, 1991).

Most of the common animal domains (Figure 1.1), including any resemblance to an immunoglobulin domain, have not been found in plants. This may be attributable to the relatively low number of reported plant and fungal sequences, or it could be because the sequences have changed to the extent where relationships can only be recognised at the three-dimensional structural level, as has been shown with some bacterial homologues. However, there is no indication that the somatomedin and haemopexin domains are changing any less rapidly than the others (Doolittle, 1995). It should also be noted that a large proportion of the common domains have immunological applications in animals, plants have different systems for dealing with infection. Lectins are found in both animals and plants (Singleton and Sainsbury, 1987).

Plants do have well known highly repetitive proteins that appear to be unique to plants. These include seed storage proteins in barley, rye and wheat (Kreis *et al.*, 1985). It seems reasonable that new domains have evolved during the time since plants and animals shared a common ancestry (Doolittle, 1995). As for fungi, the size and breadth of the ongoing yeast sequencing projects shows that the notion of not enough data being collected to show where the missing domains are is not correct. Surprisingly a third to a half of all the open reading frames appear to be unidentifiable and unique to fungi (Koonin *et al.,* 1994). These open reading frames may code for structurally homologous proteins, but because the sequences may have evolved so much, the changes are too big for standard sequence searching (Doolittle, 1995). A curious feature of these unidentified open reading frames is that a large fraction of them resemble other unidentified open reading frames (Koonin *et al.,* 1994). These findings tend to confirm an earlier observation of an extensive family of proteins typified by a 34-residue repeat that also appeared limited to fungi (Sikorski *et al.,* 1990).

(1.4.4) Mobile Domains in Bacteria.

It has been shown that numerous ubiquitous proteins were constructed as domains in ancient times from the observation of obviously rearranged domains by X-ray crystallography. Multidomain proteins have continued to evolve throughout the history of the prokaryotes as shown by many examples of module shuffling observed on the basis of primary structure similarities (Doolittle, 1995).

A class of small molecule binding proteins found in the periplasmic space of Gram stain negative bacteria are known as periplasmic binding proteins. A total of 8 different prokaryotic subclasses that bind to sugars, amino acids and anions have been identified, and crystal structures have been determined for six of these (Tam and Saier, 1993). Periplasmic binding proteins consist of two nonequivalent α-helix/β-sheet domains joined by polypeptide links which flank a ligand-binding site in a large cleft between them.

AmiC (Chapter 5) is a soluble cytoplasmic protein that functions as an amide sensor and negative regulator of the amidase operon. AmiC controls the activity of the transcription antitermination factor AmiR, which in turn regulates expression of the amidase enzyme system. *AmiE* is the gene which corresponds to the amidase enzyme, and *amiB* and *amiS* appear to form a membrane transport system for the importation of amide into the bacteria (Drew and Wilson, 1992; Wilson *et al.*, 1995). The combination of secondary structure predictions and fold recognition analyses indicated that, despite only 17% amino acid sequence identity, AmiC had the same protein fold as the leucine-isoleucine-valine binding protein (LivJ) of *Escherichia coli* (Sack *et al.*, 1989a; Wilson *et al.*, 1993). LivJ corresponds to the Cluster 4 subclass of periplasmic binding proteins (Table 5.1; Tam and Saier, 1993). The prediction was confirmed by the crystal structure of AmiC bound to its substrate acetamide (Pearl *et al.*, 1994).

It has also been determined that the monomers of the *lac* repressor core tetramer (LacR) of *E. coli* when complexed with IPTG have a tertiary structure similar to the structures of the Cluster 2 subclass of periplasmic binding proteins (Table 5.1; Tam and Saier, 1993; Friedman *et al.,* 1995). This structure is highly homologous to the arabinose binding protein, and changes in its conformation within this domain affect the ability of the DNA binding domain to interact with *lac* operator. AmiC, however, interacts with AmiR which is an RNA binding protein. Such systems where a ligand receptor directly regulates the activity of an RNA binding protein remains unique at present (Wilson *et al.*, 1996).

AmiC exhibits distinct functional properties in that it controls AmiR in response to a signal from acetamide, while the periplasmic binding proteins transport small molecules within the inner bacterial membrane. A similar relationship with LivJ has also been identified for the extracellular domain of the eukaryotic protein glutamate receptor, which is involved in neurotransmitter activity. It is thought the domain closing on binding a ligand may cause an analogous conformational change that initiates signal transduction through the covalently linked transmembrane domains. Periplasmic binding proteins interact with but are not covalently linked to membrane components (O'Hara *et al.*, 1993; Stern-Bach *et al.*, 1994). Because sequence similarities are low (19-20 %) among bacterial periplasmic binding proteins and between periplasmic binding proteins in Gram stain negative bacteria and ionotropic glutamate receptors in mammals, it is possible that both groups share a common architecture (hence ancestor) despite large evolutionary distances (Kuryatyov *et al.*, 1994).

There are numerous instances of repetitive or transposed parts of proteins that occur in different settings (Doolittle, 1995). The *E. coli* DNA repair protein Uvr was shown to consist of a series of repeated segments that were shown to be homologous to the active transport system of a class of periplasmic binding proteins. An example of one with this homology (which transports histidine) is found in *Salmonella typhimurium* (Gilson *et al.*, 1982; Doolittle *et al.*, 1986).

The timing of some domain exchange events can be gauged by examining how similar repeated segments are in different genomes. This problem is complicated because many recently shuffled domains seem to have also been involved in horizontal gene transfer. In many soil bacteria, carbohydrate-binding domains are found at widely differing locations in a diversity of extracellular glycohydrolases (Figure 1.12; Gilkes *et al.*, 1991; Fujii *et al.*, 1993; Meinke *et al.*, 1991). Some of these domains have also been found in some eukaryotes (Ramalingam *et al.*, 1992). Fn3 domains are also found in many of the same enzymes (Figure 1.12; Bork and Doolittle, 1992; Hansen, 1992). These bacterial Fn3 domains are too similar in sequence to animal Fn3 domains to have a common ancestor (Bork and Doolittle, 1992; Little *et al.*, 1994). These proteins show all the characteristics of "exon shuffling without introns" (Doolittle, 1995).

There are a number of prokaryotic structural proteins which are built from domains (Doolittle, 1995). These include wall-attachment proteins, some of which contain coiled-coil motifs (Engel *et al.*, 1992; Lupus *et al.*, 1994). Others are highly repetitive and exhibit sequence motifs observed in other microbial proteins (Foster, 1993).

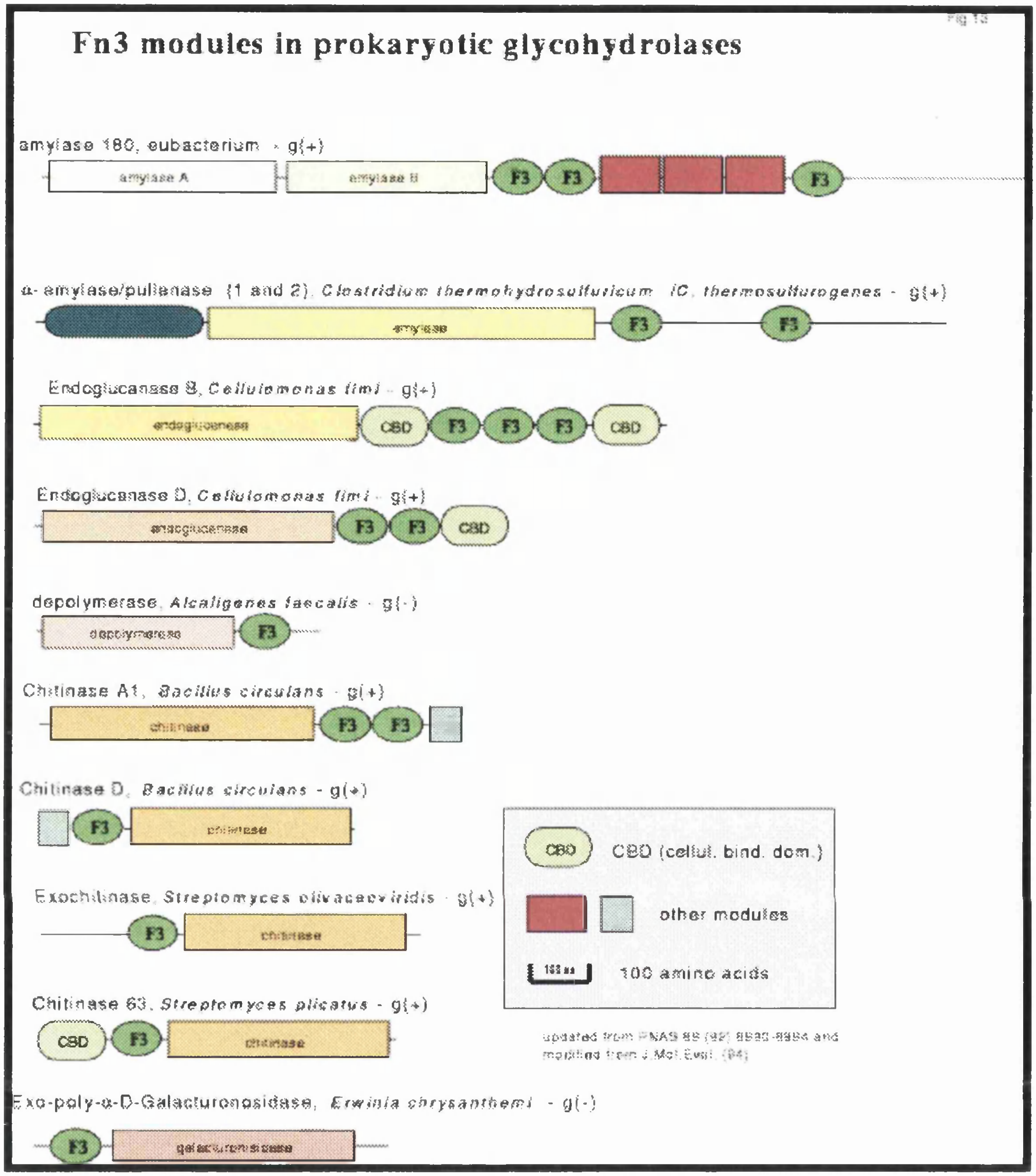


Fig 13
Fn3 modules in prokaryotic glycohydrolases
amylase 180, eubacterium - g(+)
amylase A
amylase B
F3
F3
F3
α-amylase/pullanase (1 and 2), Clostridium thermohydrosulfuricum (C. thermosulfurogenes - g(+)
amylase
F3
F3
Endoglucanase B, Cellulomonas fimi - g(+)
endoglucanase
CBD
F3
F3
F3
CBD
Endoglucanase D, Cellulomonas fimi - g(+)
endoglucanase
F3
F3
CBD
depolymerase, Alcaligenes faecalis - g(-)
depolymerase
F3
Chitinase A1, Bacillus circulans - g(+)
chitinase
F3
F3
Chitinase D, Bacillus circulans - g(+)
F3
chitinase
Exochitinase, Streptomyces olivaceoviridis - g(+)
F3
chitinase
Chitinase 63, Streptomyces plicatus - g(+)
CBD
F3
chitinase
CBD
CBD (cellul. bind. dom.)
other modules
100 aa
100 amino acids
updated from PNAS 89 (92) 8990-8994 and
modified from J.Mol.Evol. (94)
Exo-poly-α-D-Galacturonosidase, Erwinia chrysanthemi - g(-)
F3
galacturonosidase

Figure 1.12: Bork and Bairoch, (1995), domain cartoons showing carbohydrate binding domains and Fn3 domains in prokaryotic glycohydrases. (Taken from the EMBL Heidelberg World Wide Web Site http://www.embl-heidelberg.de).

A good example of domain shuffling in bacteria is shown in a number of membrane proteins that bind animal blood plasma proteins. A number of these are known to bind specific animal proteins with great specificity, including fibrinogen, fibronectin, immunoglobulins, and plasma albumin (Doolittle, 1995). Staphylococcal protein A binds the Fc portion of IgG and is of great biotechnological significance for purifying antibodies. There are also streptococcal proteins which bind to IgG but also bind to plasma albumin. For example, Protein G binds IgG with its amino-terminal domain and albumin with its carboxyl-terminal domain (Guss *et al.*, 1986). Certain strains of *Peptostreptococcus* are able to bind immunoglobulin light chains with protein L, (Kastern *et al.*, 1992). Other strains of this bacterium have a protein (PAB) that is clearly the result of a very recent joining together of domains from proteins G and L that binds albumin (de Chateau and Bjorck, 1994; Doolittle, 1995). This protein is thought to have arisen from homologous or nonhomologous recombination or perhaps a conjugative plasmid (de Chateau and Bjorck, 1994).

There are many other examples of rearranged domains in bacterial proteins, which show that domain shuffling has been a persistent activity among bacteria (Doolittle, 1995). Examples include the components of various nitrogen fixation pathways, (Ouzounis *et al.*, 1994), or the prokaryotic phosphotransferase system (Saier and Reizer, 1990).

(1.5) Summary and Conclusions.

There are a great number of proteins which can be shown by structural studies or sequence analysis to be made up from defined subunits or domains. If the function of the domains of interest are known, it may then be possible to infer the function of an unknown multidomain protein. With the various genome projects generating huge amounts of sequence data, database searches for possible domains and molecular modelling techniques will be very useful in determining putative structure and functions of the open reading frames. However, as there are only a few hundred known structures for different domain types, it may be some time before predictive modelling of open reading frames is possible.

CHAPTER 2

BIOPHYSICAL TECHNIQUES

(2.1) Importance of Structure Determination.

All biological macromolecules are composed of a specific arrangement of amino acids, carbohydrates, nucleic acids and lipids. These are organized to produce a unique three-dimensional structure that performs a particular functional role in the living organism. The spatial arrangements of the residues are very important to the function of the macromolecule and a change in conformation may result in loss of activity. In order to understand the function of a macromolecule, a detailed knowledge of structure is necessary.

(2.2) Methods of Structure Determination and Analysis.

Biological structure determination can be divided into two categories, namely those at high and low resolutions. High resolution implies a structural determination down to the atomic level, as exemplified by X-ray crystallography, and more recently by neutron crystallography and nuclear magnetic resonance (NMR). Over 7000 sets of protein atomic coordinates are currently listed in the Brookhaven Protein Data Bank (February 1998). Resolutions of 0.14 nm have been achieved using X-ray crystallography on large proteins [e.g. rubredoxin from *Desulfovibrio gigas* (Frey *et al.*,1987); cytochrome b_{562} from *Escherichia coli* (Hamada *et al.*, 1995)] and resolutions of 0.1 nm have been obtained with nucleic acids. Neutron crystallography is complementary to X-ray crystallography and is used to refine X-ray diffraction maps (reviews: Jacrot, 1987; Moore, 1985). Neutrons have the ability to discriminate between protons and deuterons, as the large difference in scattering amplitudes of hydrogen and deuterium permits the location of individual protons or water molecules in the protein structure (Bradshaw 1995).

Low resolution techniques include electron microscopy, hydrodynamic measurements and small-angle solution scattering. Electron microscopy can visualize the whole macromolecule *in vacuo* down to a resolution of around 2 nm, although the images are studied in a vacuum after staining which may cause artefacts. Hydrodynamic measurements give information on M_r and the shape of a macromolecule in solution. Small-angle solution scattering also is used to study gross shape parameters of macromolecules in solution to a resolution of 2 to 4 nm, and is discussed in some detail below. These techniques can only give a limited amount of structural information such as overall length of the molecule or its molecular weight. These techniques are, however, very useful for situations where high resolution techniques are not applicable. Comparisons of NMR, X-ray crystallography, small angle X-ray scattering and small angle neutron scattering are detailed in Table 2.1 and 2.2.

Spectroscopic techniques such as Fourier transform infrared spectroscopy (FTIR) and circular dichroism spectroscopy (CD) can give information on the types and proportions of secondary structure present in a protein but will not give a low resolution three-dimensional structure. CD spectroscopy is better at determining α-helical content than FTIR and FTIR is better at determining β-sheet content than CD, so the two techniques are largely complementary. FTIR spectroscopy utilises the ability of molecules to absorb certain wavelengths of infrared radiation that correspond to selected types of vibrations. Experimental studies with proteins and polypeptides of known atomic structure have shown very good correlation between the frequency of the amide I band (approximately 1690-1600 cm^{-1}) and the different types of secondary structure that are present. Spectral deconvolution of the broad amide I band reveals subcomponents that can be assigned to α-helix, β-sheet and random structures (Haris and Chapman, 1994), and quantification will lead to the relative percentages of these structures. CD spectroscopy corresponds to the difference in the absorption of left and right circularly polarized light. It can be considered as the absorption spectrum measured with left circularly polarized light minus the absorption spectrum measured with right circularly polarized light. To exhibit CD properties a sample must be optically active, which in turn requires that the molecule is not super-imposable on its mirror image (Drake, 1994). Because proteins consist of L-amino acids and are therefore optically active, this makes them suitable for this technique. The appearance of the CD spectrum is the direct consequence of the absolute spatial aspect of molecular shape. Two types of CD spectroscopy can be distinguished, that related to related to biopolymer backbones which is derived from amide-amide interactions, and that due to optical activity of chromophores (e.g ring structures on amino acids; ligands; haem groups) (Drake, 1994). The reliable calculation of CD spectra associated with specific secondary structures from first principles remains difficult, since different protein conformations will have different amide-amide orientations and therefore exhibit different CD spectra. In practice, reference to model polypeptides or correlation of the CD spectra of proteins with their known X-ray structures has led to a consensus set of spectra which can be treated as a set of fingerprints for the different secondary structures. The measured CD spectrum corresponds to a linear combination of these fundamental spectra and, from this the proportions of secondary structure can be determined (Drake, 1994).

Table 2.1: Stages in the determination of a protein structure by X-ray crystallography and NMR. (Based on MacArthur *et al.*, 1994).

X-Ray	**NMR**
Crystallization and derivative preparation ↓ Data collection and processing ↓ Location of heavy atoms (or molecular replacement) ↓ Calculation of phases and electron-density mapping ↓ Chain tracing and interpretation ↓ Model building ↓ Refinement of model	Sample preparation with possible isotope labelling ↓ Data collection (NOESY and COSY spectra) ↓ Sequential assignments ↓ Analysis and quantification of NOE peak intensities and conversion to approximate proton-proton separations ↓ Generation of models consistent with the NOE-derived separations ↓ Generation of models consistent with the NOE-derived separations and torsion-angle ranges from coupling constants, usually by distance geometry and simulated annealing algorithms ↓ Model improvement by inclusion of initially unassigned NOE distances and stereospecific assignments
PROS ●Well-established technique. ●More mathematically direct image construction. ●More objective interpretation of data. ●Raw-data processing highly automated. ●Quality indicators available (resolution, R-factor). ●Mutants, different ligands and homologous structures (as low as 25% sequence identity) readily compared by difference Fourier methods. ●Large molecules and assemblies can be determined, e.g. virus particles. ●Surface water molecules relatively well defined. ●Produces a single model that is easy to visualize and interpret.	**PROS** ●Closer to biological conditions. ●Can provide information on dynamics and identify individual side-chain motion. ●Secondary structure can be derived from limited experimental data. ●Free from artefacts resulting from crystallization. ●Can be used to monitor conformational change on ligand binding. ●Good for checking the correct fold of mutants. ●Ideal for small domains. ●Solution conditions can be explicitly chosen and readily changed, e.g. pH, temperature, etc. ●Useful for protein-folding studies.
CONS ●Protein has to form stable crystals that diffract well. ●Need heavy-atom derivatives that form isomorphous crystals. ●Crystal production can be difficult and time consuming and often impossible. ●Unnatural, nonphysiological environment. ●Difficulty in apportioning uncertainty between static and dynamic disorder. ●Surface residues may be influenced by crystal packing. ●May not wholly represent structure as it exists in solution. ●Less useful for large flexible modular proteins, e.g complement factors B, I and H. ●Model represents a time-averaged structure where details of mobility are unresolved.	**CONS** ●Requires concentrated solution - therefore danger of aggregation or oligomerisation. ●Currently limited to determination of relatively small proteins (<20KDa) ●Lack of established quality indicators of data and model, such as resolution and R-factor ●A weaker and more subjective interpretation of the experimental data than in X-ray crystallography ●Surface residues generally less well defined than in X-ray crystallography ●The distinction between flexibility and lack of data is not always easy ●Produces an ensemble of possible structures rather than one model. Time averaged structure. ●Conformational variability can make interpretation difficult ●Complete structure determination required if homology is less than ~60% sequence identity

(2.2.1) Crystallography.

X-rays are a form of electromagnetic radiation. The wavelength commonly used by crystallographers is 0.154 nm. This wavelength is used because it is comparable to atomic dimensions (e.g. 0.15 nm for C-C single bonds). The major problem with all crystallographic techniques is the initial requirement of a crystal, since scattering from an individual molecule is far too weak to be detected (MacArthur *et al.*, 1994). Good crystals are grown from a protein solution under a narrow range of conditions which are empirically determined by trial and error. Once a suitable crystal has been obtained, diffraction experiments will determine the intensity of the scattered waves, but further techniques are required to elucidate their phase. Crystals constrain the scattering exclusively in discrete directions to produce the diffraction pattern in accordance with the Bragg equation:

$$n\lambda=2\,d\sin\theta$$

where n is an integer, d is the spacing between lattice planes and θ is the angle of incidence at the wavelength λ. The intensity of the scattered radiation falls off with increasing θ value. There is a minimum value of d (d_{min}) that corresponds to the maximum observed θ. The resolution of a crystal structure is determined by the d_{min} value because it defines the ability to distinguish between adjacent structural features. The basic structural unit of a crystal, or unit cell, can be "seen" in all parts of the crystal as it is repeated infinitely in three dimensions. The unit cell is characterised by the vectors a, b and c and the angles α, β, γ that form the edges of a parallelepiped (Chang, 1981; MacArthur *et al.*, 1994). It was shown in 1850 by Bravais that there are only 14 unit cell types known as the *Bravais lattices* (Chang, 1981). The amplitude of a wave scattered by the contents of the unit cell in a given reflection in the diffraction pattern is described by the structure factor F, whose magnitude is obtained directly from the experimentally observed intensity. Reconstruction of the electron density of the molecule requires knowledge of the phase as well as the amplitude.

An electron density map is a contoured representation of the electron density at various points in a crystal structure. Electron density is highest at atoms. The map may be calculated by Fourier summation from the experimental structure amplitudes, F_{obs} and an appropriate set of phases α_{hkl}:

$$\rho(xyz)=\frac{1}{V}\sum_{\mathrm{hkl}} F_{obs}(hkl)\ \cos(2\pi(hx+ky+lz)-\alpha_{hkl})$$

where ρ(xyz) is the electron density at the point xyz, which are the fractional coordinates measured from the unit cell origin; hkl are integers characteristic of a given reflection, and V is the unit-cell volume. To solve the phase problem, diffraction from at least two heavy-atom (for example, cobalt or uranium salts) derivatives must be measured. Another method of solving the phase problem is by using molecular replacement where the phases are calculated by fitting the known structure of a structurally homologous protein to the observed intensities. Once an electron-density map has been constructed, the molecular structure is then derived using molecular graphics software (MacArthur *et al.*, 1994).

The initial electron density map is very approximate, so the model from this is refined until the best agreement is found between the observed structure amplitude (F_{obs}) and those back-calculated from the model (F_{calc}) and an R-factor of agreement is calculated. The initial model produced from an electron density map may have poor stereochemistry, implausible non-bonded contacts, and bond lengths and angles that show excessive deviations from ideal values, i.e. the model represents an unstable structure possessing high potential energy. The procedure used in model refinement is to lower the energy by adjusting the above parameters until they reach acceptable levels near their preferred values but still maintaining the experimentally determined conditions (MacArthur *et al.*, 1994). The R-factor is an index that gives a measure of the disagreement between the calculated and the observed structures, and is a measure of the correctness of the derived model summed over all reflections:

$$R=\frac{\sum |F_{obs} - F_{calc}|}{\sum |F_{obs}|}$$

In application to multidomain proteins, bacterial proteins are not glycosylated and are therefore more amenable to crystallisation and may be used to produce recombinant unglycosylated proteins. Two examples include the amide binding protein AmiC (Chapter 5) and the vWF type A domain of complement factor B. Of the multidomain complement components, the only ones that have been crystallized so far are fragments or single domains. These include C3a (Huber *et al.*, 1980), C-type lectin domains of mannose binding protein (Weis *et al.*, 1991a, b) factor D (Narayana *et al.*, 1991a, b) and the vWF type A domain of complement receptor type 3 (CR3) (Lee *et al.*, 1995). In general, crystallisation of complement proteins is problematic because of the high degree of glycosylation of complement proteins. For example C$\overline{1}$ inhibitor has a carbohydrate content of 26% by weight (Perkins *et al.*, 1990d). In addition, many complement proteins are composed of domains that show interdomain mobility (Reid and Day, 1989). The concentration in human plasma of the complement proteins is low, making it difficult to obtain sufficient yields necessary for crystallography, although recombinant DNA techniques are available to produce such proteins at high concentration (see Chapter 4). Recombinant methods also allow single domains to be studied.

(2.2.2) Nuclear Magnetic Resonance (NMR) Spectroscopy.

Nuclear magnetic resonance (NMR) spectroscopy can be carried out on atomic nuclei that possess a magnetic moment. These are typically nuclei with odd-numbered masses (e.g ^{1}H, ^{11}B, ^{13}C, ^{15}N, etc.). Such a nucleus may be regarded as a spinning, positively charged unit, and like any rotating electric charge it generates a tiny magnetic field along its spinning axis. If an external magnetic field is placed around the nucleus, the nucleus will rotate around an axis that is either parallel or antiparallel to the direction of the applied field and thus corresponding to upper and lower energy levels. The energy difference between these two spin states is characteristic of the particular type of nucleus and the strength of the applied magnetic field. The application of radio-frequency electromagnetic radiation to an ensemble of spins at the appropriate frequency will induce resonance between the two spin states, leading to the generation of detectable transverse magnetization (MacArthur *et al.*, 1994; Evans, 1995). The real value in NMR lies in the fact that the magnetic field experienced by the nucleus is not the applied field because the applied field is modified by the fields of neighbouring atoms. This effect is known as the chemical shift gives rise to the different positions of NMR signals in the spectrum that are characteristic of the environment of a specific proton (MacArthur *et al.*, 1994; Evans, 1995).

In conventional pulse-Fourier experiments, the sample is subjected to a short pulse whose frequency is centred in and covers the spectral region of interest. The generated output signal (or free induction decay) contains oscillating signals from all resonances as a function of time t. This is then converted by a Fourier transformation to give a one-dimensional spectrum of signal intensity as a function of resonance frequency (Chang, 1981; Evans, 1995; Willard *et al.*, 1981). By repeating the experiment many times, in which an additional incrementable time interval t_1 is set between the first pulse and a second short pulse, the data can be processed to obtain the NMR spectrum as a function of two frequencies (i.e a two-dimensional NMR spectrum) (Willard *et al.*, 1981; Evans, 1995).

Two-dimensional NMR can be used to study low molecular weight (M_r) proteins (up to 12 kDa) to atomic resolutions. NMR is the only high resolution technique to determine the three dimensional structure in solution in an environment near to physiological. Problems in the separate resolution of NMR signals arise with larger proteins for reason of signal overlap. However, new techniques including 3-D NMR and the isotopic labelling of protein samples have been used to determine structural information on proteins of M_r up to 27.5 kDa (Fesik *et al.*, 1989). Although the structures of proteins studied by both crystallography and NMR tend to be similar, significant differences have been identified between the solution and crystal states, attributable in part to the effects of crystal packing and the use of unphysiological buffers to promote crystallization (Chazin *et al.*, 1988; Nettesheim *et al.*, 1988). By 1989, the solution structure of more than 20 different globular proteins of 33 to 108 residues had been determined (Wright, 1989). By February 1998, over 1100 NMR structures had been deposited in the

Brookhaven Protein Database. These methods are ideal for large multidomain proteins which can be studied using NMR by isolating each independently folded domain. This has been achieved for the 5th SCR (not deposited in the Brookhaven Protein Database) and the 15th and 16th SCRs of factor H as single domains and a pair of domains (Barlow *et al.*, 1991; Barlow *et al.*,1992; Norman *et al.*, 1991). This is therefore of use in studying other complement proteins since most consist of small domains of less than 100 residues.

(2.2.3) Analytical Ultracentrifugation.

(2.2.3.1) Sedimentation Velocity Experiments.

One of the two major types of analytical ultracentrifugation experiment is the sedimentation velocity experiment. Particles suspended in solution are pulled downward by the effect of gravity. This movement is partially offset by the buoyancy of the particle. Since the earth's gravitational field is weak, a solution containing macromolecules is usually homogenous as a result of the random thermal motion of the molecules. The rate of sedimentation of the particles increases with the mass of the particle and the strength of the gravitational field, where the latter is changeable by spinning the solution in a centrifuge tube. The centrifugal force acting on the solute particle of mass m is $m\omega^2r$, where ω is the angular velocity of the rotor in radians per second (ω=θ/t where θ is the angle, t is the time, and ω is in rad s^{-1}), r is the distance from the centre of rotation to the particle, and ω^2r is the centrifugal acceleration of the rotor (Figure 2.1). In addition to the centrifugal force, the particle is subjected to buoyancy as a result of the displacement of the solvent molecules by the particle. This buoyancy reduces the force on the particle by ω^2r times the mass of the displaced solvent. The net force acting on any mass, m, is given by:-

net force = centrifugal force - buoyant force

$$= \omega^2 rm - \omega^2 rv\rho$$

The partial specific volume, $\bar{v}$, is defined as the increase in volume when 1 g of the dry solute is dissolved in a large volume of the solvent. It can be shown that for a particle of volume v and density ρ.

$$M = \frac{sN_0kT}{D(1-\bar{v}\rho)} = \frac{sRT}{D(1-\bar{v}\rho)}$$

M is the molar mass of the solute, N_0 is Avogadro's number, s is the sedimentation coefficient (unit = Svedberg) and the frictional coefficient f = kT/D where k is Boltzmann's constant, T is the absolute temperature, kT is a measure of the energy of the molecule, and D is the diffusion coefficient of the diffusing substance. The sedimentation coefficient is defined as s where:-

$$s = \frac{dr/dt}{\omega^2 r} = \frac{m(1-\bar{v}\rho)}{f} = \frac{M}{N_0}\ \frac{1-\bar{v}\rho}{f}$$

where dr/dt refers to the sedimentation velocity. The sedimentation coefficient for a given molecule is independent of the angular velocity of the rotor. As ω^2r increases, so does dr/dt and the ratio remains constant (Chang, 1981).

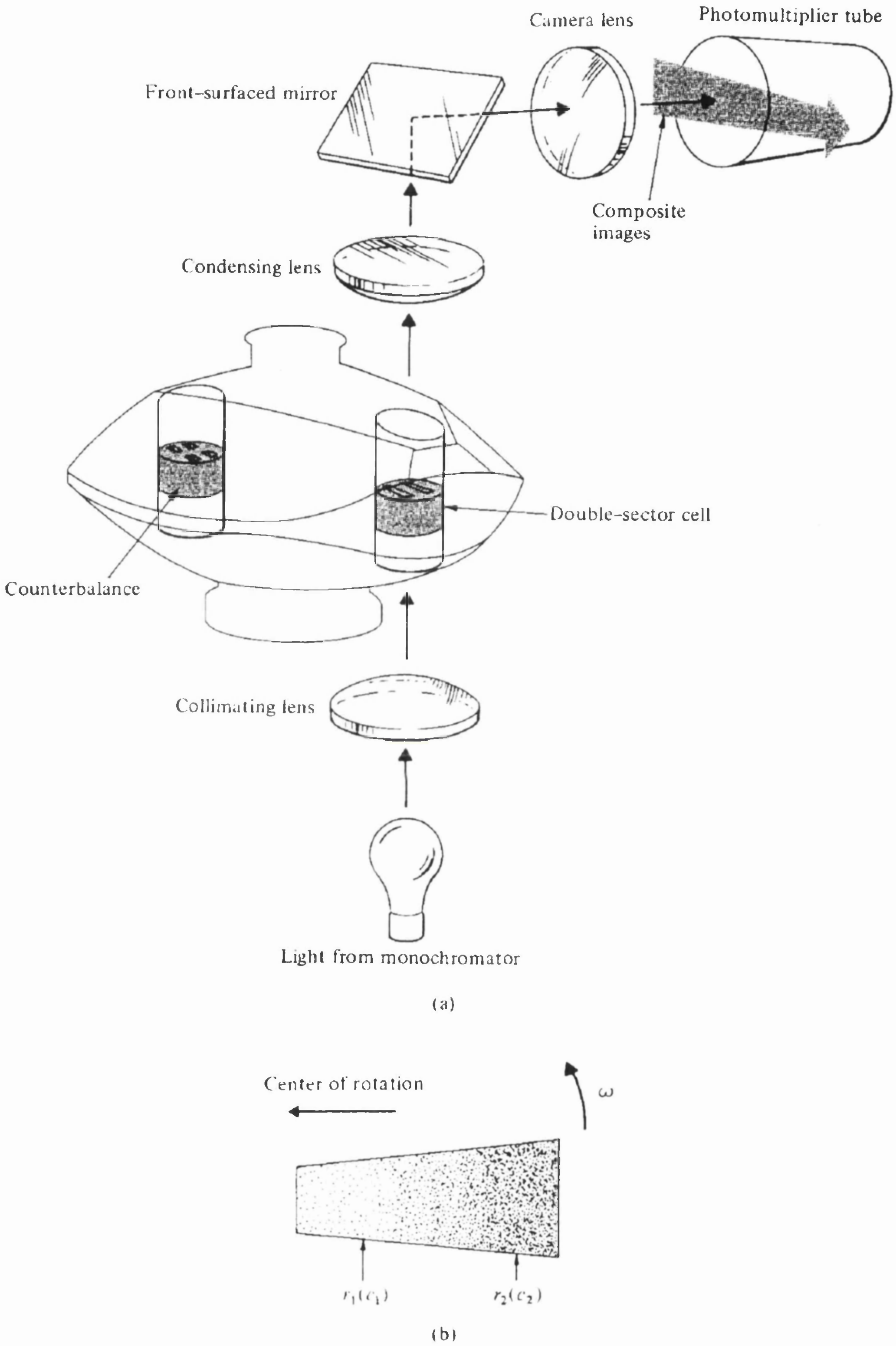


Figure 2.1: (a) Schematic diagram of an ultracentrifuge (Chang, 1981). (b) Concentration gradient established in the sample cell (Chang, 1981).

D and $\bar{v}$ can be determined by separate experiments or by calculation. The only quantity that needs to be measured in determining the molar mass M is the sedimentation coefficient re-expressed as:-

$$s\ dt = \frac{1}{\omega^2}\ \frac{dr}{r}$$

Integration over the distance travelled by the particle from r_0 (t=0) to r (t=t) gives:-

$$s = \frac{1}{t\omega^2}\ \ln\frac{r}{r_0}$$

Because ω is known, s and M are calculable (Chang, 1981). Suitable optical means such as refractive-index measurements using classical schlieren optics or scanning absorption optics are used to measure the movement of the sedimenting boundary in a given time and thereby obtains the sedimentation coefficient (Harding, 1994a).

Sedimentation coefficients are of little interest these days if they are used to determine only frictional ratios and the corresponding equivalent ellipsoid of revolution. Instead hydrodynamic theories have been developed to allow modelling of structures using the sedimentation coefficient and related parameters (Harding, 1994a). Sedimentation coefficients have been used to distinguish between possible solution models for several complement proteins using small spheres (Perkins, 1989), and intact active antibodies from which no high-resolution structural information from X-ray crystallography or NMR was available (Gregory *et al.*, 1987). Velocity centrifugation can also give information on how the molecules interact with each other (self-association) and on the flexibility of the molecule. This includes the contour length L, the persistence length a, and the characteristic ratio C_∞. The ratios, particularly L/a, can be useful for representing the conformations of linear biopolymers (Harding, 1994a).

(2.2.3.2) Sedimentation Equilibrium Experiments.

The second major type of analytical ultracentrifuge experiments involves sedimentation equilibrium for determining molecular weights. One of the most fundamental parameters describing a biological macromolecule is its molecular weight M (unit of g/mol), or the dimensionless relative molecular mass M_r. The molecular weight is straightforward to determine for a homogeneous system. If a protein sequence is known, it is simple to determine the molecular weight by simply adding up the weights of the amino acids. Glycosylation of the protein may make this procedure not possible, and M can only be estimated. Some macromolecular systems may be polydisperse, giving rise to a solution containing molecules of different molecular weights (Chapter 5). Sometimes the molecular weight is difficult to define for a heterogeneous system. If the system is self associating in the experimentally-studied concentration range, it is possible to determine an association constant (Harding, 1994b).

During the progression of an ultracentrifugation experiment a concentration gradient is created. When the rotor speed is great enough, (~60,000 rpm), all the solute molecules will eventually collect in the bottom of the cell. If the rotor speed is lowered to about 10,000 rpm a perfect balance between sedimentation and diffusion processes can be achieved. In diffusion, solute molecules move from a higher concentration to a lower one, while sedimentation reverses this process. When an equilibrium is established, no net flow occurs. At equilibrium, the diffusion rate is equal to the sedimentation rate, so that:-

$$c\frac{dr}{dt}=\frac{RT}{fN_0}\frac{dc}{dr}$$

or

$$c\omega^2rm(1-\bar{v}\rho)=\frac{RT}{N_0}\frac{dc}{dr}$$

Rearranged:-

$$\frac{dc}{c}=\frac{M\omega^2r(1-\bar{v}\rho)}{RT}dr \qquad M=mN_0$$

Integration between $r_1(c_1)$ and $r_2(c_2)$ yields:-

$$\ln\frac{c_2}{c_1}=\frac{M(1-\bar{v}\rho)\omega^2}{2RT}(r_2^2-r_1^2)$$

As with sedimentation velocity ultracentrifugation, optical techniques measure the protein concentrations c_1 and c_2 at r_1 and r_2. If $\bar{v}$, ρ, and ω are known, M can be calculated. Unlike velocity centrifugation the technique does not require any knowledge of the shape of the molecule or its diffusion coefficient. It is therefore one of the most accurate methods for the determination of molecular mass (Chang, 1981).

Sedimentation equilibrium is accurate to ±3% for absolute molecular weights.

It can determine subunit compositions for multisubunit proteins, and average molecular weights for heterogenous systems. Limitations of the technique include the time taken to reach sample equilibrium (2-96 h). Nonideality may be significant for large molecules and may be concentration dependant. Measured molecular weights at a finite concentration will be "apparent" molecular weights which may require dilution series to be performed to see if the "apparent" molecular weight changes (e.g AmiC Chapter 5). The maximum molecular weight of an assembly that can be measured is $M>20\times10^{6}$. This is attributed to the instability of the rotor systems at low speed (Harding, 1994b).

(2.2.4) High-flux X-ray and Neutron Solution Scattering.

Solution scattering is a diffraction technique that can be used to study the overall structure of a wide range of biological systems. Examples include liquid crystalline structures such as the crystallisation behaviour of cocoa butter in chocolate manufacture or the behaviour of detergents (van Gelder *et al.*, 1995), and the study of the overall structure of biological macromolecules in the solution state (e.g Glatter and Kratky, 1982; Perkins, 1988a; Perkins 1988b; Chamberlain *et al.*, 1997; Chamberlain *et al.*, 1998a; Chamberlain *et al.*, 1998b). X-rays and neutrons interact with matter in different ways. X-rays are diffracted by electrons and neutrons are diffracted by nuclei, however the physical principles are the same.

Under ideal conditions solution scattering views structures in random orientations to a resolution of about 2-4 nm in a Q range between about 0.05 and 3 nm^{-1} (Figure 2.2) (Perkins, 1994). This resolution is of course low compared to the atomic detail of X-ray

Table 2.2: Stages in the determination of a protein structure by SAXS and SANS. (Perkins 1994; Mayans *et al.*, 1995; Beavil *et al.*, 1995).

SAXS	SANS
Sample preparation ↓ Data collection ↓ Data analysis and determination of molecular weight, radius of gyration and radius of cross-section ↓ Model building using an automated curve fitting procedure.	Sample preparation ↓ Data collection in 100% H_2O, 100% ^2H_2O and in ratios of $H_2O/^2H_2O$ (solvent contrast experiments) ↓ Data analysis and determination of absolute molecular weight, radius of gyration and radius of cross-section ↓ Model building using an automated curve fitting procedure.
PROS ●Closer to biological conditions. ●Raw-data processing highly automated. ●Scattering curves can be calculated from crystal structures. ●Constrained modelling improves data interpretation. ●Quality indicators available (R-factor - comparing low resolution models to low resolution data). ●Useful for large flexible multidomain proteins, e.g complement factors B, I and H. ●Free from artefacts resulting from crystallization. ●Determination ofabsolute molecular weight. ●Can be used to determine quaternary structure. ●Can be used to monitor conformational change on ligand binding e.g periplasmic binding proteins. ●Solution conditions can be explicitly chosen and readily changed, e.g. pH, temperature, etc.	**PROS** As for SAXS plus:- ●Solvent contrasting:- can visualise the separate protein, lipid or carbohydrate component depending on the % of ^2H_2O used in the scattering experiment. ●Neutrons are non destructive :- sample can be used again. ●Dimensions of the macromolecule correspond to the macromolecular structure observed by protein crystallography.
CONS ●Low resolution technique, much lower than crystallography and NMR. ●Requires a concentrated protein solution. ●Modeldoes not represent a unique structure. ●X-rays can cause radiation damage making the protein unusable after measuring.	**CONS** ●Low resolution technique, much lower than crystallography and NMR. ●Requires a concentrated protein solution. ●Model does not represent a unique structure. ●^2H_2O may cause protein to behave differently to that seen in H_2O buffer.

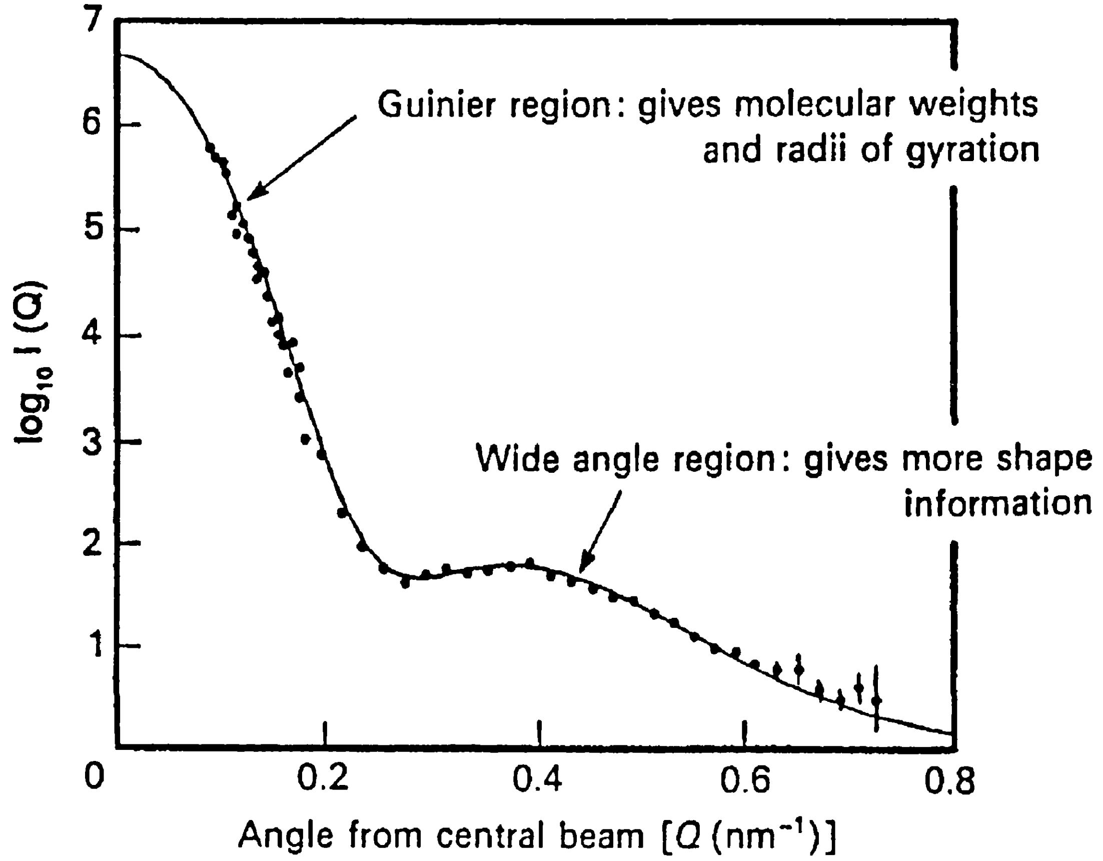


Figure 2.2: General features of a solution scattering curve I(Q) measured over a Q range (Perkins, 1988). The neutron scattering curve of a protein in 100% ^2H_2O buffer is analyzed in two regions, that at low Q, which gives the Guinier plot from which the overall radius of gyration (R_G) and the forward 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. The curve was measured at 2 distances (2.7 and 10.7 m), where the shorter distance gives rise to the maximum Q range.

crystallography and NMR (see Tables 2.1 and 2.2 for comparisons). Here $Q = 4\pi \sin \theta/\lambda$, where 2θ = scattering angle and λ = wavelength. Q is a measure of the scattering angle. Note that $Q = 2 \pi/d$, where *d* is the diffraction spacing specified in Bragg's Law of Diffraction: $\lambda=2 d \sin \theta$.

Analyses of the scattering curve I(Q) measured over a range of Q lead to the molecular weight and the degree of oligomerization from I(0), the overall radius of gyration R_G (and in certain cases, those of cross-section and the thickness), and the maximum dimension of the macromolecule. Using these parameters, small angle scattering can also be used to monitor conformational changes of (for example) periplasmic binding proteins upon binding a ligand (see Chapter 5), of the quaternary structure and allosteric activity of aspartate transcarbamylase (Fetler *et al.*, 1995) and protein folding and denaturation (e.g Kataoka *et al.*, 1995; Doniach *et al.*, 1995; Konno *et al.*, 1995). X-ray and neutron scattering has been successfully used to determine the conformations of myosin subfragment 1 ATPase intermediates (Mendelson *et al.*, 1995), and protein-DNA interactions (van Holde and Zlatanova, 1995; Olah *et al.*, 1995).

Scattering analyses can be quantitatively compared with other physical data in order to check and refine the results. These other methods include electron microscopy, determinations of sedimentation or diffusion coefficients, crystallography, and molecular graphics modelling. The main advantage of solution scattering in biology is that it is the only method that offers a multiparameter characterization of the gross structural features of macromolecules in a physiological environment. Electron microscopy does have the ability to view structures directly, but it has the disadvantages in that the preparative techniques required to prepare the sample, electron beam damage, magnification errors and the need to work *in vacuo* may cause the sample to be prone to artefacts and difficult to interpret. Hydrodynamic methods only give a single structural parameter based on the sedimentation coefficient, which reports only on the degree of structural elongation. As mentioned above, even though protein crystallography will give structures at atomic resolution, crystallization buffers are usually unphysiological, and there will be no information on how the structure behaves in solution. Solution conformation is particularly important when considering multidomain structures with flexible linkers between the domains which may make crystallisation impossible (Perkins, 1994).

(2.2.4.1) <u>Comparison Between X-rays and Neutrons.</u>

X-ray and neutron scattering techniques are complementary in many respects. X-ray scattering has the following characteristics (Perkins, 1994):-

1. Most biological macromolecules are studied in high positive solute-solvent contrasts. This contrast corresponds to the situation in which the scattering density of the macromolecule is significantly higher than that of the solvent. It has been found that this minimises the systematic errors in the curve modelling of the proteins that result if internal density fluctuations in the protein are neglected.
2. Good counting statistics are obtained in this contrast despite high background levels in the buffer curves, unlike neutron scattering in H_2O where the high incoherent scattering background of the buffer is a handicap.
3. Errors caused by wavelength polychromicity and beam divergence are not significant for synchrotron X-ray scattering, so Guinier and wide-angle analyses are not affected by

systematic errors caused by instrument geometry.

4. The hydrated dimensions of the macromolecule are studied, so the structure is larger by an additional depth of 0.36 nm at the surface to correspond to a monolayer of bound water molecules. For proteins this is usually about 0.3 g H_2O/g macromolecule.

Neutron scattering has the following characteristics (Perkins, 1994).

1. Contrast variation in mixtures of H_2O and ^2H_2O permits the analysis of hydrophobic and hydrophilic regions within proteins and glycoproteins, and the elucidation of the disposition of detergents or lipids with solubilized membrane proteins, or that of DNA or RNA in complexes with proteins (Figure 2.3). Deuteration of components in a multicomponent system can extend these methods.

2. No radiation damage effects are encountered. This can be a severe problem with synchrotron X-rays . The neutron samples can normally be recovered for other studies.

3. The dry dimensions of the macromolecule are studied and correspond to the macromolecular structure observed by protein crystallography.

4. Absolute molecular weight calculations are obtained from neutron data in H_2O, or by the use of a deuterated polymer standard, in place of the relative determinations by synchrotron X-ray scattering. The latter are based on a protein of known molecular weight with a reliable 280 nm absorption coefficient to determine concentrations.

5. Background scattering effects are very low in ^2H_2O buffers, even in the presence of high salt concentrations, and this permits studies of macromolecules at low concentrations (0.5 mg/ml). ^2H_2O is, however, a promoter of macromolecular aggregation if hydrogen bond interactions with water are important for solubility.

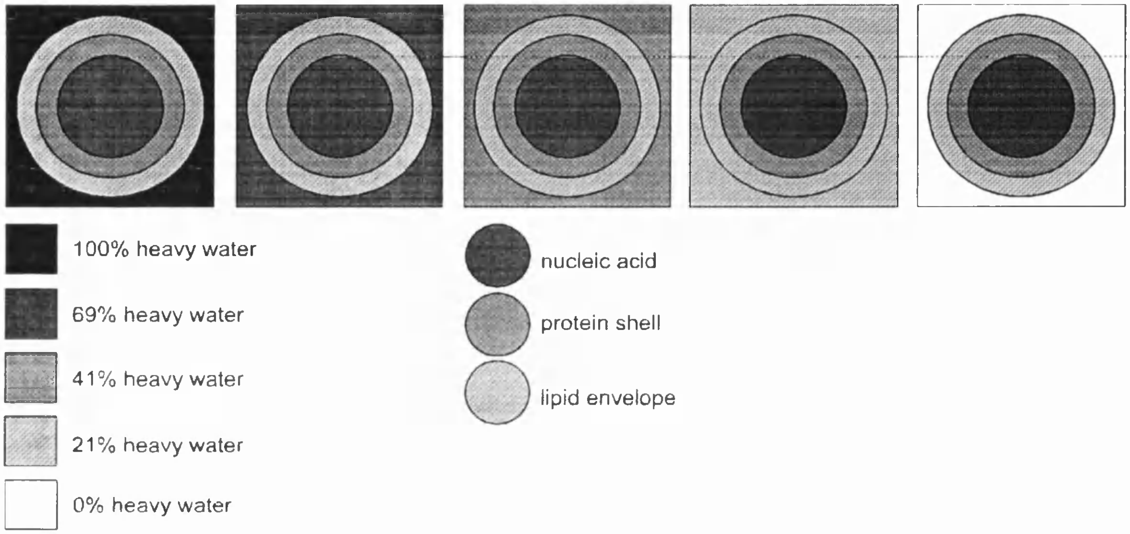


Figure 2.3: Contrast matching. By varying the $^2H_2O/^1H_2O$ ratio of a solution over a series of neutron experiments, different components can be matched out in turn. The enveloped "virus" shown above is examined in five different $^2H_2O/^1H_2O$ ratios. Each of the intermediate concentrations have been chosen to have the same neutron scattering density as one of the principle components. (Based on Bradshaw 1995).

6. Guinier analyses at low scattering angles are not significantly affected by beam divergence or wavelength polychromicity (9-10% on D11 at the ILL). However, intensities at large Q are noticeably affected and this requires consideration in curve simulations.

(2.2.4.2) <u>Sample Requirements.</u>

Samples need to be at biochemical standards of purity in monodisperse solution at a concentration high enough for a scattering curve to be observable in the required solute-solvent contrast. For studies on a single preparation, 0.5 ml of material at 10 mg/ml is ideal for synchrotron X-ray work, and 1.5 ml at 10 mg/ml for neutron work at 3 contrasts. Because scattered intensities are proportional to the square of the molecular weight at low Q, it is essential to remove all traces of aggregates prior to measurement by gel filtration (microfiltration through 0.2 or 0.5 μm membranes is not acceptable) and reconcentration of the samples if the sample is prone to aggregation (Perkins, 1994). This may of course mean that samples at 10 mg/ml are impossible so a minimum usable starting concentration for a sample is 2 mg/ml and such concentrations had to be used for some of the proteins studied in this thesis

Dialysis of samples prior to measurements was essential for accurate subtraction of buffers from the samples, so the final dialysate was always used as the buffer. Slight differences in the electron densities of the buffer for X-rays, or exchangeable protons for neutron work, can invalidate the buffer subtraction. Buffer content can also affect scattering intensities of a solution. For X-ray work, the closer a buffer is to pure water, the higher the sample transmission becomes, and the better the counting statistics.

Phosphate buffered saline is commonly used (12 mM phosphate, 140 mM NaCl, pH 7.4). For neutron work a reduction of the proton content of the buffer improves the counting statistics because of the strong incoherent scattering of ^1H nuclei. Proteins are usually measured in 0, 80, and 100% ^2H_2O buffers. Sample concentrations must be accurately known prior to measurements for molecular weight and neutron matchpoint calculations (Perkins, 1994). Neutron matchpoints correspond to the percentage ^2H_2O in which the scattering density of the macromolecule is the same as that of the solvent (Figure 2.3) (Bradshaw, 1995).

(2.2.4.3) Instrumentation.

Instrumental requirements for solution scattering are based on the irradiation of a solution of path thickness 1-2 mm with a collimated, monochromatized beam of X-rays or neutrons, and recording the scattering (or non-crystalline diffraction) pattern with a detector linked to a computer. Detectors (or scattering cameras) are provided and maintained as a multiuser facility by the institute providing the X-ray or neutron beams (Perkins, 1994).

(2.2.4.4) X-ray Scattering at SRS, Daresbury.

A synchrotron X-ray detector (such as those at Stations 2.1 or 8.2 at SRS, Daresbury (Figure 2.4); Nave *et al.*, 1985; Towns-Andrews *et al.*, 1989; Worgan *et al.*, 1990) takes a "white" beam of X-rays, which is emitted tangentially by the electrons circulating at sub-light speeds in the storage ring of the synchrotron (Figures 2.5 and 2.6). Electrons are emitted from a hot cathode and accelerated in a linear accelerator (LINAC) to 12 million electron volts (12 MeV). The electrons are further accelerated

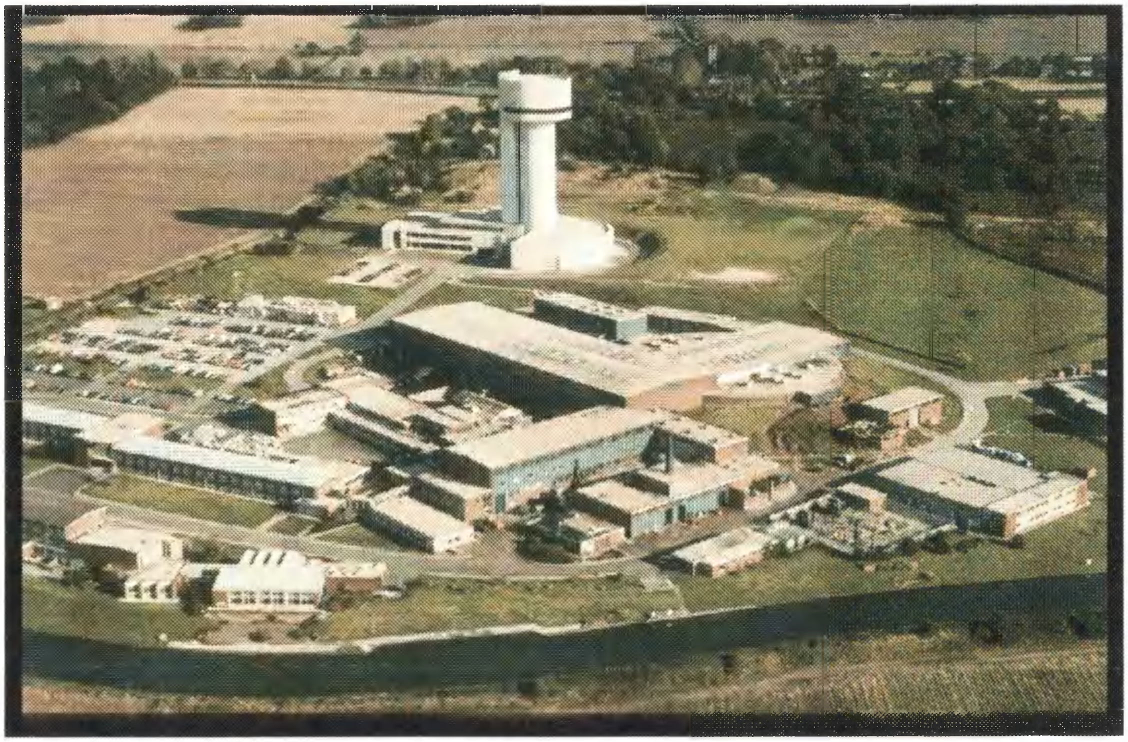


Figure 2.4: Aerial view of the Daresbury Laboratory, Warrington, U.K. (Taken from the Daresbury Laboratory World Wide Web Site http://www.dl.ac.uk).

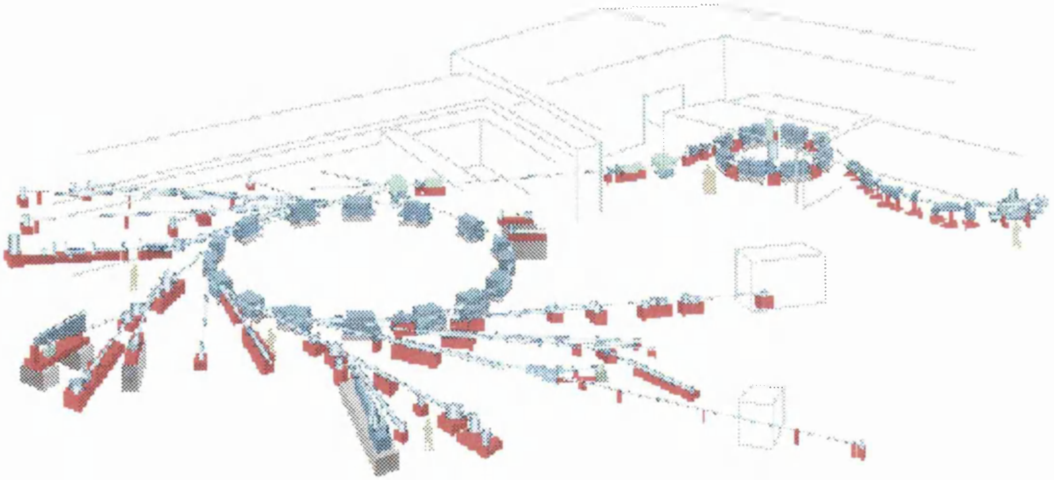


Figure 2.5: Schematic diagram of the Synchrotron Radiation Source at Daresbury, Warrington. The LINAC is on the right, the booster in the middle and the storage ring, beamlines and experimental areas on the left. (Taken from the Daresbury Laboratory World Wide Web Site http://www.dl.ac.uk).

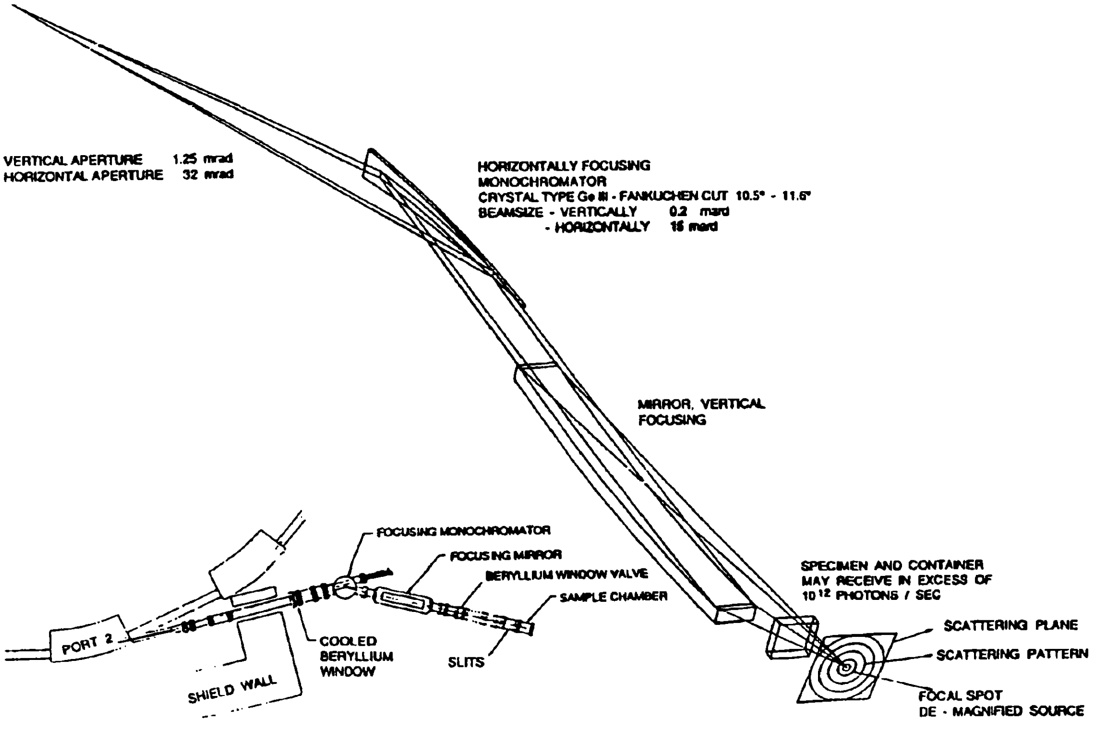


Figure 2.6: Schematic layout of the X-ray solution scattering camera at station 2.1 at the SRS Daresbury (Towns-Andrews *et al.*, 1989; Perkins, 1994). The station operates at 0.154 nm using a monochromator-mirror optical system, which reduces the heat loading on the mirror in the more conventional mirror-monochromator arrangement. A focal spot of size 0.3 × 2.5 mm^2 is produced, with a beam cross-section of 1 × 5 mm^2 at the sample position. The optics are *in-vacuo* 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 with either a linear, quadrant, or area detector that is interfaced to a computer. Inset at the lower left is an overall view showing how the X-ray beam is taken from the synchrotron storage ring.

in a booster synchrotron to 600 MeV and then injected into the storage ring and finally accelerated to 2000 MeV (2 GeV, 99.999997% of the speed of light) by a high power radio-frequency accelerating system. This also maintained the energy at this level. The beam current varied in the range of 122-173 mA. The path followed by the electrons in the storage ring is bent into a circle by 16 dipole magnets and has a 96 m circumference. The electrons travel round the circumference 3.12 million times a second and can remain in orbit for up to 30 h. The small angle scattering instruments are located on beamlines 2 (Figure 2.6), 7 and 8. Synchrotron radiation is emitted by the electrons when they are deflected by the magnetic field. The X-ray beam is horizontally focussed and monochromated to a wavelength of 0.154 nm by a single perfect crystal of (usually) Ge or Si, then vertically focussed by a curved mirror, and collimated by slits before and after the monochromator and mirror (Towns-Andrews *et al.*, 1989). Wavelength spread is negligible as a result of the monochromatisation.

Samples (1 mm path length; surface area 2 × 8 mm; total volume 25 μl) are held in a PTFE and perspex sample cell with 10-20 μm thick ruby mica windows. The sample cell is held in a brass thermostatically-controlled sample holder. The sample holder is aligned in the beam by the use of "green paper," which turns red when exposed to X-rays. Sample detector distances (0.5-5.0 m) can be used depending on the desired Q range. For all the studies presented in this thesis, sample-detector distances of 3.14 m or 3.17 m were used. The position sensitive detector is a quadrant detector, which measures the scattered intensities in a two-dimensional angular sector (70°) of a circle with the nominal position of the main beam located at the centre of the circle. Quadrant detectors can give improved counting statistics at large scattering angles, and a larger Q range is available, compared to the older linear detectors. A quadrant detector was used on station 2.1 for all the studies presented here. Beam exposures are monitored by the use of ion chambers positioned before the sample to check the quality of the beam and after the sample to enable normalisation of the scattering. This is the equivalent to the sample transmission measurements of neutron scattering. The detector and data logging system is interfaced with a Sun computer system for data storage and on-line processing to assess the experimental data as it is being recorded. The X-ray scattering apparatus is inside a radiation-shielded hutch, protected by safety interlocks to avoid accidental lethal exposures to users.

Before data collection started, the instrument was calibrated. The Q range on the detector is defined using slightly wet stretched collagen (rat tail tendon). The response of the detector channels are not uniform. This was calibrated by exposure for several hours to a uniform radioactive ^{55}Fe source when there was no beam. Because the main beam diminishes in intensity during a session and the background intensity is high at low Q, small buffer subtraction errors occur frequently. These errors were minimised by running the samples in duplicate with a buffer run in between them (all in the same cell with the same mica windows) for 10 minutes. Radiation damage was monitored by recording the data in sets of 10 time frames during the measurement and examining the 10 subcurves for time-dependent effects (Perkins, 1994). It was common to find that the data from some proteins was unusable due to radiation damage, but use of the first time frame alone would avoid this problem. It has been proposed that additives such as 100 mM formate be added as a means of reducing damage and aggregation caused by X-ray induced free radicals (Zipper *et al.*, 1985; Durchschlag and Zipper, 1988).

(2.2.4.4.1) Preliminary Data Reduction.

Data reduction at Daresbury utilized the FORTRAN software OTOKO (Figure 2.7) (Bendell, P., Bordas, J., Koch, M.H.C., and Mant, G.R., EMBL Hamburg and CLRC Daresbury Laboratory, unpublished software) running on a Sun workstation. The .DIN procedure normalised all the spectra to the counts in the back ion chamber. This corrected for beam flux, transmission of the sample and exposure times. The buffer background was subtracted from the spectra using .ADD to give the scattering curve due to the sample only. The resultant spectra were normalised with the detector response using the .DIV and finally the 10 time frames were averaged together with .AVE. The spectra could be plotted with the .PLO (plots out averaged spectra) or .PL3 (plots out individual time frames). The x-axis contained an artificially produced gap due to the electronics (Figure 2.8). This was removed using .XSH by specifying the position of the beginning of the gap and the size of the gap in pixels. After June 1997 the station software automatically removed this gap so the .XSH procedure is not needed for processing of data after this date. The Q-axis was calibrated using .XAX from the collagen diffraction pattern. The spectrum consisted of a series of peaks of diffraction spacing 67 nm with the major peaks being the 1st, 3rd, 5th, and 9th order reflections. The position in pixels of all the peaks was determined. Since the number of pixels between any two peaks was approximately the same, the position of the 0th order peak (Q=0) could be determined by extrapolation. The pixel position and Q value ($Q=2\pi \times$ order/67) of another peak, usually the ninth order, was calculated. The pixel position and Q value of these two peaks were used by .XAX to produce a Q axis file. RECONV converted the binary OTOKO files to card image files. DOTKO combined the Q axis file with the spectrum intensity files for final analysis using SCTPL5.

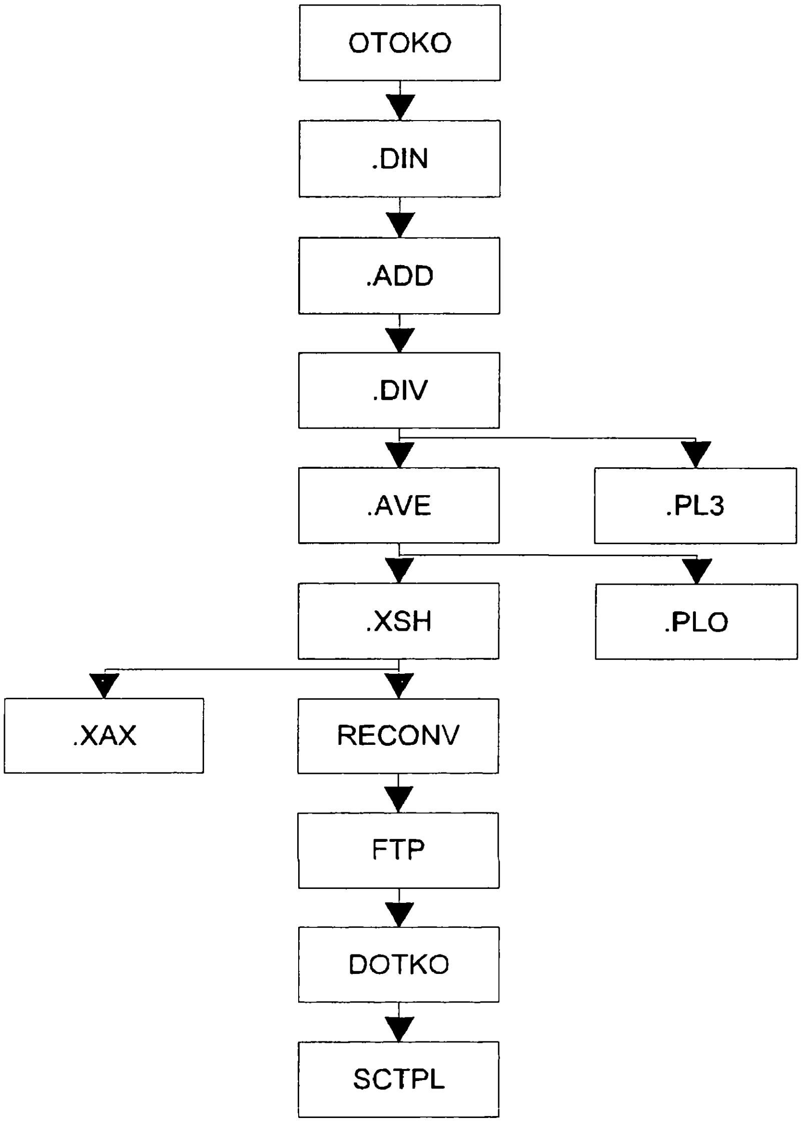


Figure 2.7: Flow diagram of the data reduction procedures for X-ray scattering. The .DIN procedure normalises all the spectra to the monitor counts measured in the back ion chamber. The buffer background is subtracted from the sample plus buffer spectra using .ADD. The resultant spectra are normalised to the detector response using the .DIV and the 10 individual time frames are averaged together with the .AVE. The spectra are plotted either with the .PLO (averaged spectrum) or .PL3 (individual time frames). The detector gap is removed using .XSH (Figure 2.8). The .XSH procedure is not required for data collected after June 1997 because the gap is automatically removed. The Q axis is calculated using .XAX from the diffraction pattern of wet, slightly stretched collagen. RECONV converts the binary OTOKO files to card image files for transfer via FTP to London. DOTKO combines the Q axis file with the spectrum intensity files for final analyses using SCTPL. (Adapted from Meyer, 1994).

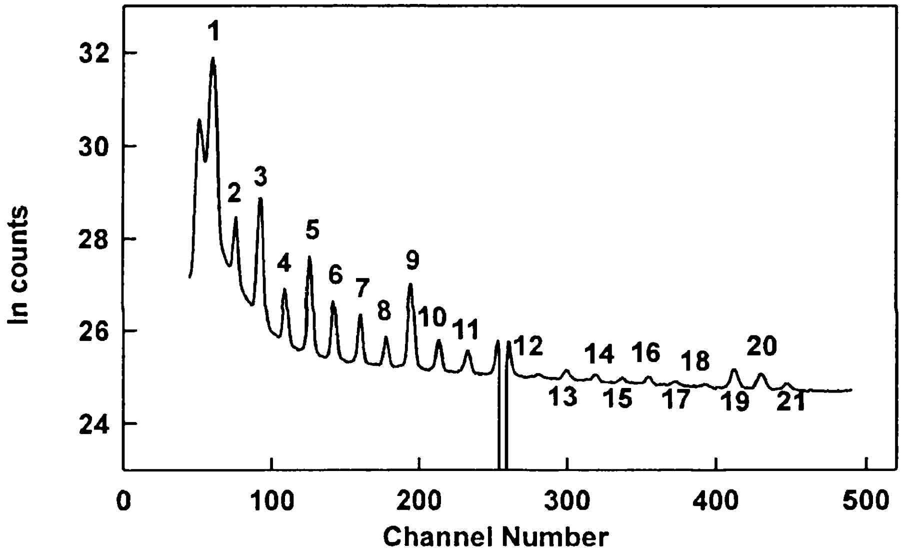


Figure 2.8: Collagen diffraction pattern. A total of 21 diffraction peaks are visible here, each of which is $2\pi/67$ nm^{-1} apart. The channel numbers are converted into Q values using .XAX. The gap between peaks 11 and 12 is artificial and is removed during data reduction by determining the position of the beginning of the gap and the size of the gap in pixels. This gap does not exist for data collected after June 1997. (Taken from Meyer, 1994).

(2.2.4.5) Neutron Scattering at ILL, Grenoble.

Neutron scattering studies were carried out at the high flux reactor at the Institute Laue-Langevin (ILL) in Grenoble, France (Figures 2.9, 2.10 and 2.11). High energy neutrons were produced by the fission of U^{235}. The thermal neutron flux was moderated by either a hot (graphite) or cold (deuterium at 25K) source to enhance neutron intensities at low wave lengths ($0.4 < \lambda < 0.08$ nm) or at longer wavelengths used for solution scattering of $\lambda > 0.4$ nm, respectively. Neutron guides transferred the neutrons from the reactor core to the external instruments in the guide hall (Figures 2.12 and 2.13). Neutrons were selected using velocity selectors and were collimated (Figures 2.14 and 2.15) before reaching the sample areas. The small angle scattering instruments used were D11 (Figures 2.16 and 2.17) and D22 (Figures 2.18 and 2.19).

(2.2.4.5.1) Preliminary Data Reduction.

A number of calibration and normalisation measurements were performed during the processing of any ILL data run (Ghosh, 1989) as shown in Figure 2.20. DETEC listed the raw counts from the detector cell by cell and was used to calculate the position of the beam stop. RNILS calculated and stored the radial intensity function I(Q) of the detector. Individual cells were averaged at a given radial step length spacing of 1 cm to give a mean Q at that step and the intensity I(Q). SPOLLY (Figure 2.21) normalised and combined individual spectra to obtain the final output. All spectra were normalised with respect to the monitor counts (counting time). The role of the cadmium sample was to block the direct beam of neutrons and hence estimated the neutron and electronic background counts in the guide hall to be subtracted from each spectrum. Each sample spectrum was corrected by subtracting the buffer background. These were normalised

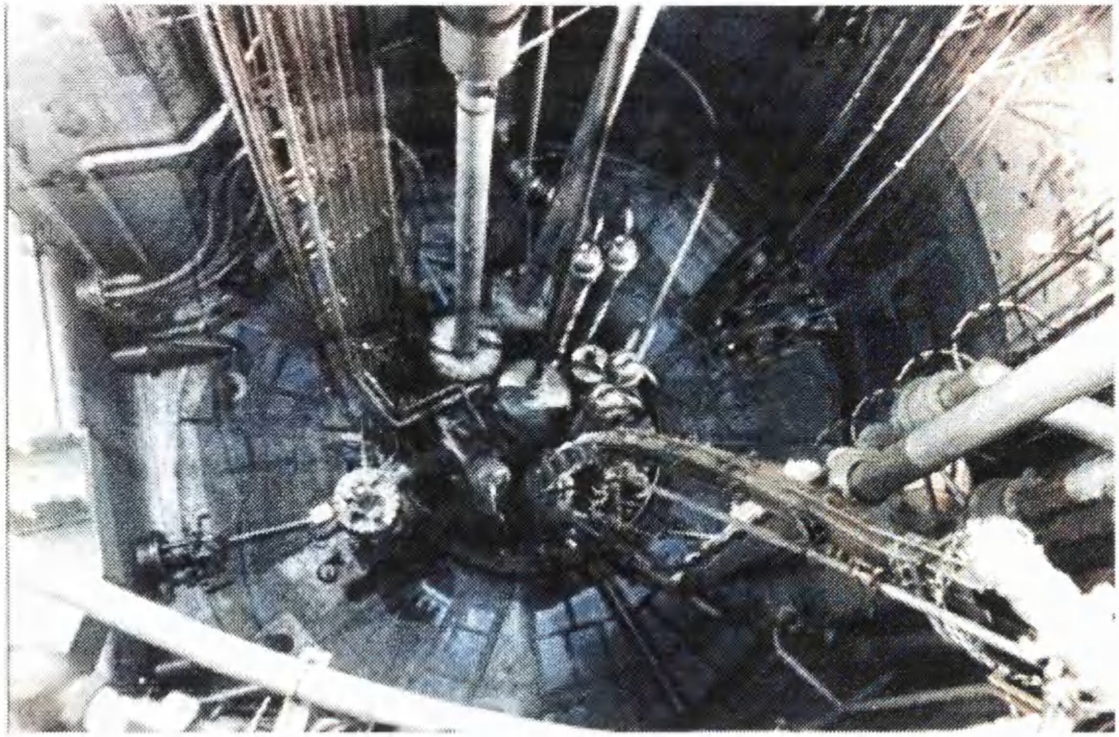
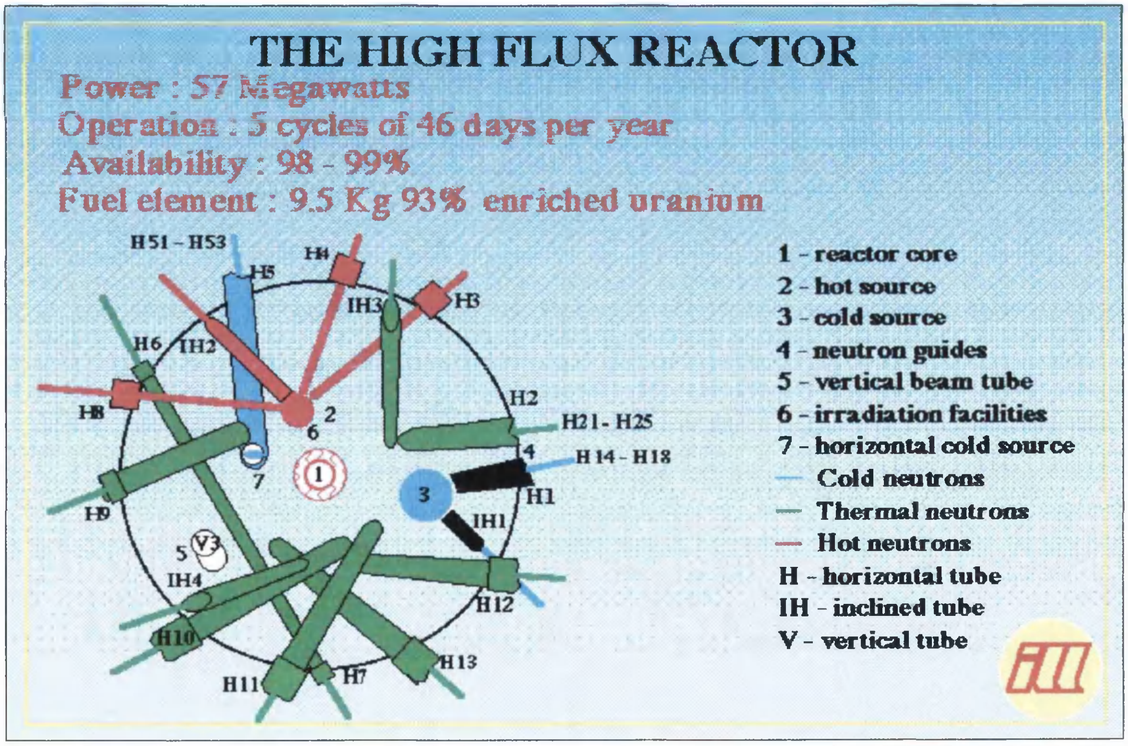


Figure 2.9: The view from above the heavy water vessel. The blue light is Cerenkov radiation. (Taken from the ILL World Wide Web Site http://www.ill.fr).



THE HIGH FLUX REACTOR

Power : 57 Megawatts

Operation : 5 cycles of 46 days per year

Availability : 98 - 99%

Fuel element : 9.5 Kg 93% enriched uranium

1 - reactor core

2 - hot source

3 - cold source

4 - neutron guides

5 - vertical beam tube

6 - irradiation facilities

7 - horizontal cold source

— Cold neutrons

— Thermal neutrons

— Hot neutrons

H - horizontal tube

IH - inclined tube

V - vertical tube

Figure 2.10: The high-flux reactor uses a single fuel element with an operating cycle of 46 days (usually 5 cycles per year). There are hot (red), thermal (yellow), and cold (blue) neutrons available from different beam tubes. (Taken from the ILL World Wide Web Site http://www.ill.fr).

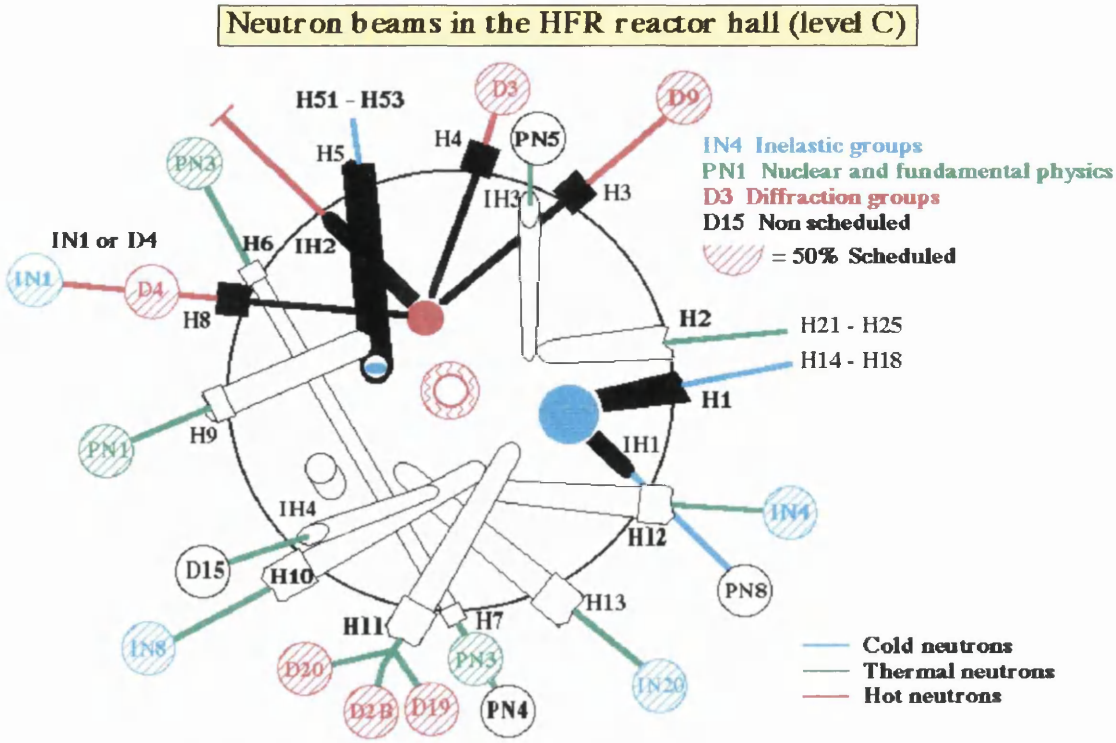


Figure 2.11: Beam tube arrangement at the high-flux reactor. (Taken from the ILL World Wide Web Site http://www.ill.fr).

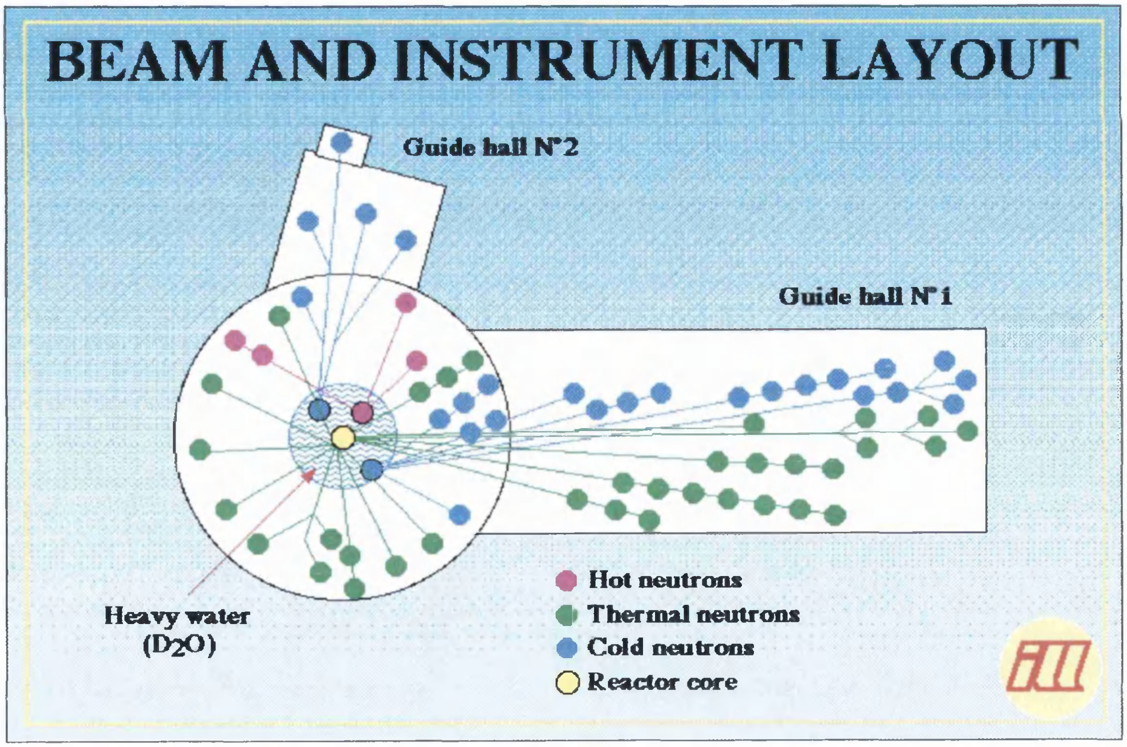


Figure 2.12: Schematic diagram of the guide halls at the ILL in relation to high-flux reactor. (Taken from the ILL World Wide Web Site http://www.ill.fr).

Figure 2.13: Neutron guide hall 1 at the ILL. (Taken from the ILL World Wide Web Site http://www.ill.fr).

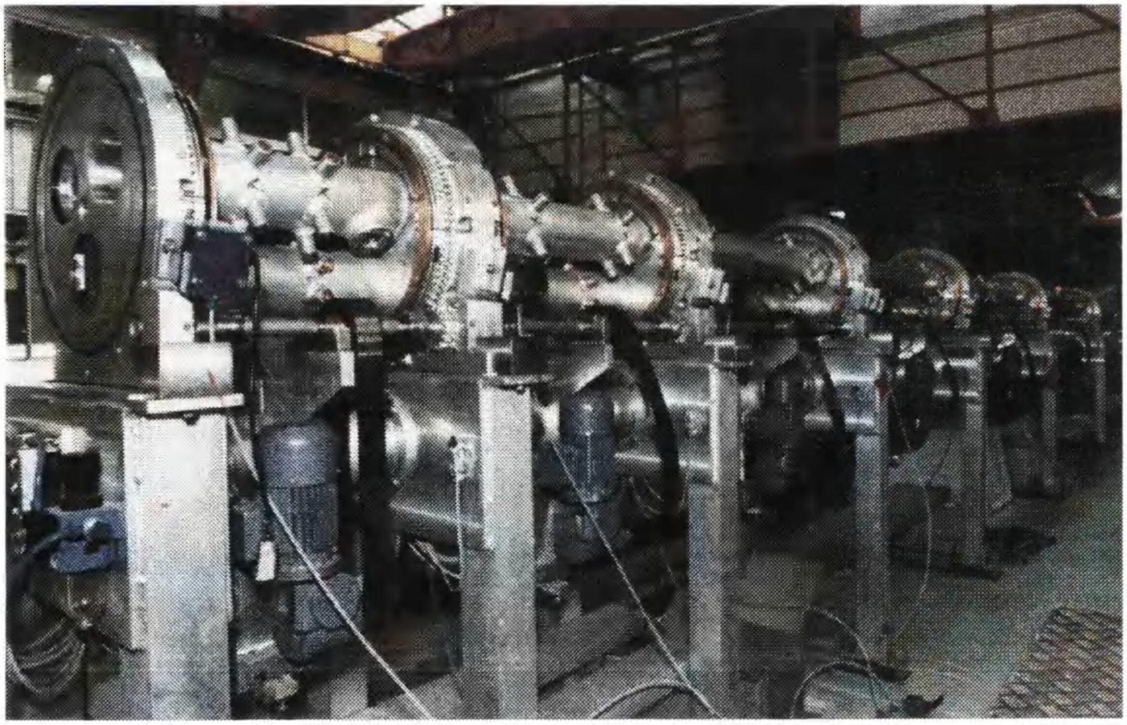


Figure 2.14: Picture of the primary collimation system on D22. (Taken from the ILL World Wide Web Site http://www.ill.fr).

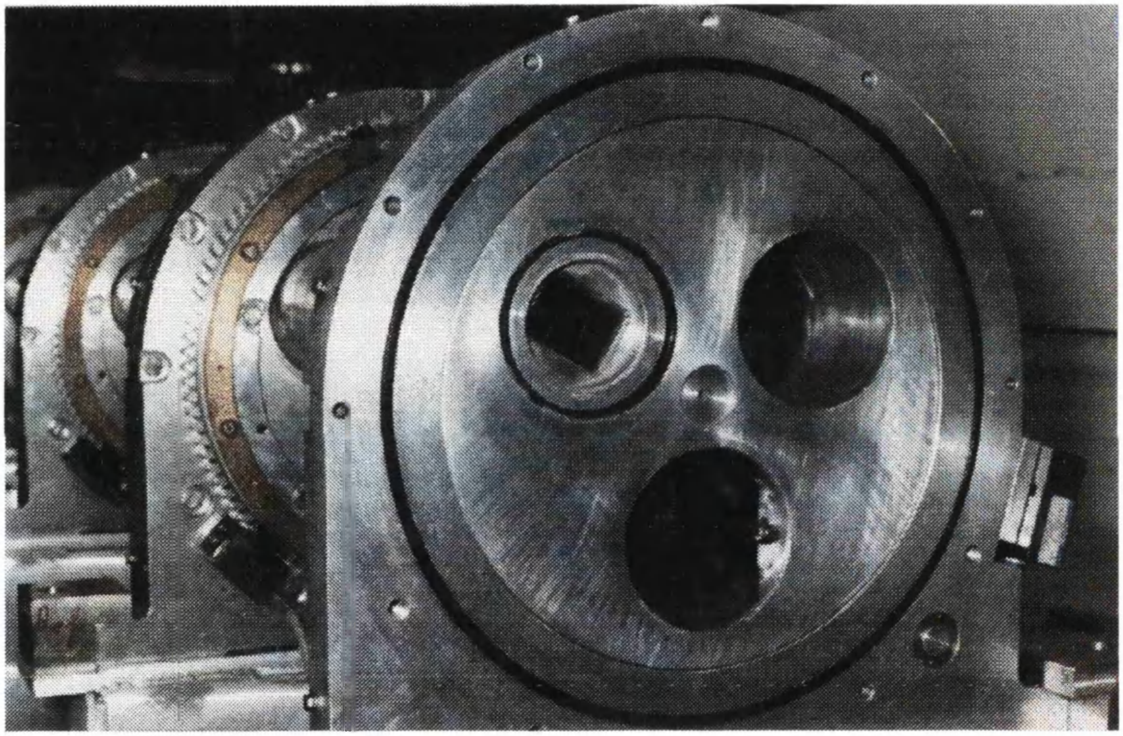
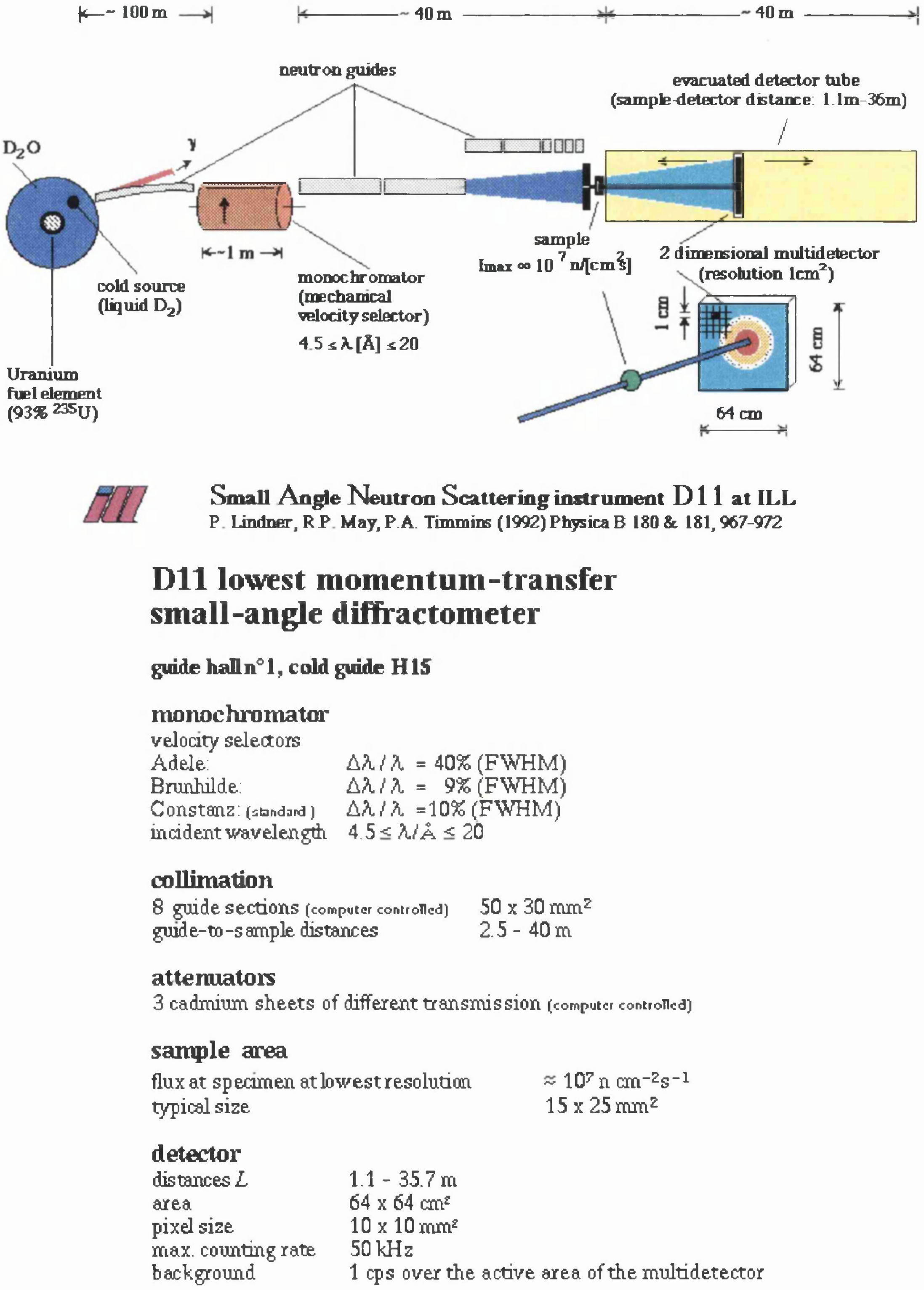


Figure 2.15: Detailed view of the primary collimator. (Taken from the ILL World Wide Web Site http://www.ill.fr).



Small Angle Neutron Scattering instrument D11 at ILL
P. Lindner, R.P. May, P.A. Timmins (1992) Physica B 180 & 181, 967-972

D11 lowest momentum-transfer small-angle diffractometer

guide hall n°1, cold guide H15

monochromator

velocity selectors

Adele:	Δλ / λ = 40% (FWHM)
Brunhilde:	Δλ / λ = 9% (FWHM)
Constanz: (standard)	Δλ / λ =10% (FWHM)
incident wavelength	4.5 ≤ λ/Å ≤ 20

collimation

8 guide sections (computer controlled)	50 x 30 mm^2
guide-to-sample distances	2.5 - 40 m

attenuators

3 cadmium sheets of different transmission (computer controlled)

sample area

flux at specimen at lowest resolution	≈ 10^7 n cm^{-2}s^{-1}
typical size	15 x 25 mm^2

detector

distances *L*	1.1 - 35.7 m
area	64 x 64 cm^2
pixel size	10 x 10 mm^2
max. counting rate	50 kHz
background	1 cps over the active area of the multidetector

Figure 2.16: Schematic diagram and characteristics of D11. (Taken from the ILL World Wide Web Site http://www.ill.fr). A helical velocity selector is used for monochromatization. Movable neutron guides give collimation distances of 2, 5, 10, 20, and 40 m. The beam size at the sample is defined by a diaphragm. The 64 × 64 element BF_3 detector is housed within a 40 m evacuated tube and can be moved by remote control between 2 and 38 m sample-detector distances. The detector is interfaced with a minicomputer for data collection and storage (Ibel, 1976; Lindner *et al.*, 1992).

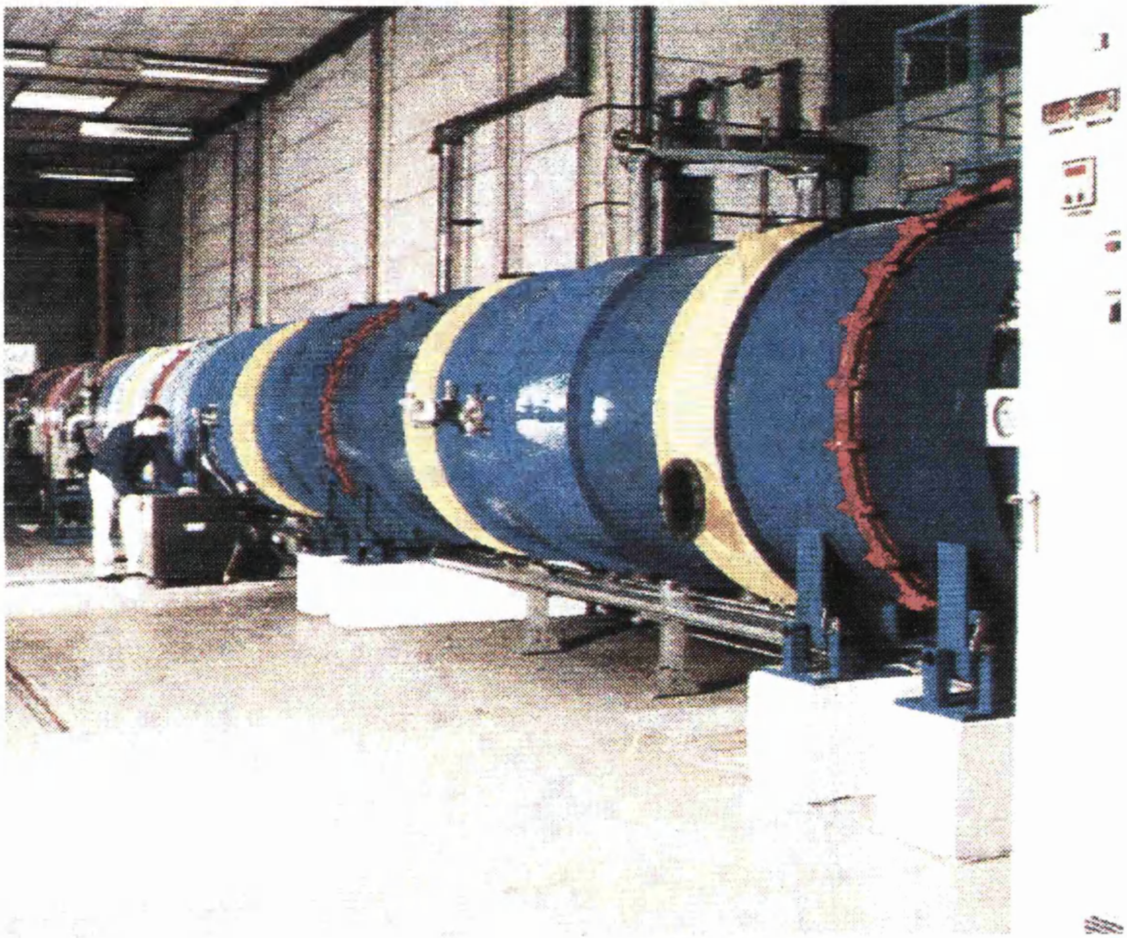
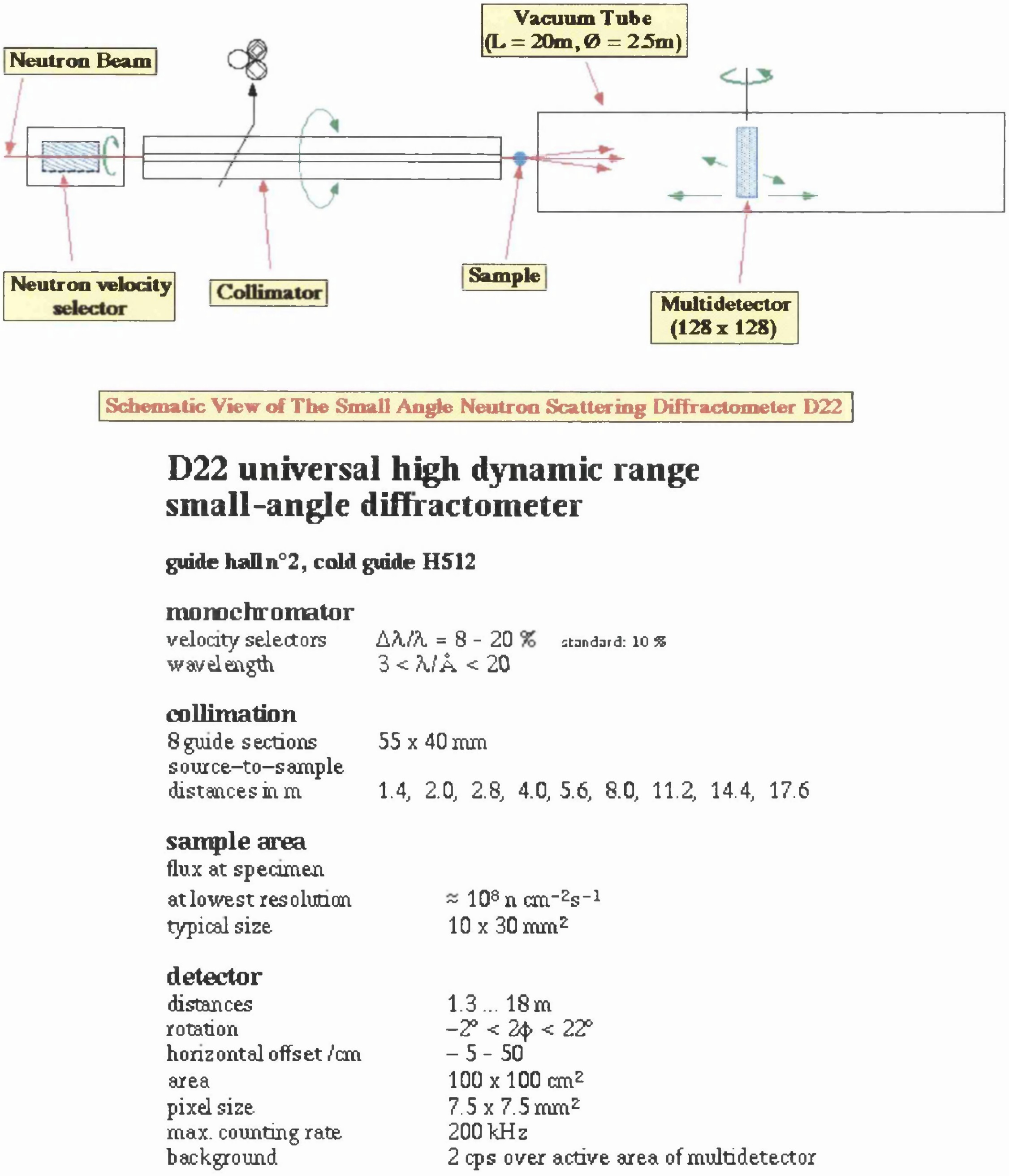


Figure 2.17: The detector tube of D11. (Taken from the ILL World Wide Web Site http://www.ill.fr).


Vacuum Tube
(L = 20m, Ø = 2.5m)
Neutron Beam
Neutron velocity selector
Collimator
Sample
Multidetector
(128 x 128)
Schematic View of The Small Angle Neutron Scattering Diffractometer D22

D22 universal high dynamic range small-angle diffractometer

guide hall n°2, cold guide H512

monochromator

velocity selectors	$\Delta\lambda/\lambda$ = 8 - 20 %	standard: 10 %
wavelength	3 < λ/Å < 20	

collimation

8 guide sections	55 x 40 mm
source–to–sample distances in m	1.4, 2.0, 2.8, 4.0, 5.6, 8.0, 11.2, 14.4, 17.6

sample area

flux at specimen	
at lowest resolution	≈ 10^8 n $cm^{-2}s^{-1}$
typical size	10 x 30 mm^2

detector

distances	1.3 ... 18 m
rotation	–2° < 2ϕ < 22°
horizontal offset /cm	– 5 - 50
area	100 x 100 cm^2
pixel size	7.5 x 7.5 mm^2
max. counting rate	200 kHz
background	2 cps over active area of multidetector

Figure 2.18: Schematic diagram and characteristics of D22. (Taken from the ILL World Wide Web Site http://www.ill.fr).

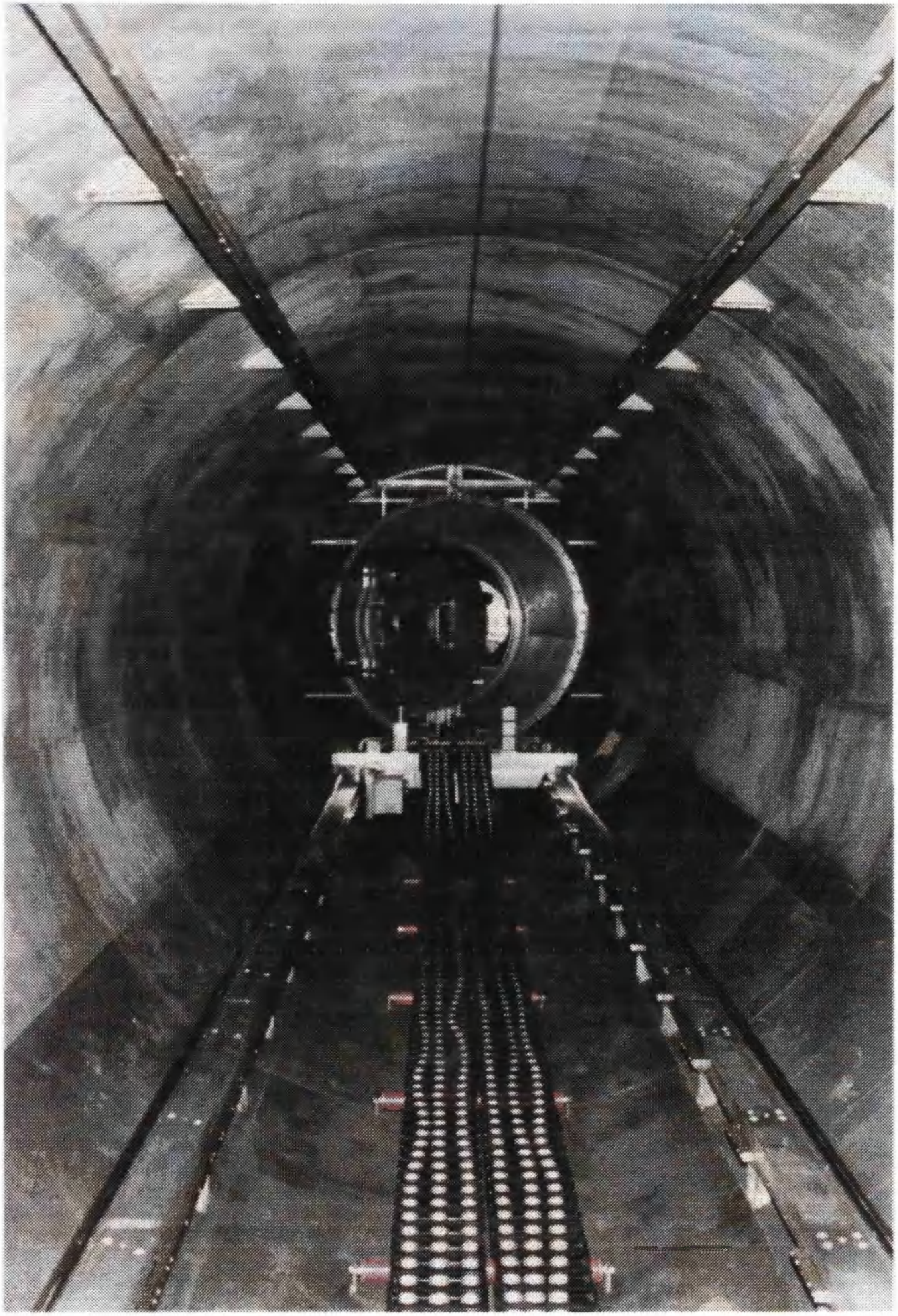


Figure 2.19: D22 multidetector housing inside the vacuum tube. (Taken from the ILL World Wide Web Site http://www.ill.fr).

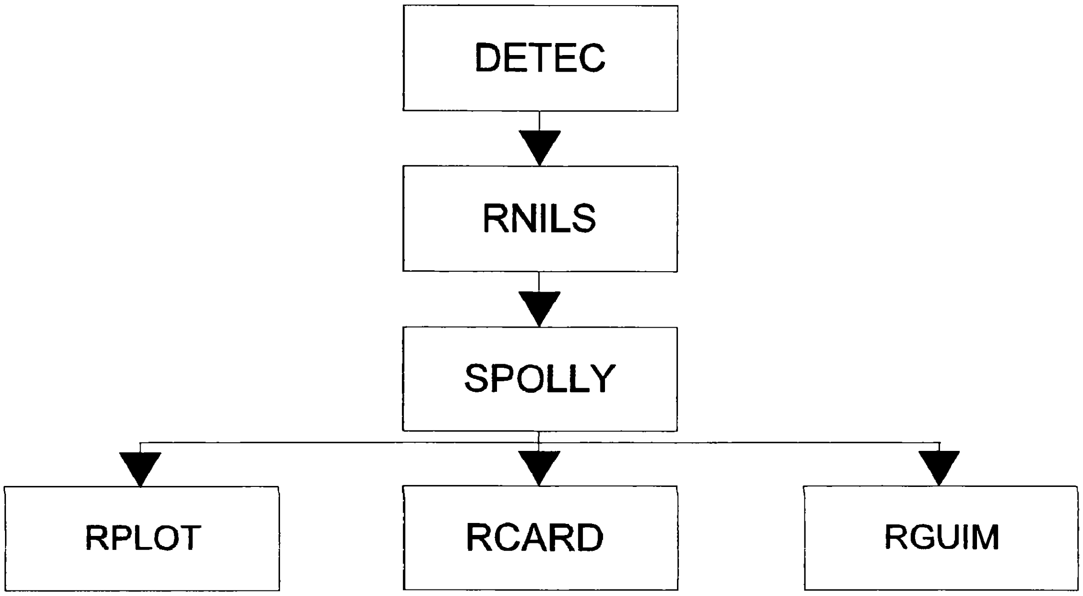


Figure 2.20: Flow diagram showing the data reduction procedures for neutron scattering at D11 and D22. (Adapted from Meyer, 1994).
DETEC Lists the raw counts from the detector cell by cell.
RNILS Calculates the radial intensities function I(Q) of the detector.
SPOLLY Corrects and normalises the sample spectrum I(Q).
RPLOT Plots out two spectra on the same graph.
RCARD Produces a card image disc file of formatted data.
RGUIM Calculates the Guinier plot.

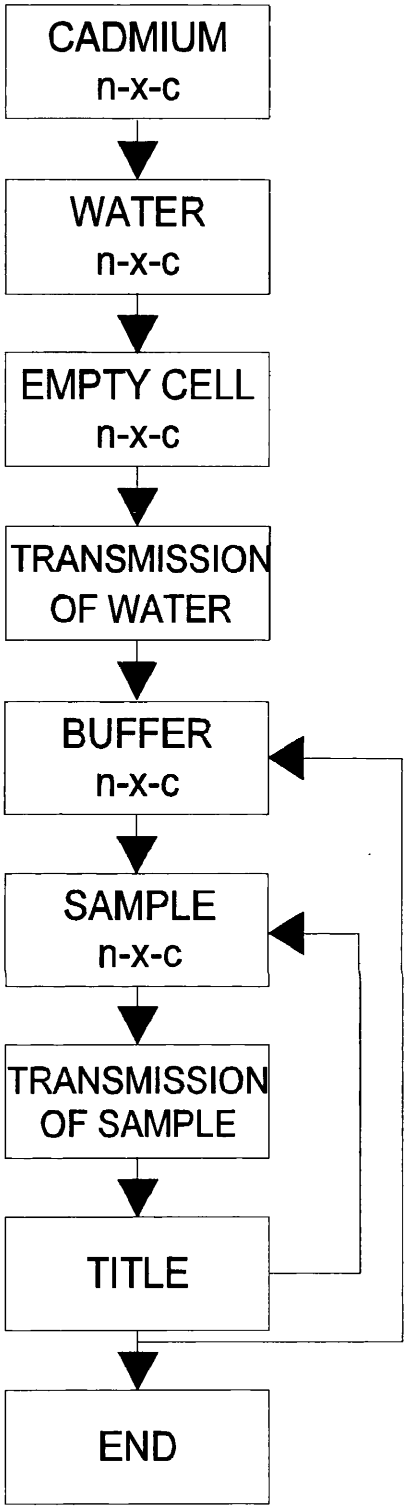


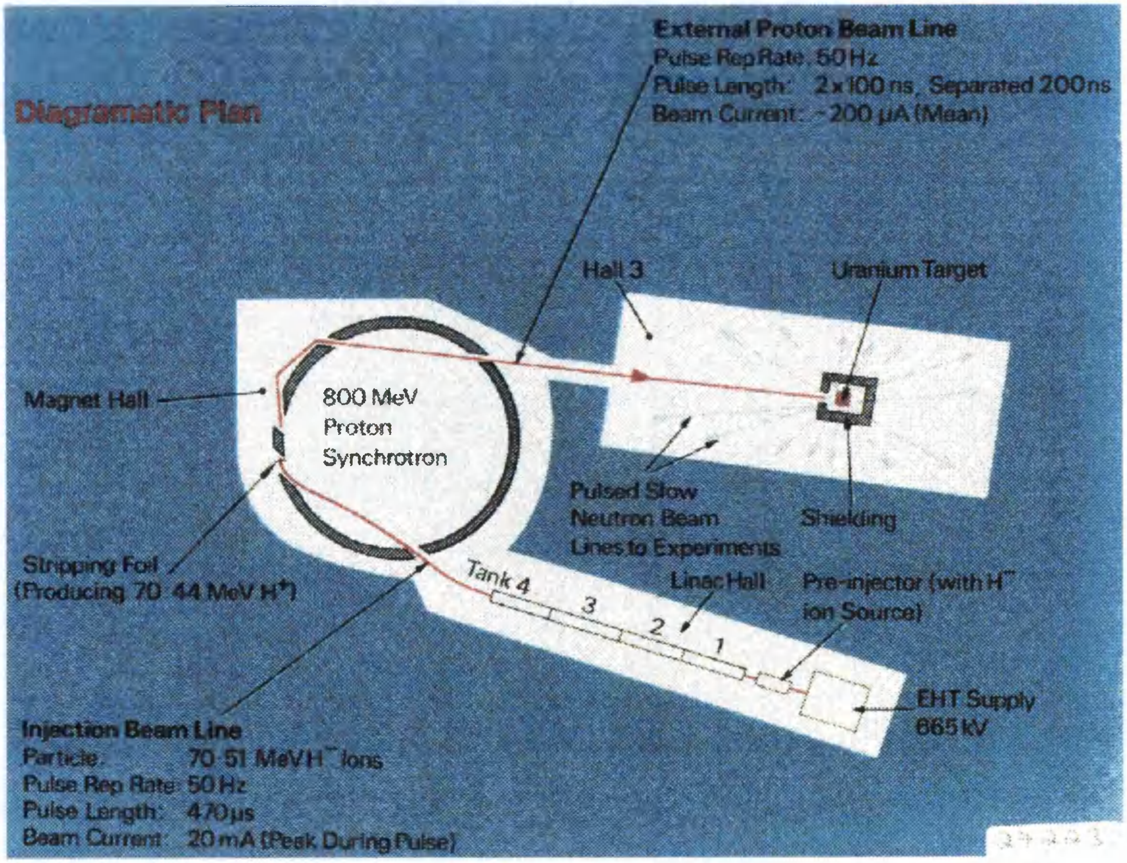
Figure 2.21: Detailed flow chart of SPOLLY. The abbreviation n-x-c refers to the individual run number, its extension number and the monitor counts (proportional to the time in the beam). The transmission of the sample plus buffer and buffer alone are assummed to be equal and the default value of one is used for data reduction. (Adapted from Meyer, 1994).

using the water and empty cell runs. Preliminary R_G and I(0) values were calculated using RGUIM during data acquisition to check the samples had been counted for a sufficient amount of time, and again after full data reduction. RPLOT was a general graphics program and could plot two spectra together. This was used to ensure spectra of the same sample from two different instrument configurations overlapped sufficiently in regions of similar Q. RCARD produced output containing Q, I(Q) and error of I(Q) for transfer via FTP to London. The spectra were reanalysed in more detail in London using the interactive graphics program, SCTPL5.

(2.2.4.6) Neutron Scattering at ISIS, RAL, Didcot.

Neutron scattering data were also obtained on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, U.K. (Figure 2.22) (Heenan and King, 1993). At ISIS (Figures 2.23 and 2.24), the pulsed neutron beam (typically at 170 μA) is emitted from a uranium or tantalum target after proton bombardment at 50 Hz and is moderated by a liquid $^2H_2$ cold source on station LOQ (Figures 2.25 and 2.26). This cools down the neutrons which then travel at slower speeds and so increases the number of neutrons at longer wavelengths, which is important for Guinier measurements. The wavelength of the neutrons is kept between 0.2 to 1.0 nm by a supermirror bender which cuts out the wavelengths of lower than 0.2 nm, and an overlap mirror which removes neutrons with wavelengths greater than 1.0 nm. A chopper cuts out every other neutron pulse (Figure 2.27), so LOQ operates at 25 Hz. The resulting polychromatic pulsed beam can then be correctly resolved into wavelengths in the range 0.2 - 1.0 nm using time of flight techniques to achieve monochromatisation. A ^{3}He ORDELA detector was employed to record neutron scattering patterns at a fixed

Figure 2.22: Aerial view of the ISIS Facility at the Rutherford-Appleton Laboratory, Didcot. (Taken from the Rutherford-Appleton Laboratory Web Site http://www.rl.ac.uk).



Diagramatic Plan
External Proton Beam Line
Pulse Rep Rate: 50 Hz
Pulse Length: 2 x 100 ns, Separated 200 ns
Beam Current: ~200 μA (Mean)
Hall 3
Uranium Target
Magnet Hall
800 MeV Proton Synchrotron
Pulsed Slow Neutron Beam Lines to Experiments
Shielding
Stripping Foil (Producing 70 44 MeV H⁺)
Tank 4
3
2
1
Linac Hall
Pre-injector (with H⁻ ion Source)
EHT Supply 665 kV
Injection Beam Line
Particle: 70 51 MeV H⁻ Ions
Pulse Rep Rate: 50 Hz
Pulse Length: 470 μs
Beam Current: 20 mA (Peak During Pulse)

Figure 2.23: Layout of the LINAC, synchrotron and target station at ISIS. (Taken from the Rutherford-Appleton Laboratory Web Site http://www.rl.ac.uk).

Figure 2.24: Picture of the ISIS instrument hall. (Taken from the Rutherford-Appleton Laboratory World Wide Web Site http://www.rl.ac.uk).

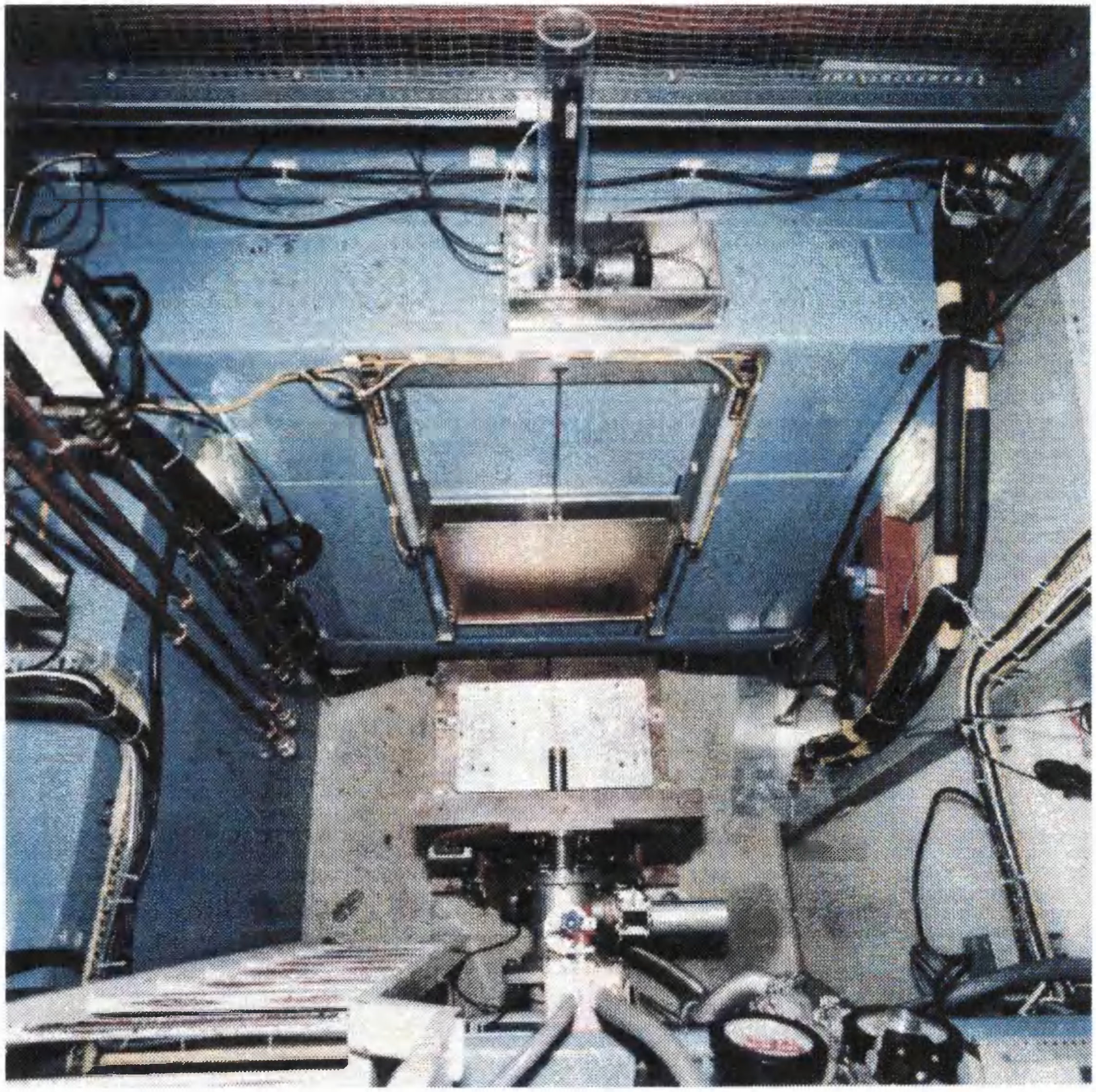


Figure 2.25: View from above the LOQ sample pit. The sample changer (Figure 2.28) is not in place. (Taken from the Rutherford-Appleton Laboratory World Wide Web Site http://www.rl.ac.uk).

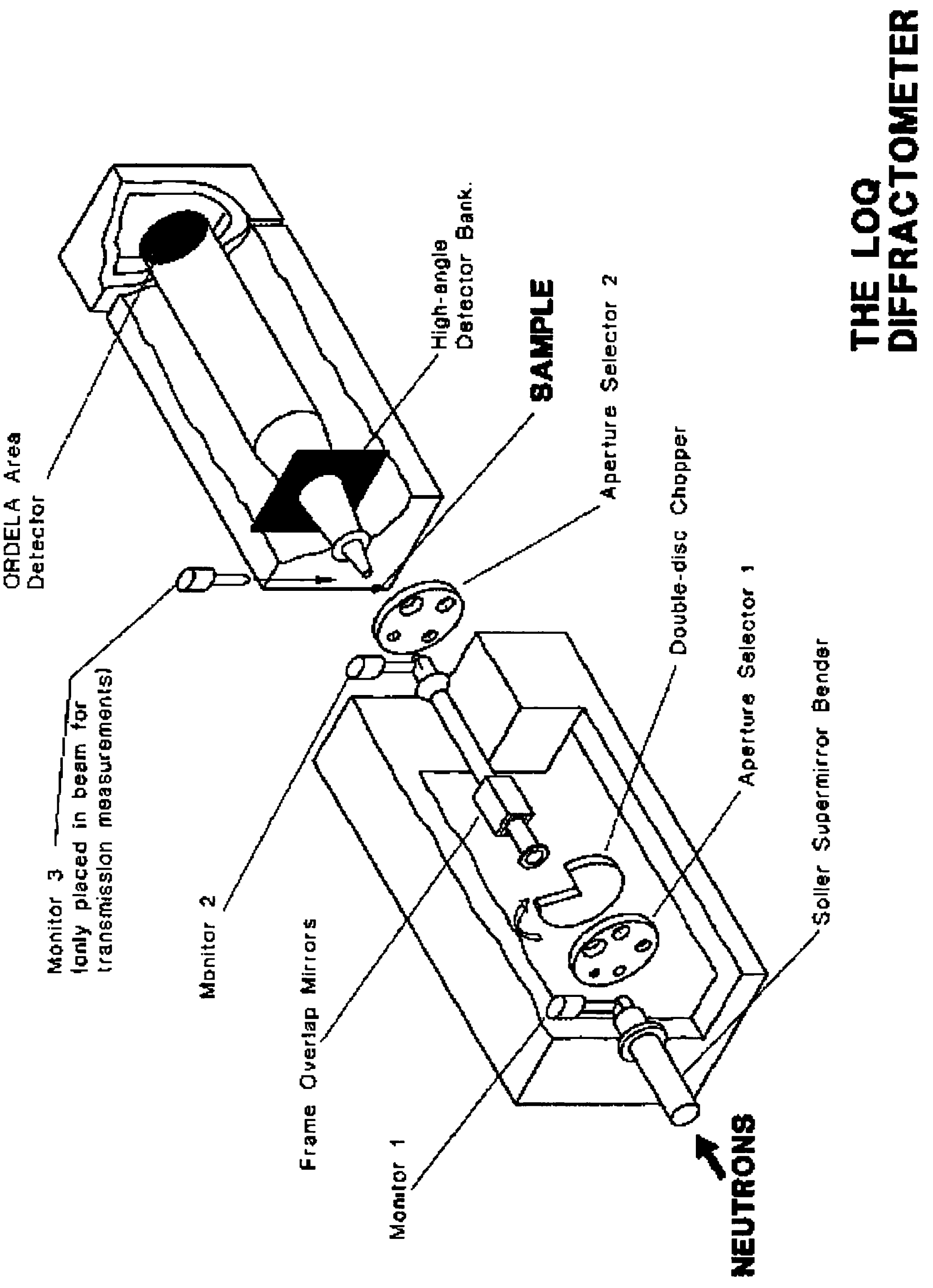


Figure 2.26: Schematic diagram of the LOQ diffractometer at the Rutherford-Appleton Laboratory, Didcot. (Taken from the Rutherford-Appleton Laboratory World Wide Web Site http://www.rl.ac.uk).

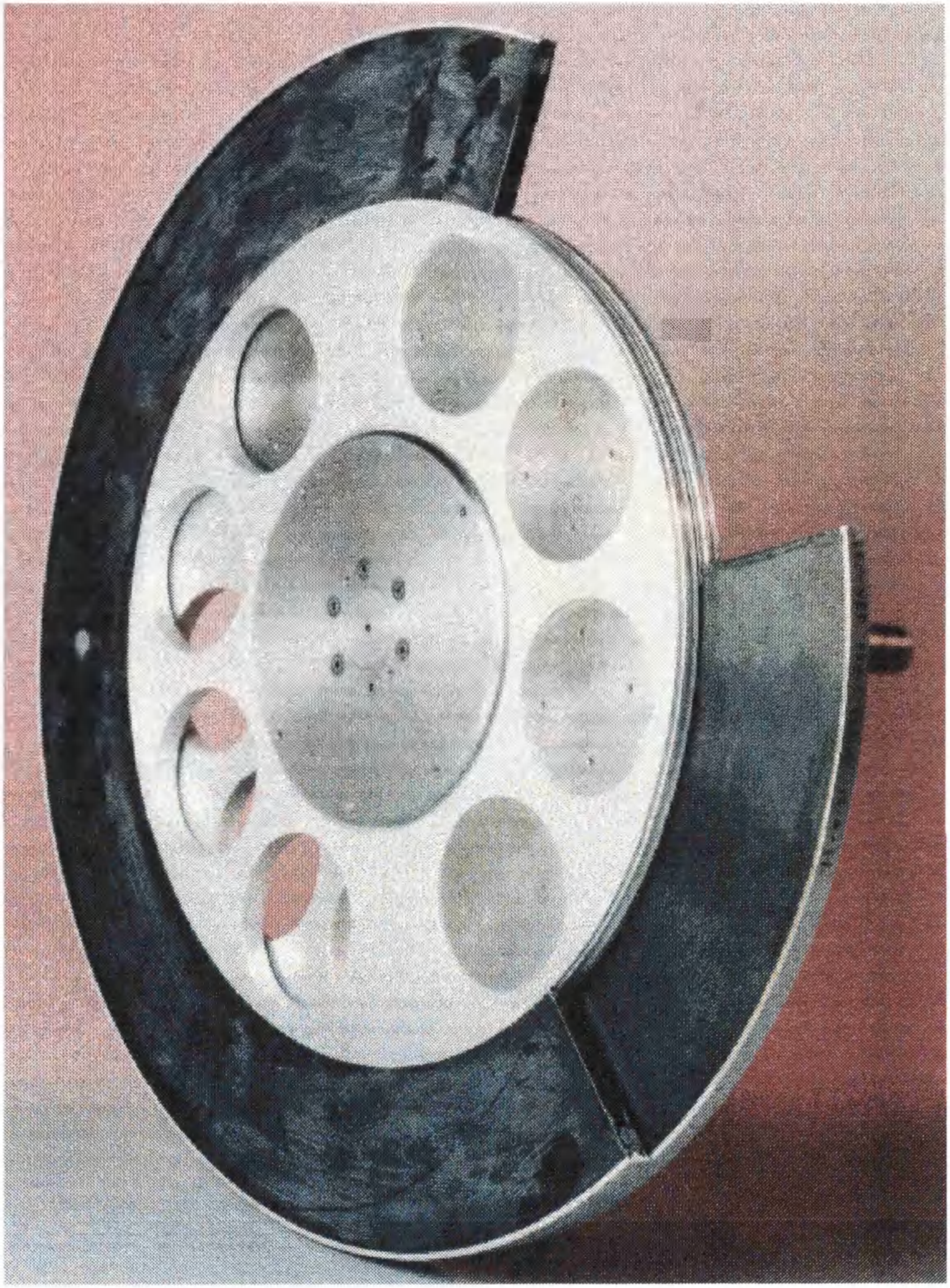


Figure 2.27: Picture of the double-disc chopper from the LOQ diffractometer. (Taken from the Rutherford-Appleton Laboratory World Wide Web Site http://www.rl.ac.uk).

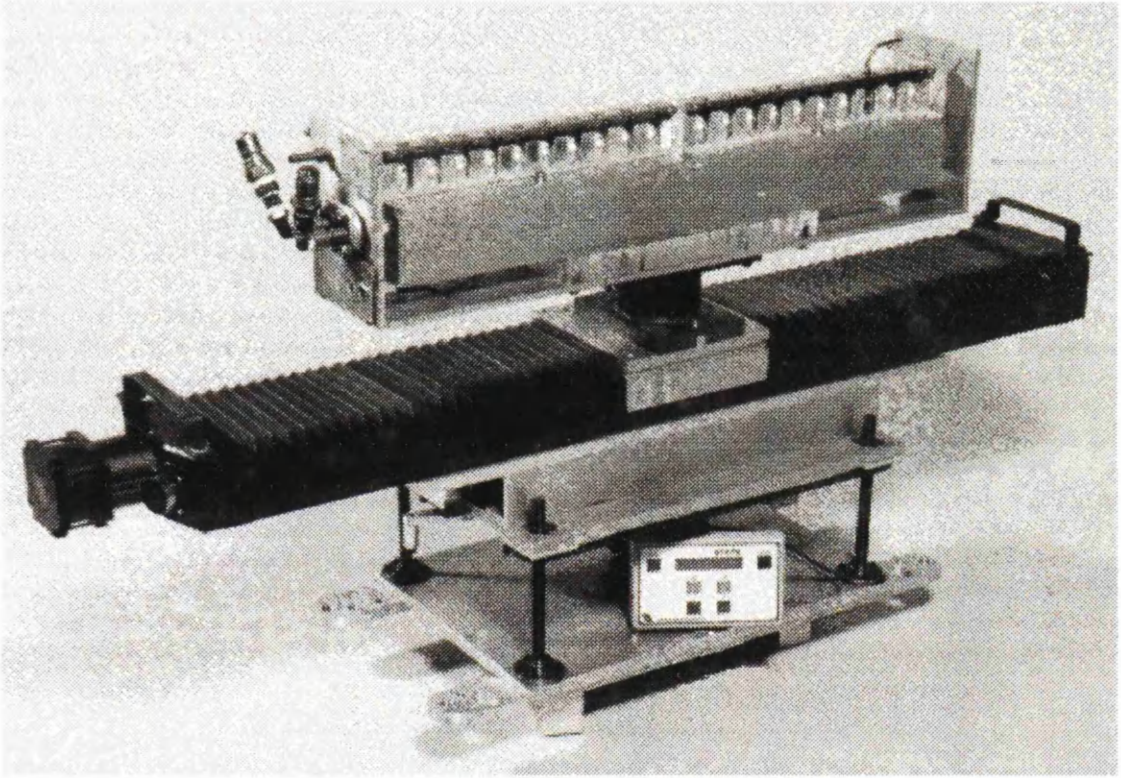


Figure 2.28: Picture of the LOQ sample changer. (Modified from a picture taken from the Rutherford-Appleton Laboratory World Wide Web Site http://www.rl.ac.uk).

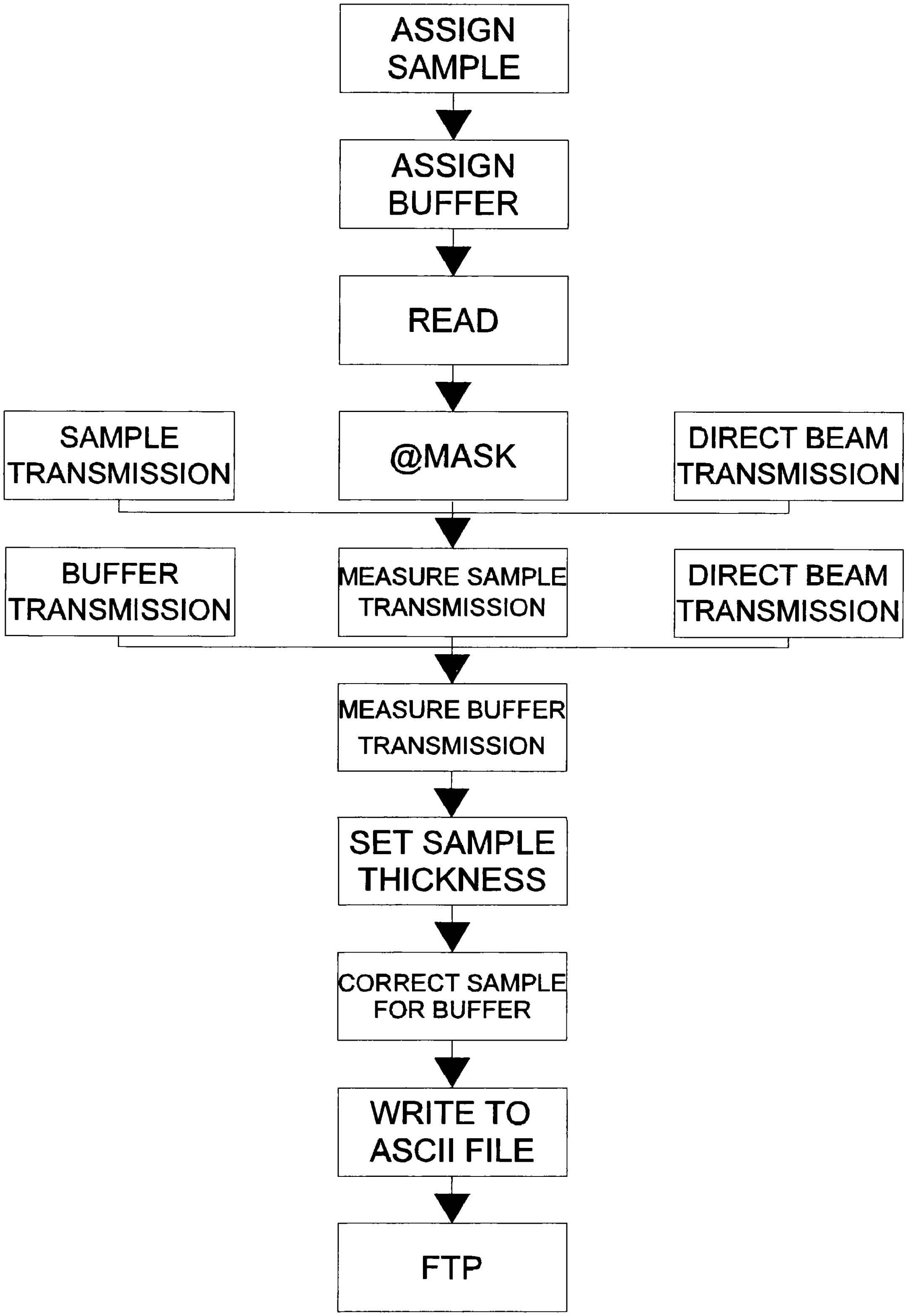


Figure 2.29: Detailed flow diagram of COLETTE. The diagram follows the stages of data reduction from the raw data files to the final transfer of ASCII files to London. @MASK executes a file (mask.com), which is updated by the instrument scientists to account for fluctuations in the behaviour of the detector and changes in instrument configuration. (Adapted from Meyer, 1994).

sample to detector distance of 4.3 m, and the usable Q range was 0.1 to 2.0 nm^{-1}. The faster neutrons of shorter wavelengths result in data at high Q, whilst slower neutrons result in data at low Q. Samples and their corresponding buffers in ^2H_2O were measured in 2 mm thick rectangular silica Hellma cells positioned in a temperature controlled rack at 15°C for 1h at a range of protein concentrations (Figure 2.28). Spectral intensities were normalised relative to the scattering from a standard calibrated partially deuterated polystyrene sample. Transmissions were measured for all the samples and buffers, together with the polymer standard and the empty beam position.

(2.2.4.6.1) Preliminary Data Reduction

Data reduction of the raw data collected in 100 time frames of 64 x 64 cells utilised the standard ISIS software package COLETTE (Heenan *et al.*, 1989) (Figure 2.29). Scattered intensities were binned into individual diffraction patterns based on wavelengths of 0.22 - 1.0 nm in linear steps of 0.02 nm or logarithmic steps of 0.08%, and corrected for the wavelength dependence of the transmission measurements. These patterns were merged to give the full scattering curve in a Q range between 0.05 - 2.2 nm^{-1}. The Q range was based on 0.04% or 0.08% logarithmic increments which was optimal both for Guinier R_G and I(0) analyses at low Q and better signal noise ratios at large Q.

(2.2.5) Analysis of Scattering Data.

(2.2.5.1) Guinier Plots.

Plots of ln I(Q) against Q^2 (Guinier plots) were performed on all X-ray and low angle neutron data using the Fortran program SCTPL5. Guinier analyses at low Q gives the radius of gyration R_G and the forward scattering I(0) (Glatter and Kratky, 1982).

$$\ln I(Q) = \ln I(0) - \frac{R_G^2 Q^2}{3}$$

For a spherical macromolecule this expression is valid in a Q range extending to a Q.R_G of 1.3. The program calculated the R_G from the gradient of the straight line fit, where the R_G is a measure of particle elongation and the internal arrangement of different scattering densities. I(0)/c (c = sample concentration) is calculated from the Y-axis intercept and is proportional to the molecular weight M_r. The M_r is calculated as an absolute value for neutron data, since it is referenced to the incoherent scattering from water as a known standard on D11 and D22, or to the forward scattering I(0) of the polymer standard on LOQ. The M_r is calculated as a relative value from X-ray data (i.e. referenced to I(0)/c measured from other samples in the same beamtime session, and normalised with the same ^{55}Fe detector response). The I(0)/c parameters may have marked concentration dependences (e.g. AmiC, Chapter 5), which can be seen by plotting I(0)/c against protein concentration.

(2.2.5.2) Cross-Sectional Plots.

If the molecule of interest is elongated, the mean radius of gyration of the cross-sectional structure R_{xs} and the mean cross-sectional intensity at zero angle $[I(Q)Q]_{Q\rightarrow 0}$ (Hjelm, 1985) can be obtained from:

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

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:

$$L=\sqrt{[12(R_G^2-R_{XS}^2)]}$$

where L is its length (Glatter and Kratky, 1982). Alternatively, L is given by (Perkins, *et al*, 1986):

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

The two semi-axes, A and B, of the elliptical cylinder are calculated by combining the dry or hydrated volume V (V= π ABL) with the R_{XS} value:

$$R_{XS}^2=\frac{(A^2+B^2)}{4}$$

The hydrated volume is obtained on the basis of a hydration of 0.3 g of water / g protein and 0.0245 nm^3 per water molecule (Perkins, 1986). Data analysis was performed using the program SCTPL5.

(2.2.5.3) Distance Distribution Function.

Indirect transformation of the scattering data in reciprocal space I(Q) into that in real space P(r) was carried out using GNOM (Svergun *et al.*, 1988; Semenyuk and Svergun, 1991; Svergun, 1992).

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

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. The calculation of P(r) from D11 and D22 necessitated joining together the scattering curves from the small angle and wide angle instrumental positions, whereas a curve from Station 2.1 or LOQ contains all the small and wide angle points within one spectrum. GNOM employs a regularisation procedure with an automatic choice of the transformation parameter α to stabilise the P(r) calculation (Svergun, 1992). 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.

CHAPTER 3

MODELLING OF BIOLOGICAL MOLECULAR STRUCTURES

(3.1) Introduction.

Sometimes it is not possible to obtain high resolution 3-dimensional structures of proteins. This could be because the protein will not crystallise or the crystals do not diffract or the protein is too large for nuclear magnetic resonance studies. In situations where a 3-dimensional structure is required the only option left is molecular modelling. Molecular modelling can be defined as being concerned with ways to mimic the behaviour of molecules and molecular systems (Leach, 1996). By combining molecular modelling techniques with physical data from spectroscopic techniques and low resolution non-crystalline diffraction studies (SAXS and SANS), it is possible to build a representative structure. There are several different techniques available for molecular modelling, namely secondary structure predictions, fold recognition methods, homology modelling and scattering curve modelling. Each is described in turn below. These are used for the modelling of the structures of AmiC (Chapter 5), Factor I (Chapter 7) and RuvA (Chapter 8).

(3.2) Secondary structure prediction methods.

A total of about 90% of residues in proteins are found in either α-helices (38%), β- strands (20%), or reverse turns (32%). The practice of predicting secondary structure from amino acid sequence on the way to predicting total protein structure is very widespread (Creighton, 1993) The methods discussed vary from statistically based approaches to neural networks, all of which have advantages and disadvantages. By combining the results of each technique the reliability of the overall prediction is increased, and it can then be seen which α-helices and β-sheets are consistently predicted by all these methods.

(3.2.1) The Chou and Fasman method.

The Chou and Fasman predictive method relies on a table of conformational preferences that were statistically derived from the occurrence of these features in 15 proteins (Chou and Fasman, 1978). Along with an extension of the analysis (Table 3.1), Chou and Fasman also composed a set of empirical rules governing the folding of secondary structural elements.

The empirical rules involved the classifying of amino acids as favouring, breaking or being indifferent to each type of conformation. The prediction method only takes into account α-helix and β-strand parameters and assumes the rest of the chain to be random coil. The conformational parameters are as follows:

$$P_\alpha = \frac{f_\alpha}{\langle f_\alpha \rangle}$$

$$P_\beta = \frac{f_\beta}{\langle f_\beta \rangle}$$

where, f_α and f_β are the frequency of residues in the α-helix and β-regions, and $<f_\alpha>$ and $<f_\beta>$ are the average frequency of residues in the α-helix and β-regions. Simple averaging of P_α and P_β values for a given residue gives the α-helix and β-strand potential for that residue.

An α-helix is said to be initiated when a cluster of four out of six adjacent helical favouring residues occur ($<P_\alpha> \geq 1.03$ and $<P_\alpha> > <P_\beta>$). As these conditions continue, any segment longer than six residues will also be predicted as helical. For the initiation

P_α			P_β			P_t		f_i		f_{i+1}		f_{i+2}		f_{i+3}	
Glu	1.51	H_α	Val	1.70	H_β	Asn	1.56	Asn	0.161	Pro	0.301	Asn	0.191	Trp	0.167
Met	1.45		Ile	1.60		Gly	1.56	Cys	0.149	Ser	0.139	Gly	0.190	Gly	0.152
Ala	1.42		Tyr	1.47		Pro	1.52	Asp	0.147	Lys	0.115	Asp	0.179	Cys	0.128
Leu	1.21		Phe	1.38	h_β	Asp	1.46	His	0.140	Asp	0.110	Ser	0.125	Tyr	0.125
Lys	1.16	h_α	Trp	1.37		Ser	1.43	Ser	0.120	Thr	0.108	Cys	0.117	Ser	0.106
Phe	1.13		Leu	1.30		Cys	1.19	Pro	0.102	Arg	0.106	Tyr	0.114	Gln	0.098
Gln	1.11		Cys	1.19		Tyr	1.14	Gly	0.102	Gln	0.098	Arg	0.099	Lys	0.095
Trp	1.08		Thr	1.19		Lys	1.01	Thr	0.086	Gly	0.085	His	0.093	Asn	0.091
Ile	1.08		Gln	1.10		Gln	0.98	Tyr	0.082	Asn	0.083	Glu	0.077	Arg	0.085
Val	1.06		Met	1.05		Thr	0.96	Trp	0.077	Met	0.082	Lys	0.072	Asp	0.081
Asp	1.01	I_α	Arg	0.93	i_β	Trp	0.96	Gln	0.074	Ala	0.076	Thr	0.065	Thr	0.079
His	1.00		Asn	0.89		Arg	0.95	Arg	0.070	Tyr	0.065	Phe	0.065	Leu	0.070
Arg	0.98	i_α	His	0.87		His	0.95	Met	0.068	Glu	0.060	Trp	0.064	Pro	0.068
Thr	0.83		Ala	0.83		Glu	0.74	Val	0.062	Cys	0.053	Gln	0.037	Phe	0.065
Ser	0.77		Ser	0.75	b_β	Ala	0.66	Leu	0.061	Val	0.048	Leu	0.036	Glu	0.064
Cys	0.70		Gly	0.75		Met	0.60	Ala	0.060	His	0.047	Ala	0.035	Ala	0.058
Tyr	0.69	b_α	Lys	0.74		Phe	0.60	Phe	0.059	Phe	0.041	Pro	0.034	Ile	0.056
Asn	0.67		Pro	0.55		Leu	0.59	Glu	0.056	Ile	0.034	Val	0.028	Met	0.055
Pro	0.57	B_α	Asp	0.54	B_β	Val	0.50	Lys	0.055	Leu	0.025	Met	0.014	His	0.054
Gly	0.57		Glu	0.37		Ile	0.47	Ile	0.043	Trp	0.013	Ile	0.013	Val	0.053

P_α, P_β and P_t are conformational parameters of helix, β-sheet and β-turns, f_i, f_{i+1}, f_{i+2}, f_{i+3} are bend frequencies in the four positions of the β-turn.
Helical assignments: H_α, strong α former; h_α, α former; I_α, weak α former; i_α, α indifferent; b_α, α breaker; B_α, strong α breaker.
β-sheet assignments: H_β, strong β former; h_β, β former; I_β, weak β former; i_β, β indifferent; b_β, β breaker; B_β, strong β breaker (Adapted from Prevelige and Fasman, 1989).

Table 3.1 Conformational Parameters for α-Helical, β-sheet, and β-turn Residues in 29 Proteins, used in the Chou and Fasman method.

of a β-strand, any three out of four or five adjacent residues needs to favour β-strand conformation. After β-strand initiation, any segment longer than three residues, where $<P_\beta> \geq 1.05$ and $<P_\beta> > <P_\alpha>$, will be predicted as β-strand. These regions of secondary structure will continue in either direction until opposite conformation is reached or sets of tetrapeptide breakers ($<P_\alpha>$ or $<P_\beta> < 1$) are encountered (Chou and Fasman, 1978). These conformational preferences P_α, P_β, and P_i are limited in that they apply to instances where a single amino acid occurrence is considered.

(3.2.2) <u>The GOR method.</u>

The GOR method (Garnier, Osguthorpe, and Robson) is a different statistical approach to structure prediction and is simpler in concept than the Chou and Fasman method as the result of the formalised use of information theory. Originally, this technique was based on 26 protein X-ray crystal structures (Garnier *et al.*, 1978) but has been updated to include a database of 75 structures (Gibrat *et al.*, 1987). The database has been defined in terms of eight different secondary structure types (Kabsch and Sander, 1983a), but this has been simplified into three secondary structures that are used within a three state predictive method. These structures are α-helix (H), β-strand or extended (E) and coil (C) (Gibrat *et al.*, 1987). The basis of the technique relies on the theory that the conformation of a particular amino acid is dependent upon the conformation of all the other residues in the sequence. Eight residues either side of this particular amino acid have been shown to exert the greatest influence on the conformational state of the amino acid (Robson and Pain, 1971; Robson and Suzuki, 1976). In the Robson method the information (I) on the conformational state (S_j), which occurs in only two ways (X or not X), at position j of residues (R_1, R_2,...., R_n) is defined by the equation:

$$I\ (S_j{=}X{:}\overline{X};R_{j-m},\ldots,R_j,\ldots,R_{j+m}) \simeq \sum_m I\ (S_j{=}X{:}\overline{X};R_{j+m})$$

This is one of two approximations and is named the directional information and specifies the information that a residue at position j + m (with m ≤ 8) brings to bear on the conformational state of the residue at position *j* (Robson and Suzuki, 1976; Garnier *et al.*, 1978). The other approximation is a combination of self information, that is the information carried by the residue itself concerning its own conformation, and pair information, which is the information carried by a residue at j+m on the conformation of the residue at *j* taking into account the type of residue at j. This is a better approximation but it is less accurate due to the smaller number of observations in the database (Gibrat *et al.*, 1987). Each conformational state (H, E and C) was evaluated for each residue within the database to give informational values for all 20 amino acids. These values are natural logs of the probabilities or probability ratios and are expressed subdivisions of a natural unit, known as centinats or cnats. These values give useful data for preferences of a particular amino acid to a particular conformation. If the values are plotted for -8 ≤ m ≤ +8 for each amino acid, four groups can be defined:

Group I

Those that have symmetrically distributed information values with a maximum at m=0 favour that conformation e.g. Ala favours α-helix (Figure 3.1a) and Val favours β-strand

Group II

Symmetrical information values that have a minimum at m=0 will disfavour that conformation at that particular residue e.g. Gly disfavours α-helix (Figure 3.1b), Leu and Val disfavour turn.

Group III

Values that are asymmetrical and are positive for $m < 0$ and negative for $m > 0$ favour the conformational state when located toward the N-terminal end of the secondary structure e.g. Asp and Glu (Figure 3.1c) in the α-helix.

Group IV

Values that are asymmetrical and are negative for $m < 0$ and positive for $m > 0$ favour the conformational state when located toward the C- terminal end of the secondary structure e.g. Lys at the C-terminus of helices (Figure 3.1d).

When the information values obtained for the 1978 and the 1987 databases were compared, they were similar. Differences were noted for the α-helical propensities for Arg, Cys and Trp; the β-sheet propensities for Gln and Pro; the β-turn values for Cys, Gln, Ile and Trp; and the random coil propensities for Trp (Garnier and Robson, 1989).

These predictions may be improved by a number of measures. Decision constants, which can be subtracted from either α-helix or β-sheet values, are used to favour a particular conformation if experimental information is available (e.g. from either circular dichroism or Fourier transform infrared spectroscopy) (Garnier *et al.*, 1978).

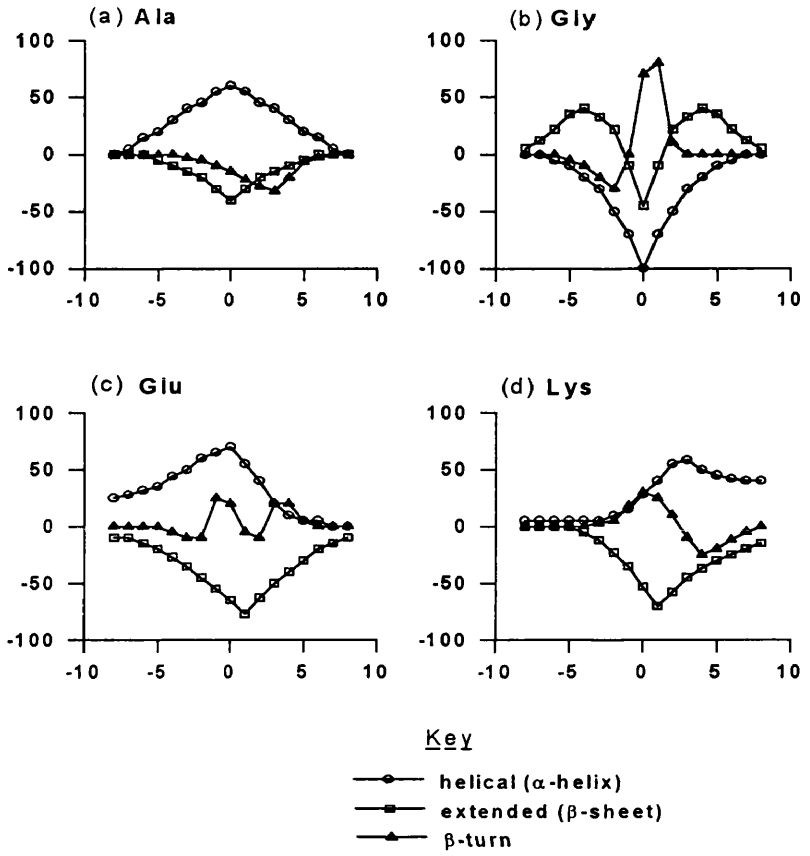


Figure 3.1: Directional informational values used in the GOR method:

$$I(S_j=X:\overline{X};R_{j+m})$$

(Y-axis) plotted against position relative to *j* (X-axis) for the residues (a) Ala, (b) Gly, (c) Glu and (d) Lys. From these plots, it can be seen that Ala favours an helical conformation at *j* but Gly disfavours this conformation. Glu favours the helical conformation at the N-terminus of the helix whereas Lys favours the C-terminus (Reproduced from Brissett, 1997).

Studies have shown the Robson and the Chou-Fasman prediction methods work better for proteins containing a single type of secondary structure, i.e. all α-helix or all β-sheet, than for proteins of a mixed type. In one example, the distribution of hydrophobic residues was used in association with these programs to improve the predictions (Busetta and Hospital, 1982). Subsequent secondary structure prediction algorithms have made use of this technique.

(3.2.3) The PHD method.

The profile network system from EMBL-Heidelberg (PHD) is a secondary structure prediction algorithm based on the output from neural networks. PHD uses the profiles from multiple sequence alignments as an input into a three-layered network (Figure 3.2). The resulting secondary structure predictions had an accuracy >70% when cross-validated on >100 unique proteins (Rost and Sander, 1993b; 1994). Based on the statement that homologous proteins have the same three-dimensional fold and approximately equivalent secondary structure profiles at around a level of 25-30% identical residues, a multiple sequence alignment of the protein family can contain more structural information than a single sequence (Rost and Sander, 1993a).

The first layer (sequence-to-structure net) is concerned with classifying strings of adjacent residues into three states of secondary structure, helix (α), strand (β), and loop (L). Each position at the centre of a window (w) of 13 consecutive residues in the multiple sequence alignment is analysed for the frequency of occurrence of each of the twenty amino acids (profile generation, Figure 3.2). This frequency is used as an input into a network trained to classify mutually independent segments of residues in terms of

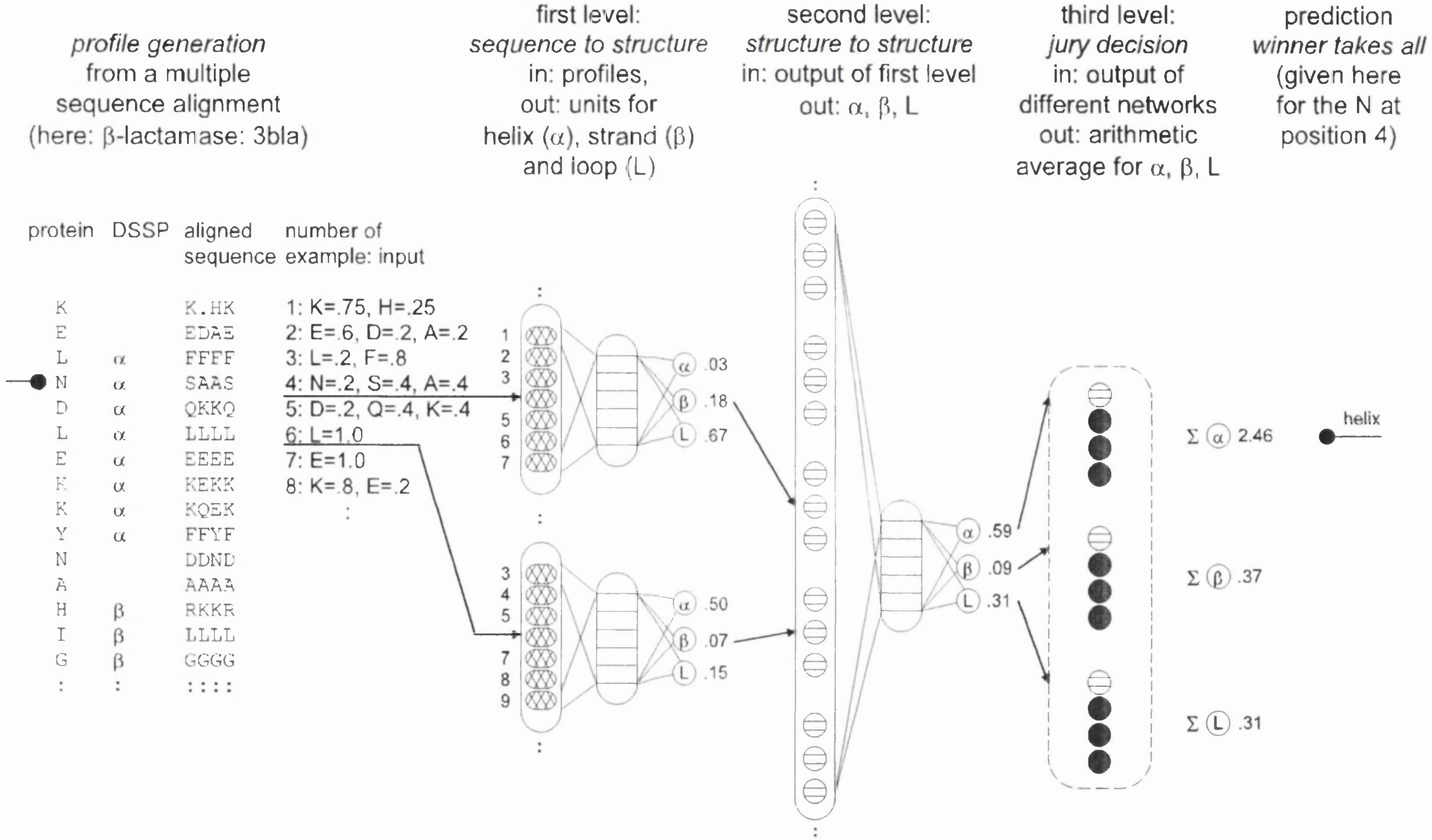


Figure 3.2: A representation of the PHD network for secondary structure prediction. The network consists of 3 layers: 2 network layers and 1 layer averaging over independently trained networks. ⊗, Basic cell containing 20+1 units to code residues at position 1 to *w* of the input window; here w=7. ⊖, Hidden units. Circled α, β and L, output units for helix strand and loop. ●, Output from architectures not shown here. –●, Example: residue N at position 4 predicted to be in helix●–. (Reproduced from Brissett 1997).

the state of a single central residue. The output from this level gives the probability of the residue occurring in one of the three states. At this stage the prediction of the three states is for each individual position and is not related to the adjacent residues (sequence to structure, Figure 3.2).

The next layer is concerned with the interpretation of these individual positions in the context of the surrounding predicted states. This structure to structure network (Figure 3.2) uses a window length (w) of 17 that predicts the secondary structure for the central residue from stretches of predictions. Here, an unlikely prediction from the first level such as HHHEEHH (H, helix; E, strand) will be altered to HHHHHHH. Even though the overall prediction accuracy is not significantly improved the length of predicted segments is more consistent to observed protein structures than the output from level one.

The third level or jury decision (Figure 3.2) is effectively a noise reduction step that comes about by the arithmetic averaging of 12 different network predictions. These networks have been trained differently on 'balanced' and 'unbalanced' datasets. In 'balanced' training the number of examples of α-helix, β-strand and loop presented in the training set are equal, as opposed to 31% α-helix, 22% β-strand and 47% loop found in the database or 'unbalanced' training set (Rost and Sander, 1993b).

(3.2.4) <u>The SAPIENS method.</u>

SAPIENS (Secondary structure and Accessibility class Prediction Including ENvironment-dependent Substitution tables) is a prediction method based on the evaluation of mean propensities and environment-dependent substitution tables for amino acid residues in aligned sequences. The program uses this data in a three step method with a multiple sequence alignment as its primary input. Unlike PHD, SAPIENS evaluates four conformational states (α-helix [H], β-strand [E], buried coil [i], and exposed coil [o]).

The principle in step one is to evaluate the preference of each residue in the protein of interest for one of the four conformational states. This is done by analysing the amino acid substitution patterns and mean propensities at equivalent sites in aligned homologous proteins. Step one (stp1) consists of three substeps (a1-3, b1-3, Figure 3.3). Initially, each residue is assigned to either one of H, E, i, and o by assessing the preference of that residue for the given conformational state derived from propensity and substitution tables (Wako and Blundell, 1994a). These assignments are re-evaluated for neighbouring residue co-operativity (giving H, h, E, e, i, and o) (a1, b1 Figure 3.3). For this purpose, if the preference of a residue (for H or E) is not the greatest at that position, but greater than any other preference multiplied by a scaling factor (ϕ), h or e is assigned. As assignment to both α-helix and β-strand is carried out independently, residues can be assigned to both h and e.

Based on these assignments secondary structure segments are predicted (a2, b2 Figure 3.3). Here, segments of residue length ω_h, exclusive of gaps are analysed for α-helix assignments (H or h). These H and h assignments are scored 1 and 0.5 respectively. This score is summed over the residues in the segment and if the score is above a certain cutoff value ρ_h, the whole segment is predicted as α-helix. If N- and C-terminal residues of the segment are not already assigned to H or h in a1 or b1, they are not assigned H. Using this rule the case of an α-helix 3 residues long can occur, e.g. cHHHc (c=i or o). β-strand prediction is carried out the same way using E, e, ω_e, and ρ_e.

In Figure 3.3, the Hs in bold type are used to illustrate this procedure. For ω_h=5 and ρ_h=3 the five amino acids SKVKA have a point score of 2.5, but KVKAL has a point score of 3 and all these residues are assigned H as the window moves along. Residues VKALJ only score 2.5 but Hs have already been assigned to VKAL. As the window moves along the point score falls below the ρ_h threshold until the window reaches residues ALJSE and LJSEL which are assigned H. After this, ρ_h<3 defines the C-terminal edge of the segment.

The next substep involves the assignment of one of the four states to the N or C-terminal residues of the secondary structure segment (a3, b3 Figure 3.3). This substep is based on a search criteria that starts at the N or C-terminal and searches for residues within the secondary structure segment that satisfy the capping propensity conditions. If no residues that satisfy this criteria are found in the segment, or if the new N or C-terminal residues are reversed (i.e. the residue number for the N-terminal residue is greater than that of the C-terminal residue), the secondary structure assignments in the segments are cancelled out and i and o are reassigned.

The final substep concerns residues that have been assigned to both α-helix or β-strand. Initially, the defined states of neighbouring residues are used to decide the state of the residues in question. If this still results in dual assignments then the mean preferences for both states are calculated over the segment, and the most preferable state is assigned to the segment. This final assignment is reported in stp1 (Wako and Blundell, 1994b).

Step two (stp 2, Figure 3.3) involves using the arrangement of solvent accessibility classes along the sequence (accs, Figure 3.3). The assignments are adjusted based on the result. The β-strand (f, Figure 3.3) assignment is just based on a simple template method, where the solvent accessibility profile of segments predicted as coil are searched for one of two patterns. The first pattern is an alternating series of i and o spanning more than six residues, and the second pattern is consecutive i for 5 or more residues. These patterns are classically seen for surface based β-strands and buried β-strands respectively. The solvent accessibility classes are worked out for each method by the method of Wako and Blundell (1994a). For the α-helix assignment (g, Figure 3.3) a Fourier transform method that detects periodicity in solvent accessibilities is utilised (Wako and Blundell, 1994b). H assignments are changed without respect to their previous states after this algorithm is applied.

Step three (stp 3, Figure 3.3) takes the average conformational state assigned across the alignment at that residue position and is the final secondary structure prediction from SAPIENS. The most dominant state is the one that is reassigned to all residues at this position. If no state dominates then the original assignments stay the same (Wako and Blundell, 1994b).

```
 protein # 1:
      1                                                                    70
sqnc) RGLTLRGSQRRTJQEGGSWSGTEPSJQDSFMYDTPQEVAEAFLSSLTETIEGVDAEDGHGPGEQQKRKIV
  a1)   h hh  hHHh HH  h    H h Hh  H    HH HHHhh  h Hh hhh hHh hh  hhH H  h
  a2)         HHHHHHH                    HHHHHHHH                   HHHHH
  a3)                                   HHHHHHHHH                 HHHHHHHHH
  b1)   eEee  eeee     eee E  E e eEeE e   E   Ee  EE eE eE           e e EE
  b2)                             EEEEEE           EEEEEEEE             EEEE
  b3)                             EEEE                                   EEE
stp1) ooiiioioooooioooooioiooooiooEEEEoiHHHHHHHHHooiooiioiooioooooHHHHHHHHEE
stp2) ooiiioioooooioooooioiooooiooEEEEoiHHHHHHHHHooiooiioiooioooooHHHHHHHHEE
   f)
   g)                                   HHHHHHHHh
accs) ooioioioooioioooooioioiooioooiiiiooooiioiioooioooioiooooooooooooioioio
stp3) ooiiioioooHHHHHoooioiooooioooEEEoiHHHHHHHHHooiooiioiooioooooHHHHHooEEE
sqnc) RGLTLRGSQRRTJQEGGSWSGTEPSJQDSFMYDTPQEVAEAFLSSLTETIEGVDAEDGHGPGEQQKRKIV
     71                                                                    99
sqnc) LDPSGSMNIYLVLDGSDSIGASNFTGAKK
  a1)  h   hh   h h h h   h    hHhh
  a2)
  a3)
  b1) E     e EEEEE     E    E Ee
  b2) E     EEEEEEE
  b3) EE      EEEEEE
stp1) EEooooioEEEEEEioooiooooiooioo
stp2) EEooooioEEEEEEioooiooooiooioo
   f)
   g)
accs) ioooooioioiiiiioooiooooioiioo
stp3) EEooooiEEEEEEEioooiooooiooioo
sqnc) LDPSGSMNIYLVLDGSDSIGASNFTGAKK
```

Figure 3.3: An example of the output from SAPIENS as an illustration of the prediction method. The protein sequence, sqnc, is from the link region between Ba and Bb of human complement factor B.

a1-3 and b1-3 are the α-helix and β-strand assignment substeps in step1, stp1.
f, is the β-strand assignment based on the accessibility arrangement in accs.
g, is the α-helix assignment based on a Fourier transform analysis of the accessibility arrangementin accs.
stp2, is the secondary structure assignment based on the substeps f and g.
stp3, is the final secondary structure assignment taken by the averaging of the conformational states across equivalent positions in the alignment.

(3.3) Fold Recognition.

In distinction to secondary structure prediction methods, fold recognition techniques (or inverted protein structure prediction) attempt to match one-dimensional information contained in sequences directly to three-dimensional folds (Bowie and Eisenberg, 1993). These methods are based on the environment of the residue as this tends to be more highly conserved than the identity of the residue itself (Jones *et al.*, 1992). Because of this, matching of a test sequence to a protein fold results in the detection of more distant sequence-to-structure relationships. This is an advantage over sequence-based methods which can fail to recognise highly similar protein structures that have a sequence similarity of <25%.

THREADER is a fold recognition program that uses a dynamic programming algorithm based on a combination of neighbour and solvation preferences (Jones *et al.*, 1992). A library of 254 unique protein chains of crystallographic resolution ≤2.8Å, was constructed. Each fold is ignored at the sequence level and considered just as a chain tracing through space. Optimal fitting ('threading') of the test sequence to the backbone coordinates of the fold then follows, and the pseudo-energy of each fitting is evaluated by the summing of pairwise interactions. The fitting process is carried out via a dynamic programming-based algorithm that is capable of optimising pairwise interaction potentials between amino acid residues. Evaluation of the pseudo-energy of a sequence in a particular conformation is not handled by classical energy potentials, but by 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.

Ranking of the library folds in order of ascending total energy is carried out with the lowest pairwise interaction energy attributed to the most probable match (Jones *et al.*, 1992). The interaction energies in THREADER are expressed as Z-scores which are defined as :

$$Z\text{-}score = \frac{Energy - mean}{standard\ deviation}$$

for the pairwise or solvation energies. The Z-score that is used for overall ranking of the folds is the pairwise interaction energy Z-score based on the calculated residue potentials for the sets of proteins with a reasonable proportion of the sequence and structure matched. For example, a Z-score <-3.5 is regarded as very significant and is probably a correct prediction.

(3.4) Homology Modelling.

In order to visualise the atomic coordinates of crystal structures as well as predicted models, the technique of molecular modelling is used. Molecular modelling involves the visual manipulation of protein structures, and various suites of programs are available that run on a variety of platforms to achieve this effect. INSIGHT II 95.0 (Biosym/MSI, San Diego, USA), is a suite of molecular graphics and computational chemistry programs that can be manipulated interactively via a mouse driven interface or from the command line in a non-interactive manner. Within INSIGHT II are a number of programs (or modules) that allow the user to build, manipulate, and simulate virtually any class of molecule whilst also studying its molecular properties.

HOMOLOGY is one program within INSIGHT II that allows the prediction of protein structure to be based on existing conformations of reference proteins. As seen in Figure 3.4, suitable reference structures are determined for the model sequence. This can be done by a variety of methods such as sequence database search/alignment methods or by fold analysis methods. The next step is to define regions of structural conservation. These structurally conserved regions are more reliable in determining protein folding than the use of sequence alignments alone, as similar sequences do not always have the same conformations. This step is only useful if a family of homologous sequences has been defined and more than one reference structure is identified. After defining the structurally conserved regions the model sequence is aligned to these regions and the coordinates are assigned to the sequence.

The areas between these conserved regions need to have coordinates assigned and a variety of methods can be used to generate these conformations. This part of the model building procedure is a fairly time consuming process as these areas tend to be loop regions that tend to be divergent in sequence. HOMOLOGY enables either a *de novo* segment to be generated or a probable loop to be identified from the crystallographic database using its own search algorithm. Search Loops is a routine that searches the pdb_select.1995-jun-01 database of fragments from 349 crystal structures at 0.2 nm resolution or better (Hobohm *et al.*, 1992; Hobohm and Sander 1994) for the regions that meet a certain geometric criteria. A C^{α} distance matrix is used to determine regions of proteins that have C^{α} distances that best fit the region of the protein under study. An added constraint in the search is that the selected region should have the same number of residues as the region under study.

HOMOLOGY uses a best fit equation to define the lowest root mean-squared distance value as seen in:

$$\left(\sum_{i=1}^{N} \frac{(x+x_0)^2+(y+y_0)^2+(z+z_0)^2}{N}\right)^{\frac{1}{2}}$$

The ten best segments are retained that are the segments with the lowest RMS value. The matrix is based not on the area to be constructed (flex region) but upon the areas pre- and post- to the flex region (Figure 3.5). The geometry of the selected segment is allowed to vary and is not a search criteria. The number of distances compared is defined by $(N^2-N)/2$ where N is the total number of pre- and post-flex residues. The next step is to search for optimum side chain conformations of residues that differ in the model protein from those in the reference protein.

(3.5) Energy Refinements.

The final step in Figure 3.4 is to subject the structure to energy minimisation in order to relax any strain that has been introduced during the modelling process. DISCOVER is a molecular simulation program within INSIGHT II that will achieve this. Routines that include energy minimisation, template forcing, torsion forcing and dynamic

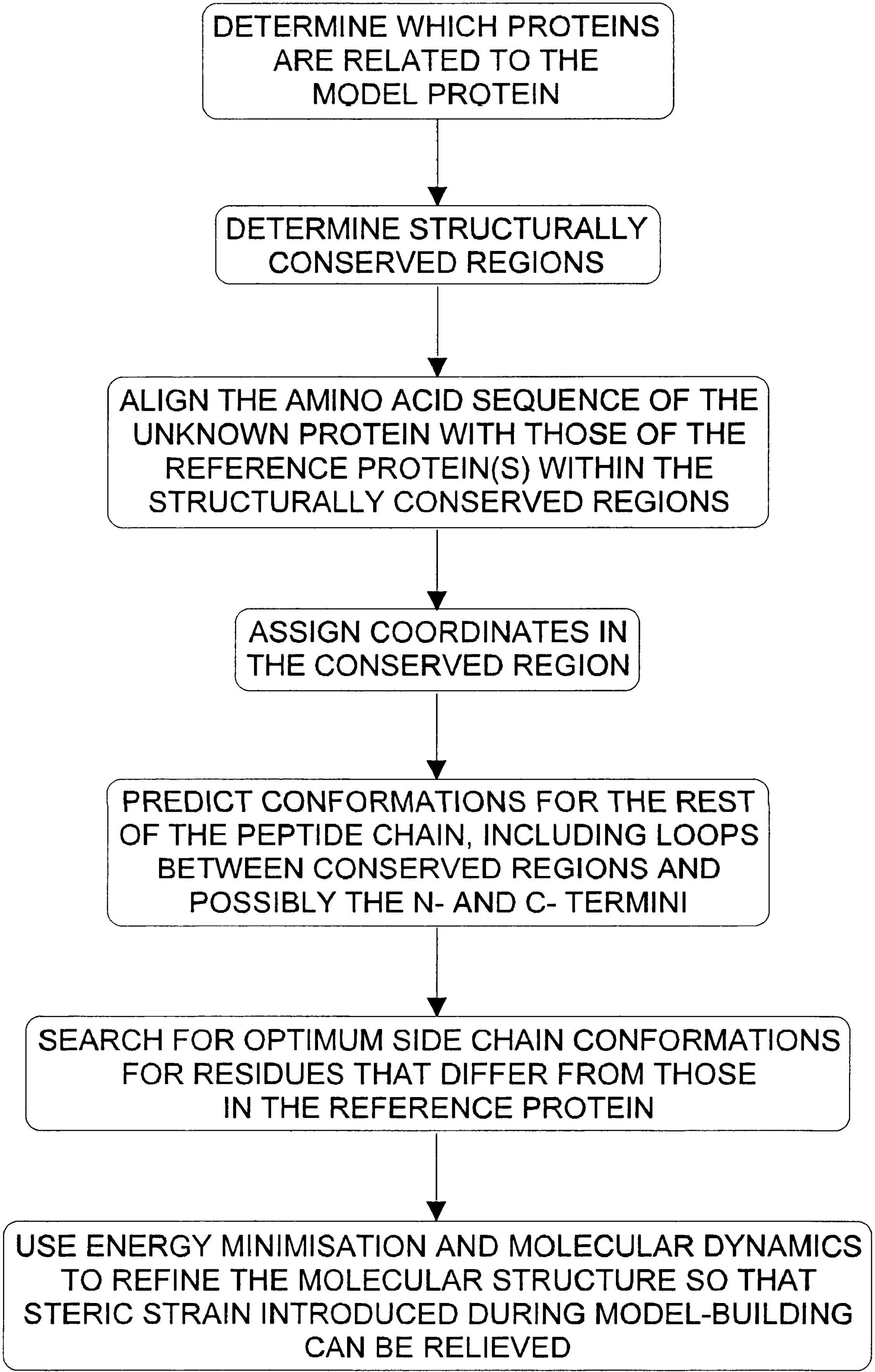

DETERMINE WHICH PROTEINS ARE RELATED TO THE MODEL PROTEIN
DETERMINE STRUCTURALLY CONSERVED REGIONS
ALIGN THE AMINO ACID SEQUENCE OF THE UNKNOWN PROTEIN WITH THOSE OF THE REFERENCE PROTEIN(S) WITHIN THE STRUCTURALLY CONSERVED REGIONS
ASSIGN COORDINATES IN THE CONSERVED REGION
PREDICT CONFORMATIONS FOR THE REST OF THE PEPTIDE CHAIN, INCLUDING LOOPS BETWEEN CONSERVED REGIONS AND POSSIBLY THE N- AND C- TERMINI
SEARCH FOR OPTIMUM SIDE CHAIN CONFORMATIONS FOR RESIDUES THAT DIFFER FROM THOSE IN THE REFERENCE PROTEIN
USE ENERGY MINIMISATION AND MOLECULAR DYNAMICS TO REFINE THE MOLECULAR STRUCTURE SO THAT STERIC STRAIN INTRODUCED DURING MODEL-BUILDING CAN BE RELIEVED

Figure 3.4: Flow diagram highlighting the main processes involved in homology model building. (Reproduced from Brissett, 1997).

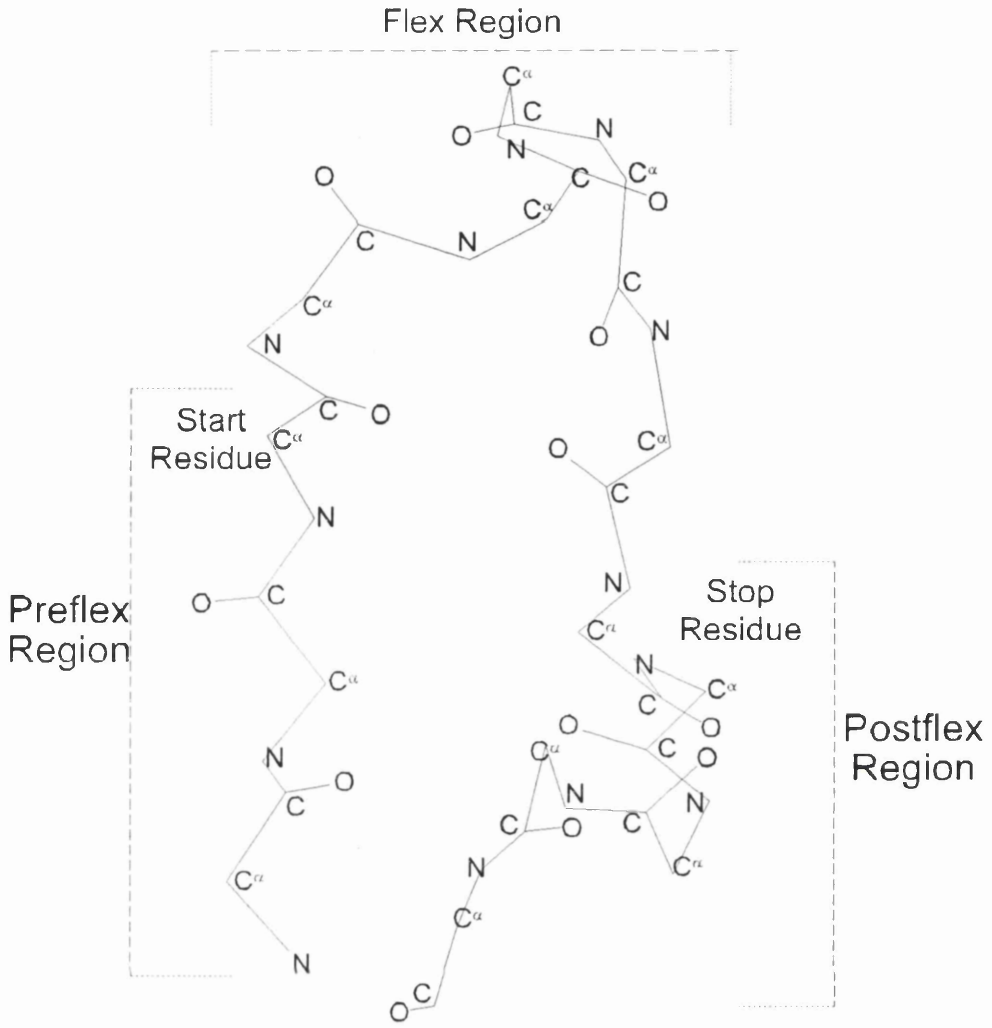


Figure 3.5: Diagram illustrates the regions involved in the loop search feature used in HOMOLOGY. The preflex region consists of residues preceeding the gap. The flex region is the number of residues that must be contained in the loop to be searched. Post flex residues are the residues after the gap. The distance between the last C^α of the preflex and the first C^α of the postflex regions is used as the search criteria. The top ten loops are fitted to the model structure by the best RMS fit to all the atoms in the two residues already specified. The loops can also be fitted be taking into consideration the RMS fit between the pre- and postflex C^αs of the model protein and selected loop. (Reproduced from Brissett, 1997).

trajectories as well as the calculation of interaction energies, derivatives, mean square displacements, and vibrational frequencies are performed by this program. Flexible control is allowed so that these calculations can be carried out at different user-defined levels of stereochemical restraint on sets of atoms, depending on the conservation of secondary structure, or if the residue has been mutated. For example the user may only wish to refine the positions of searched loops while leaving the conserved residues alone.

Two first-order minimisation algorithims that are frequently used in molecular modelling are the steepest descent and conjugate gradient methods. These gradually change the coordinates of the atoms as they move the system closer and closer to the minimum point (Leach, 1996).

The steepest descent method moves in the direction parallel to the net force. The conjugate gradient method produces a set of directions which does not show the oscillatory behaviour of the steepest descent method. In the steepest descent method both the gradients and direction of successive steps are orthogonal. In conjugate gradients, the gradients at each point are orthogonal but the directions are conjugate (this method is more properly called the conjugate directions method) (Leach, 1996).

(3.6) Modelling of Solution Scattering Data.

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, as a function of pH, temperature or other variable of interest (Perkins *et al.*, 1998).

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 (Chapter 2). 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 (Chapter 5). 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 multisubunit 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 (Perkins *et al.*, 1998).

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 (Perkins *et al.*, 1998).

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) (Perkins *et al.*, 1991). 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 tedious 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 (Perkins *et al.*, 1998).

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 (Glatter and Kratky, 1982; Perkins and Weiss, 1983).

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

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

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. The method has been tested with crystal structures for β-trypsin and α_1-antitrypsin (Smith *et al.*, 1991; Perkins *et al.*, 1993), and more recently with one for pentameric serum amyloid

P component (Ashton *et al.*, 1997). 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 by carcinoembryonic antigen (Boehm *et al.*,1986; Perkins and Weiss, 1983).

The stages of the modelling procedure are summarised in Figure 3.6. Initial trial models were generated using INSIGHT II using the atomic structures for individual domains in order to determine how best to set up an automated procedure. Full coordinate models were used. If carbohydrate was present, the oligosaccharide chains were represented by a suitable structure adapted from the Brookhaven database (Boehm *et al.*, 1996) 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 (Figures 3.7a, 3.7b, 3.7c; 3.8a, 3.8b, 3.8c). For the analyses of oligomers, symmetry constraints were used to define the location of the monomeric subunits (Figures 3.7d, 3.7e; 3.8d, 3.8e).

The atomic coordinates of each glycoprotein model were converted to spheres (Table 3.2). 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

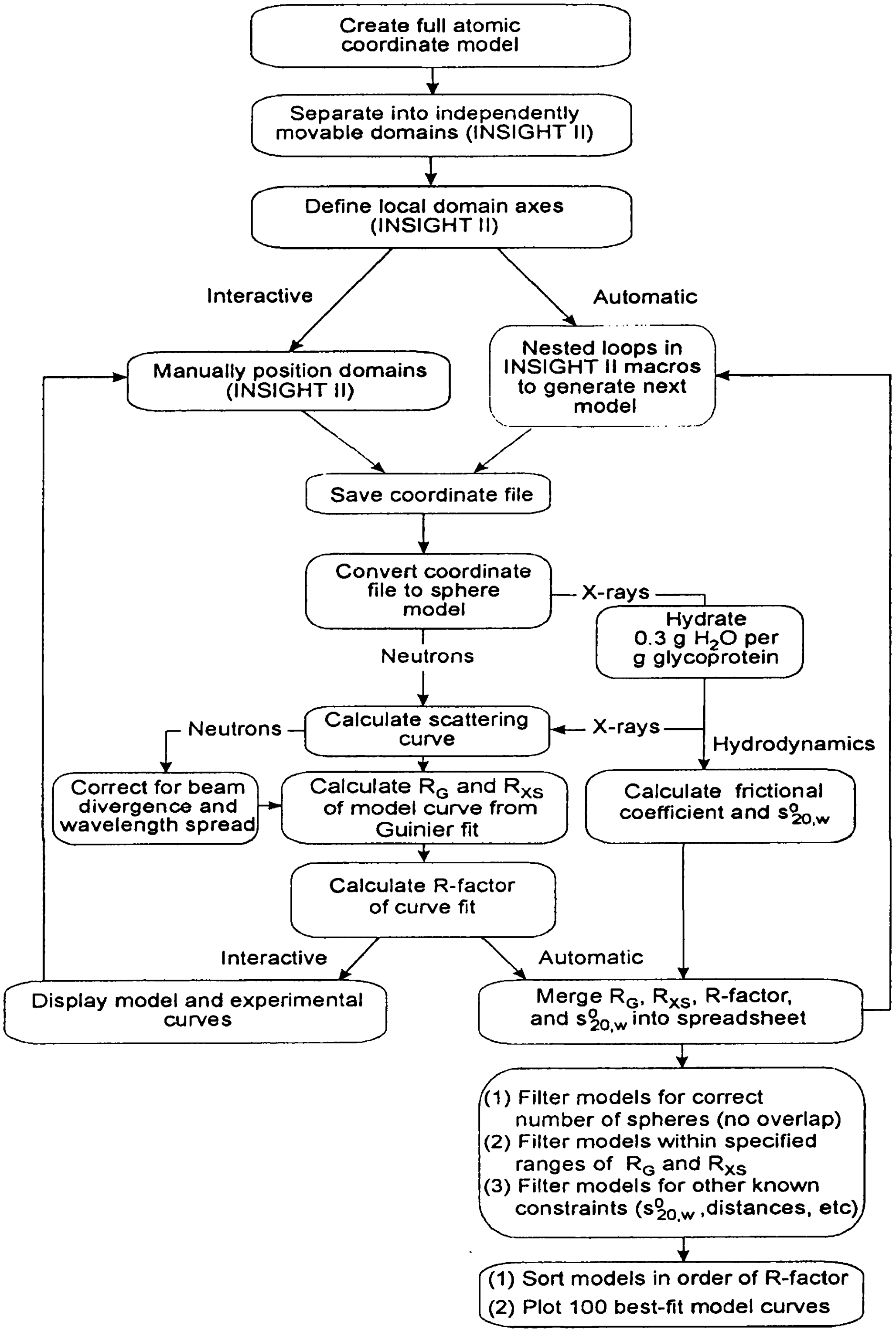

Create full atomic coordinate model
Separate into independently movable domains (INSIGHT II)
Define local domain axes (INSIGHT II)
Interactive
Automatic
Manually position domains (INSIGHT II)
Nested loops in INSIGHT II macros to generate next model
Save coordinate file
Convert coordinate file to sphere model
X-rays
Hydrate 0.3 g H_2O per g glycoprotein
Neutrons
Neutrons
Calculate scattering curve
X-rays
Hydrodynamics
Correct for beam divergence and wavelength spread
Calculate R_G and R_{XS} of model curve from Guinier fit
Calculate frictional coefficient and $s^o_{20,w}$
Calculate R-factor of curve fit
Interactive
Automatic
Display model and experimental curves
Merge R_G, R_{XS}, R-factor, and $s^o_{20,w}$ into spreadsheet
(1) Filter models for correct number of spheres (no overlap)
(2) Filter models within specified ranges of R_G and R_{XS}
(3) Filter models for other known constraints ($s^o_{20,w}$,distances, etc)
(1) Sort models in order of R-factor
(2) Plot 100 best-fit model curves

Figure 3.6: 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. (Reproduced from Perkins *et al.*, 1998).

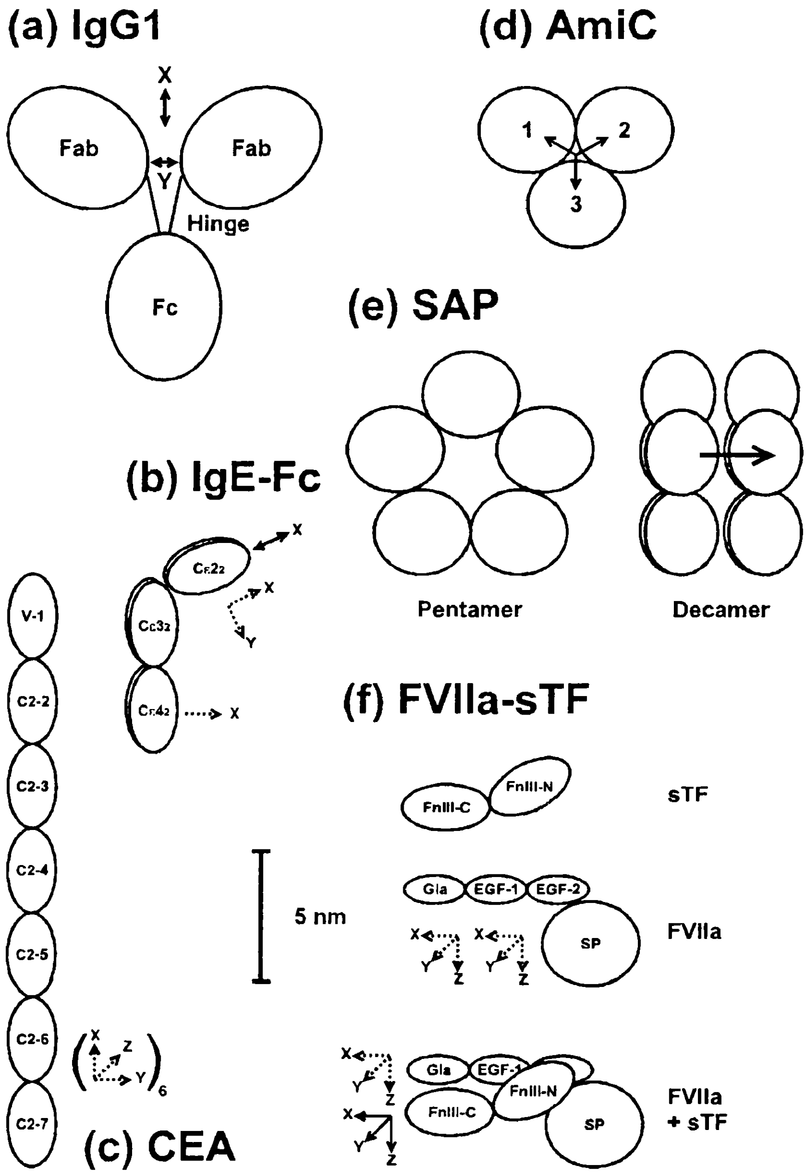


Figure 3.7: Schematic outlines of six multidomain or oligomer structures to show how domain or subunit translations and rotations were implemented during the curve fit analyses. Translations are denoted by solid arrows, and rotations by dashed arrows.

(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 $C\epsilon 2_2$ domain pair were translated along the X-axis twice, and rotated about the X- and Y-axes relative to the Cε3 and Cε4 domains. A further X-axis rotation involving the $C\epsilon 4_2$ 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 three-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. (Figure reproduced from Perkins *et al.*, 1998).

Five proteins (a) to (e)	Molecular weight	Spheres	Cube side (nm)	Search parameters	Number of models	Instrument[a] R_G (nm)	Observed R_G (nm)[b]	Fitted	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 × 9,360	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 × 4,056	St 8.2 (X-rays)	8.0 ± 0.6	8.0	0.12-2.0	4.7
					2 × 4,056	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 × 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 × 9,261	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

Table 3.2: Scattering curve fit analyses for six multidomain proteins. Table taken from Perkins *et al.*, 1998.

[a] Neutron data correspond to 100% ^2H_2O buffers.

[b] The fitted R_G values correspond to the final model depicted in Figure 3.8.

[c] Two-density models with 485 protein and 474 carbohydrate spheres were used for the final fit.

[d] Asymmetric trimers from Chamberlain *et al.*, 1997.

volume calculated from the sequence (Perkins, 1986). 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.

The dry models do not have a hydration shell and are used for neutron curve modelling as neutron scattering observes unhydrated glycoprotein structures (Ashton *et al.*, 1997; Smith *et al.*, 1990; Perkins *et al.*,1993). 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 (Perkins, 1986), 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 (Ashton *et al.*, 1997) 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 (Perkins *et al.*, 1998).

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 around 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 (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 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 (Ashton *et al.*, 1997).

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 (Figure 3.6) 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: (i) 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. (ii) Next, models were retained if the modelled R_G and R_{XS} values were within 5% or ± 0.3 nm from the experimental values. (iii) 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 (Beavil *et al.*, 1995; Smith *et al.*, 1990). 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 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 arbitraily set as 1000. The

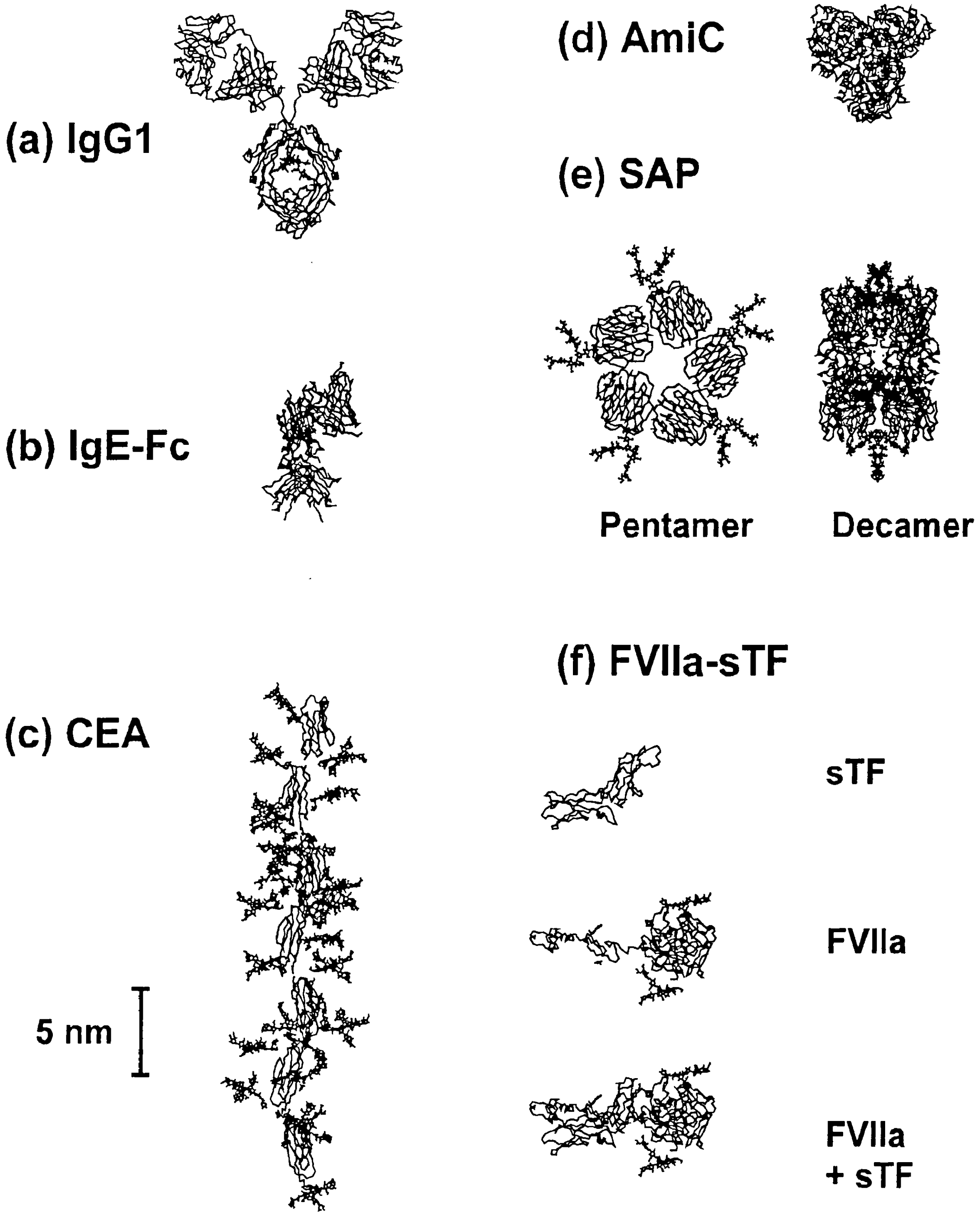


Figure 3.8: The best-fit model from each curve fitting analysis to follow that of Figure 3.7. The protein structure is denoted by an α-carbon trace, while oligosaccharides are shown in full if present. (Reproduced from Perkins *et al.*, 1998).

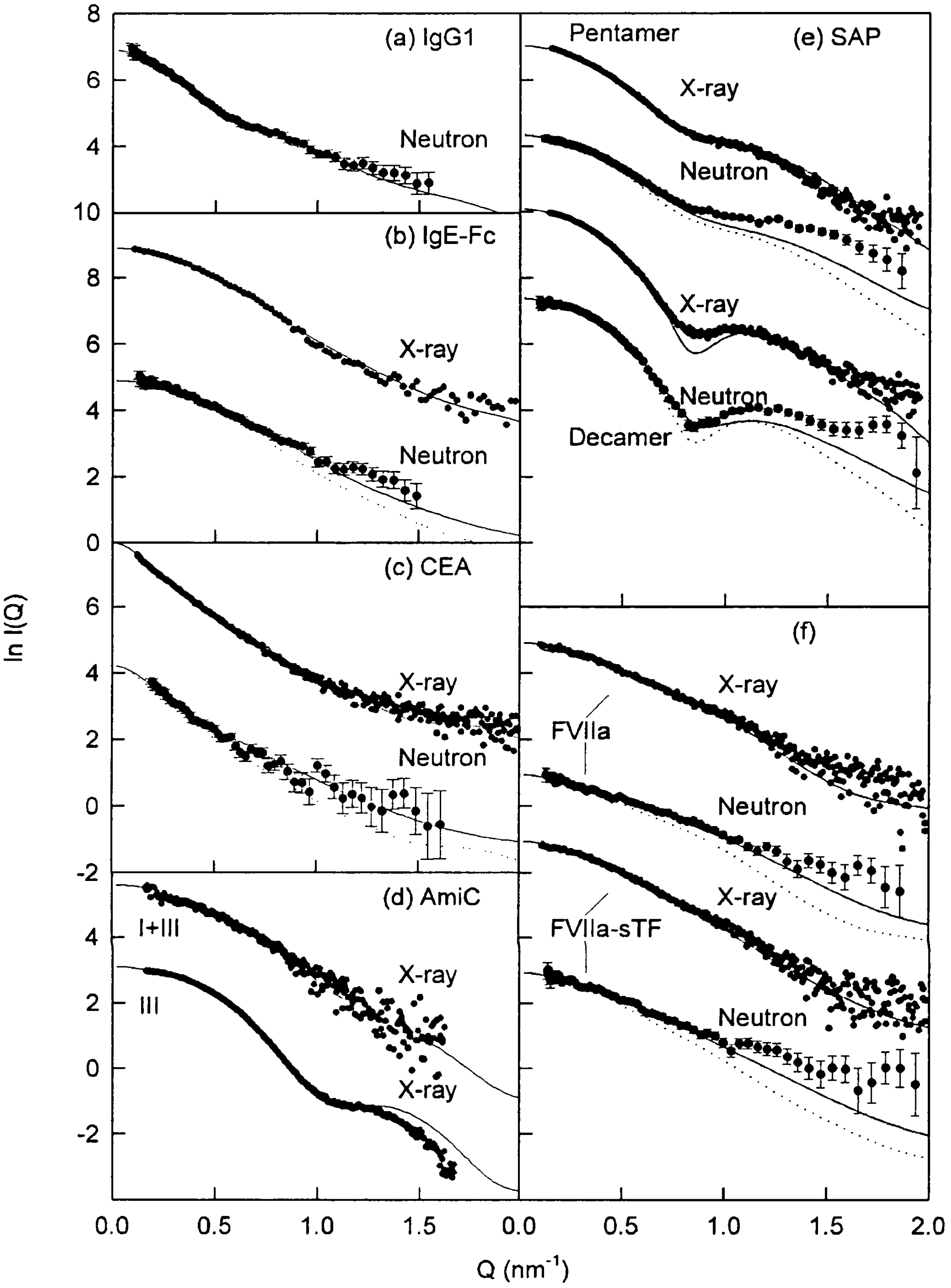


Figure 3.9: Final X-ray and neutron curve fits based on the best-fit models from Figure 3.8. The X-ray data were obtained from Stations 8.2 for IgE-Fc and CEA, and from Station 2.1 for AmiC, SAP, FVIIa and the FVIIa-sTF complex. Neutron data using 100% ^2H_2O buffer systems were obtained from LOQ. The 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. (Reproduced from Perkins *et al.*, 1998).

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 $R_{0.8}$ to $R_{2.0}$). 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 $R_{0.8}$to $R_{2.0}$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 (Figures 3.8 and 3.9).

These procedures can also be used to calculate sedimentation coefficients from analytical ultracentrifugation experiments (Figure 3.6). 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.

CHAPTER 4

EXPRESSION AND PURIFICATION OF RECOMBINANT PROTEINS

(4.1) Introduction.

Structural studies require large amounts of highly purified proteins. For example proteins from the complement cascade occur at very different levels in plasma, (see Table 4.1 for examples). Table 4.1 shows that it would be easier to purify C3 from 1 l of plasma than Factor D from 1 l of plasma. Recombinant DNA technology changed all of this. It is now possible to produce proteins normally found at levels of mg/l in plasma or other biological fluids at equivalent or higher levels of mg/ml or even at g/ml by the use of expression systems. Using an expression system then also allows the possibility of protein engineering or mutagenesis to see what individual domains or residues actually do.

Complement proteins (and complement protein domains) which have been successfully expressed by recombinant means include the vWF type A domains of factor B and complement receptor 3, C-type lectin domains of mannose binding protein, whole factor I (Chapter 6) and domains of factor I, factor D, and the SCR domains 5, 15 and 16 of factor H.

Expression systems can be grouped into 2 categories depending on the expression vector host, namely prokaryotic and eukaryotic. Each category has certain characteristics which may make them more suitable for a particular protein.

(4.2) Prokaryotic Systems.

Prokaryotic systems are ones which have a prokaryotic organism as the vector host, namely eubacteria and the Archaebacteria (or Archaea as they are now known).

	M_r (kDa)	Approx. serum conc. (mg l^{-1})	Domains
Classical Pathway			
C1q	457	80	stalks, head
C1r	172	50	2 RS, 2 SCR, EGF, SP
C1s	158	50	2 RS, 2 SCR, EGF, SP
C4	197	600	C3/C4/C5
C2	102	20	vWF, SP, 3 SCR
C3	187	1300	C3/C4/C5
Alternative Pathway			
Factor D	24	1	SP
Factor B	89	210	3 SCR, vWF, SP
Terminal Pathway			
C5	194	70	C3/C4/C5
C6	107	64	3 TSR, LDLr, PLR, EGF, 2 SCR, 2 FIM
C7	95	56	2 TSR, LDLr, PLR, EGF, 2 SCR, 2 FIM
C8	154	55	2 TSR, LDLr, PLR, EGF
C9	66	59	TSR, LDLr, PLR, EGF
Control proteins			
Plasma			
C$\bar{1}$ inhibitor	71	200	N-terminus, serpin
Factor J	20	5	
Factor I	74	35	FIM, CD5, 2 LDLr, SP
Properdin (trimer)	162	20	3 x 6 TSR
C4BP	491	250	7 x 8 SCR + 3 SCR
Factor H	150	480	20 SCR
S-Protein	83	505	
SP-40,40	70	100	coiled-coil
Carboxy-peptidase N	310	35	
Membrane bound			
MCP	45-70		4 SCR, ST, U, TM, CYT
DAF	70		4 SCR, ST, G
HRF	65		
CD59	18-20		murine LY-6 Antigen
Receptors			
CR1	160, 190 220, 250		30 SCR, TM, CYT
CR2	140		16 SCR, TM, CYT
CR3	265		vWF, 3 MB, TM, CYT
CR4	245		vWF, 3 MB, TM, CYT
C5a receptor	39		7 TM
C1q receptor	56		

Abbreviations

CYT	cytoplasmic domain	SCR	short consensus repeat
EGF	epidermal growth factor	SP	serine protease domain
G	glycolipid anchor	ST	serine/threonine-enriched area
FIM	factor I module	TM	transmembrane domain
LDLr	LDL receptor	TSR	thrombospondin repeat
MB	metal binding domain	U	unknown functional significance
PLR	perforin like region	vWF	von Willebrand Factor
RS	C1r/C1s domain		

Table 4.1: Summary of the physiological concentrations and domain structures of the complement components. Table reproduced from Smith, 1992.

The most widely used organism is the Gram-negative enteric bacterium *Escherichia coli*. This is because it is the best understood organism in nature because it has been extensively studied for nearly forty years. Recombinant DNA technology is a direct extension of the biochemical and genetic analyses carried out in the 1960's and 1970's. Even before the advent of molecular cloning, genetically altered *E. coli* strains were used to produce quantities of proteins of interest. When cloning techniques became available, most cloning vectors utilised *E. coli* as the vector host, and as progress was made *E. coli* was then used for large scale protein production. However, other bacteria have been used for molecular genetical and structural biology studies. An example of this is the negative regulator of amidase expression in *Pseudomonas aeruginosa*, (AmiC, Chapter 5). *Streptomyces lividans* is another example of a Gram-negative bacterium that has been used for the secretory production of proteins. For example the Fv fragment of mAb HyHEL10 was successfully expressed under the control of the streptomyces subtilisin inhibitor (ssi) promoter, where yields of 1 mg/l were reported (Ueda *et al.,* 1993). Secretion was achieved by linking both the V_H and V_L genes to the ssi signal protein. The Gram-positive bacterium *Bacillus subtilis* has the ability to secrete proteins directly into the medium in high concentration (Doi *et al.*, 1986) and has been used to express functional anti-digoxin scFv yielding 5 mg/l (Wu *et al.*, 1993). A major disadvantage with *B. subtilis* is that it also secretes a high level of proteases into the media as well, so multiple protease-deficient hosts have had to be developed to prevent the product being degraded. Another Gram-positive bacterium successfully used for protein secretion is *Staphylococcus carnosus*. This organism is unlike other Gram-positive bacteria in that it exhibits low exoproteolytic activity so there are less problems with product degradation. This organism has been successfully used for the expression of a fusion protein consisting of a lipase pre-proprotein gene fragment fused to a synthetic gene coding for the immunoglobulin REI variable domain (Pschorr *et al.,* 1994). Acceptable yields of 10 mg/l of soluble, correctly folded protein were reported. The fusion junction included a poly-histidine sequence to aid purification and a recognition sequence for *Neisseria gonorrhoeae* IgA protease for site-specific cleavage.

(4.2.1) General Strategy for Gene Expression in *E. coli.*

The basic approach used to express all foreign genes in *E. coli* (or any other bacterial system) begins with insertion of the gene of interest (or cDNA) into an expression vector. Vectors usually contain several elements:

1. Sequences encoding a selectable marker that assures maintenance of the vector in the host (e.g. antibiotic resistance).
2. At least one origin of replication to allow the vector to be copied from parent to daughter cells (origins are specific for different cell types).
3. A system to turn the gene of choice on or off by the use of a controllable promoter (e.g., *lac*, *trp*, or tac).
4. Translational control sequences, such as an appropriately positioned ribosome binding site and initiation codon (ATG).
5. A polylinker or multiple cloning site to simplify the orientation of the gene of choice in the correct direction. Most multiple cloning sites have been engineered to contain multiple type II restriction endonuclease sites to allow ease of insertion and rescue of the gene of choice.
6. For protein expression, a region coding for a "tag" to assist protein purification is often used (see Section 4.4.1).

Once constructed, the expression vector is introduced into an appropriate *E. coli* strain by transformation. *E. coli* was thought not to be naturally competent, where competence is the ability to take up foreign DNA. However, it has recently been reported that *E. coli* has been naturally transformed in fresh water (Baur *et al.,* 1996). Before transformation can take place the *E. coli* strain needs to be made competent. This can be achieved by chemical means (e.g. treatment with $CaCl_2$ in conjunction with heat and cold shock), or by the use of electric means to cause temporary holes in the plasma membrane which allows DNA into the cells (electroporation). Alternatively, competent cells can also be purchased from various life science companies. Once the cells have been transformed, they should be plated out and allowed to grow overnight at 37°C on a medium containing a selective agent to which the vector encodes a resistance gene to.

(4.2.1.1) Plasmid Vectors.

Bacterial plasmids are self-replicating, circular extrachromosomal molecules of DNA. Plasmids are naturally occurring, and bacterial and some eukaryotic cloning vectors are derived from them. Plasmids are desirable as cloning vectors because:-

1. Low molecular weight. Plasmids are normally a few kb in size, and are more stable, since the bigger the DNA molecule is, the easier it is to shear. Because of the low molecular weight, they also tend to occur in high copy numbers which makes purifying plasmid DNA easier.
2. Confer readily selectable phenotypes.
3. Single sites for restriction enzymes (cloning sites).
4. Cloning sites can be present in genes with selectable phenotypes, where loss of one phenotype in a two phenotype cloning system can be indicative of the presence of an

insert. For example pBR322 has tetracycline and ampicillin resistance with unique restriction enzyme sites in both resistance genes (Brown, 1991).

(4.2.1.2) Bacteriophage Vectors.

Bacteriophage vectors are based on λ viruses of *E. coli.* λ can have 2 alternative phases in its life cycle:-

1. Lytic cycle where the phage enters the host cell, makes multiple copies of itself and causes the cell to lyse. This can be seen as a plaque forming on a lawn of *E. coli.*
2. Lysogenic cycle where the phage DNA integrates into the host's chromosome. The cell lives and is resistant to further phage infection. The virus however can become lytic at a later time.

λ viruses are self replicating in an appropriate host with several genes mapped by classical genetics, and has a single linear double stranded chromosome of ~48.5 kb. Its entire sequence has been known since 1982. At each end of the chromosome are short stretches of single stranded DNA complementary in sequence which can form a circle. These ends are natural cohesive termini or *cos* sites. Wild type λ has several restriction sites. Derivatives have been found or created in which:-

1. Insertional vectors where a single restriction site for insertion of foreign DNA (e.g. λgt10). Insertion vectors can hold 0-10 kb, depending on the type used (different phage have different maxima) (Brown, 1991).
2. Replacement vectors where 2 restriction sites are used to remove a section of λ DNA which codes for recombination genes (non-essential to viral function) which is then replaced by foreign DNA (e.g. λEMBL 4). Replacement vectors can hold over the range

of 0-24 kb (different phage will have different top and bottom ranges) (Brown, 1991).

Once the vectors have inserts cloned into them, the constructs are then packaged *in vitro* to form infectious λ particles. Recombinant phage can be selected using specific *E. coli* hosts containing the *hflA* (high frequency of lysogeny by phage λ) gene (e.g *E. coli* C600*hflA*) (Brown, 1991). These particular strains do not form plaques when infected with non-recombinant λ, the virus goes into a lysogenic state. If a recombinant λ with the cI repressor inactivated is used, plaques are produced and not lysogens. λ bacteriophage are suited for genomic library work for the isolation of genes of interest rather than for protein expression.

(4.2.1.3) Cosmid Vectors.

Cosmid vectors are a derivative of λ viruses because they make use of a property of cos sites. Concatamers of λ DNA are produced if the cos sites are ~38-52 kb apart. Only a small region around the cos site is needed for recognition, therefore any sequence ~38-52 kb long between cos sites can be packaged. Such a particle can infect *E. coli* but there will be no viruses produced and no plaques formed so selection has to be based on a dominant selectable characteristic such as antibiotic resistance. Cosmid is a term used for any plasmid containing a cos site (Singleton and Sainsbury, 1987).

(4.2.1.4) Single Stranded Vectors.

A series of highly useful cloning vectors also have been developed from the single-strand DNA bacteriophage M13. When this phage infects bacteria, the strand that is packaged in the phage's capsid (the plus strand) replicates to form a double-stranded intermediate known as the replicative form (RF). The RF of M13 is structurally and functionally similar to a plasmid and can be isolated from bacterial cells using standard laboratory procedures for harvesting plasmid DNA. Foreign DNA can be inserted at any one of a number of unique restriction sites in the 7200 bp M13 genome. Depending on the orientation, one or the other strand of the inserted fragment will be packaged into progeny virus particles along with the plus strand of M13. Because the entire sequence of M13 is known, it is then a simple task to sequence or PCR out the insert.

(4.2.1.5) Shuttle Vectors.

Vectors that include replication systems derived from more than one host species are known as shuttle vector. Such vectors commonly include a replication system able to function in *E. coli*, as well as a replication system able to function in a second host which may be bacterial or eukaryotic (Zubay, 1987). The rationale behind this is that it is easy to produce large amounts of plasmid from an *E. coli* culture.

(4.3) Eukaryotic Systems.

Eukaryotic systems are ones which have a eukaryotic organism or cell type as the vector host. These include yeast, fungi, animal, plant and insect cells (and in some cases whole organisms).

(4.3.1) The Baculovirus Expression System.

(4.3.1.1) Virus Characteristics.

The family Baculoviridae consists of two sub-families, the *Eubaculovirinae* (the occluded viruses) in which the virions are embedded in a protein matrix and the

Nudibaculovirinae (the non-occluded viruses) (Francki *et al.,* 1991). The occluded viruses can be divided into two genera based on the morphology of their protein occlusion body (OB). In the genus Nuclear Polyhedrosis Virus (NPV), the occlusion bodies are polyhedra-shaped and range in size from 0.5 to 15 μm across (Bilimoria, 1991), and in the genus Granulovirius (GV) they are much smaller, ellipsoidal in shape and resemble granules ranging in size from 0.3 to 0.5 μm in length (Crook, 1991). The occlusion body protein, polyhedrin in the case of NPVs and granulin in GVs, has a Mr of approximately 29 000. The occlusion bodies of the NPVs contain many enveloped virions. In the sub-genus multiply-enveloped NPV (MNPV), the virions enclose several nucleocapsids per envelope. However, in the sub-genus singly-enveloped NPV (SNPV), the virions usually contain a single nucleocapsid per envelope. Generally, only one virion containing a single nucleocapsid is embedded in each occluded granulosis virus (GV). The occlusion body (OB) enables the virus to persist in the environment, between host larvae or between generations and may involve an interval of one or more years. In the sub-family *Nudibaculovirinae*, OBs are absent and the particle consists of a single enveloped nucleocapsid (Winstanley and Rovesti, 1993).

Baculoviruses have rod-shaped virions, approximately 250 × 50 nm in the case of the *Eubaculovirinae*, containing at least 25 different polypeptides and a double-stranded circular supercoiled DNA genome, packaged with an arginine-rich basic protein into a cylindrical capsid. This is surrounded by an envelope, with an intermediate layer between the capsid and the envelope (Winstanley and Rovesti, 1993). Genome sizes of approximately 100 to 180 kbp have been reported for GVs (Crook, 1991) and 90 to 170 kbp for NPVs (Bilimoria, 1991) and probably encode up to 100 genes.

(4.3.1.2) Insect hosts.

Baculoviruses are named after the host species from which they were isolated, e.g *Cydia pomollea* GV (CpGV) was isolated from codling moth larvae (Winstanley and Rovesti, 1993).

There are over 600 baculoviruses reported. The highest proportion have been NPVs, although the number of insect species from which GVs have been identified is currently in the order of 150 (Winstanley and Rovesti, 1993). MNPVs generally have a wider host range than SNPVs and GVs and many can be grown in insect cell culture (Grandos and Hashimoto, 1989). The most studied baculovirus, *Autographa californica* MNPV (AcMNPV) has a host range of over 39 insect species in 13 families, and can grow in several insect cultures (King and Posse, 1992; Winstanley and Rovesti, 1993). There is even evidence that AcMNPV can replicate in a mosquito cell line, having crossed the order barrier (Döller, 1985). AcMNPV, however is not regarded as a typical MNPV, having an unusually wide host range (Winstanley and Rovesti, 1993).

(4.3.1.3) Viral replication *in vivo.*

The larval or caterpillar stage of the insect's life cycle is the most susceptible to infection with NPVs. Polyhedra are ingested when the insect feeds on contaminated leaves and are dissolved in the alkaline environment of the mid-gut to release the virus particles (Figures 4.1 and 4.2). After passing through the peritrophic membrane lining the gut, the virus lipoprotein envelope fuses with the plasma membrane of the gut wall and liberates nucleocapsids into the cytoplasm (Figures 4.3-4.5) (King and Possee, 1992). The nucleocapsid is then transported into the nucleus of the cell; it is unclear if the DNA is injected through a nuclear pore or if the entire nucleocapsid enters the capsid. It was apparent in early *in vivo* experiments that virus replication was bi-phasic producing two distinct structural forms of the virus. In the infected gut cells, nucleocapsids are formed by about eight hours post-infection (h p.i.) and begin to bud through the nuclear membrane by 12 h p.i., and acquire a lipid envelope. This membrane appears to be "lost" in the cytoplasm, but the nucleocapsid gains another as it buds through the plasma membrane. During this latter process, the virus acquires a virus-encoded glycoprotein of 67 kDa (gp67; Whitford *et al.*, 1989), which is inserted into the plasma membrane. This protein most probably serves to attach the budded virus to other susceptible cells in the insect. In cell culture the budded virus is 1000-fold more infectious than virus particle released from polyhedra, which lack gp67. Budded virus is released into the haemolymph to infect other cells (2° infection). The cells infected in the second round of infection also cause the cell to produce budded virus and occluded virus in polyhedra. The accumulation of polyhedra within the insect proceeds until the host consists almost entirely of a bag of virus. In the terminal stages of infection the insect liquefies and releases polyhedra which can infect other insects (King and Possee, 1992).

(4.3.1.4) Viral replication *in vitro.*

Continuous culture of insect cells *in vitro* is relatively easy to establish. Starting material may include pupal ovarian tissue or fat bodies, haemocytes and other organs from larvae or homogenate of entire larvae (King and Possee, 1992). Using specific tissue is advantageous in that cell lines can be derived which do not simply reflect the fastest cell type and may be more useful for propagating virus or for recombinant protein production.

Cell lines which support the replication of AcMNPV have been derived from *Spodoptera frugiperda* (Sf) (Fall army worm) pupal ovarian tissue (Vaughn *et al.*, 1977) or ovaries from adult (cabbage looper) *Trichoplusia ni* (Hink, 1970; Davis *et al.*, 1992; Wickham, *et al*, 1992; Wickham and Nemerow, 1993), and *Mamestra brassicae* (cabbage moth) (King and Possee, 1992).

The study of baculovirus replication *in vitro* has greatly simplified experiments to understand the kinetics of virus gene expression and replication. It was an essential prerequisite for the development of the baculovirus expression vector system. Baculovirus gene expression has been divided (by inhibitor studies) into four phases. These are: intermediate-early (α); delayed-early (β); late (γ); very late (δ). In general, the expression levels attained in each succeeding phase is higher than that of the preceding one. Many useful AcNPV genes have been mapped, cloned and sequenced. These include the late AcNPV genes, basic viral protein and p39 and the very-late polyhedrin and p10 genes (King and Possee., 1992; O'Reilly *et al.*, 1992). Polyhedrin and p10 have been shown to be non-essential for viral infection and replication under tissue culture conditions (Smith *et al.*, 1983a). The polyhedrin protein which is the major component of the occlusion bodies has a molecular weight of 29 kDa and can accumulate to very high levels. Up to 1 mg/ml of polyhedrin may be synthesized per $1\text{-}2\times10^6$ infected cells accounting for 30-50% of the total insect protein (Summers and Smith, 1978; Gruenwald and Heitz, 1993).

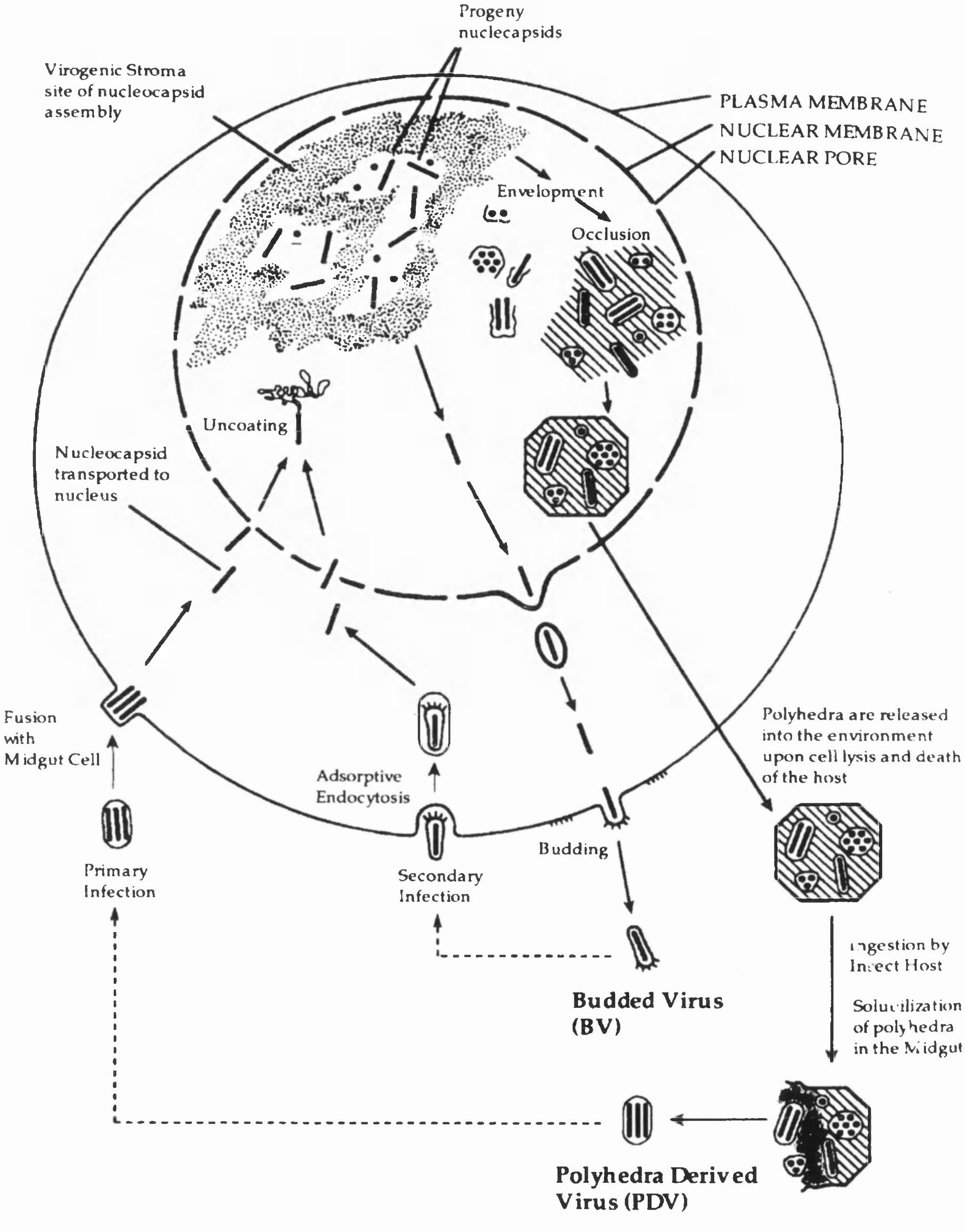

Progeny nuclecapsids
Virogenic Stroma site of nucleocapsid assembly
PLASMA MEMBRANE
NUCLEAR MEMBRANE
NUCLEAR PORE
Envelopment
Occlusion
Uncoating
Nucleocapsid transported to nucleus
Fusion with Midgut Cell
Adsorptive Endocytosis
Budding
Primary Infection
Secondary Infection
Polyhedra are released into the environment upon cell lysis and death of the host
Ingestion by Insect Host
Solubilization of polyhedra in the Midgut
Budded Virus (BV)
Polyhedra Derived Virus (PDV)

Figure 4.1: Typical cellular infection cycle of the nuclear polyhedrosis virus. (Reproduced from Winstanley and Rovesti, 1993).

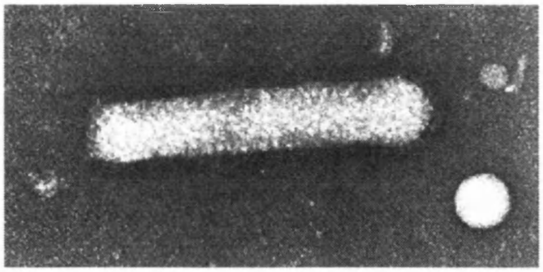


Figure 4.2: High magnification electron micrograph showing a negatively-stained baculovirus virion. Note the asymetric capsid structure and the presence of an envelope with surface projections (peplomers). (Taken from http://meds-ss10.meds.queensu.ca/~carstens/).

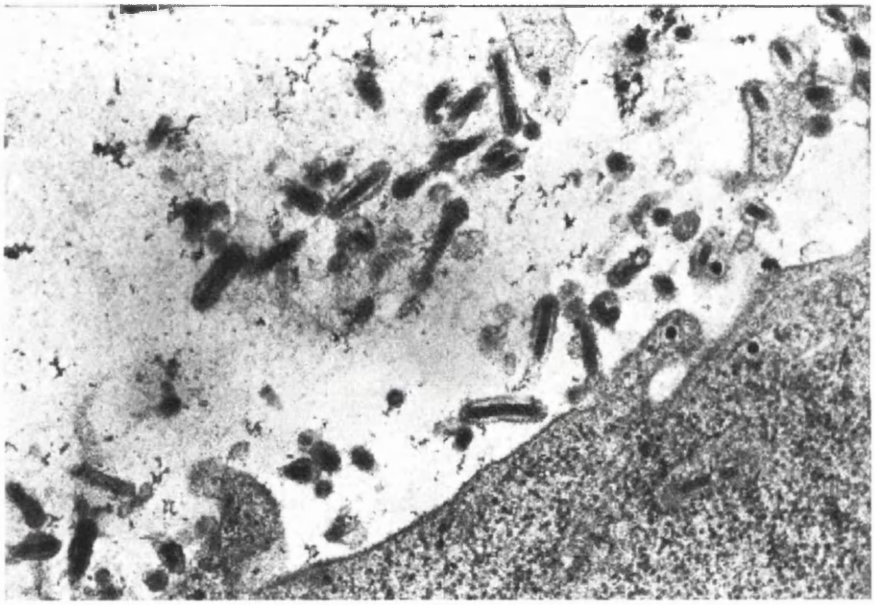


Figure 4.3: Electron micrograph showing a thin section of an insect cell infected with the baculovirus AcMNPV. A portion of the cell cytoplasm is seen in the bottom right hand corner. Many enveloped extracellular virions have budded through the cytoplasmic membrane and are visible outside the cell. (Taken from http://meds-ss10.meds.queensu.ca/~carstens/).

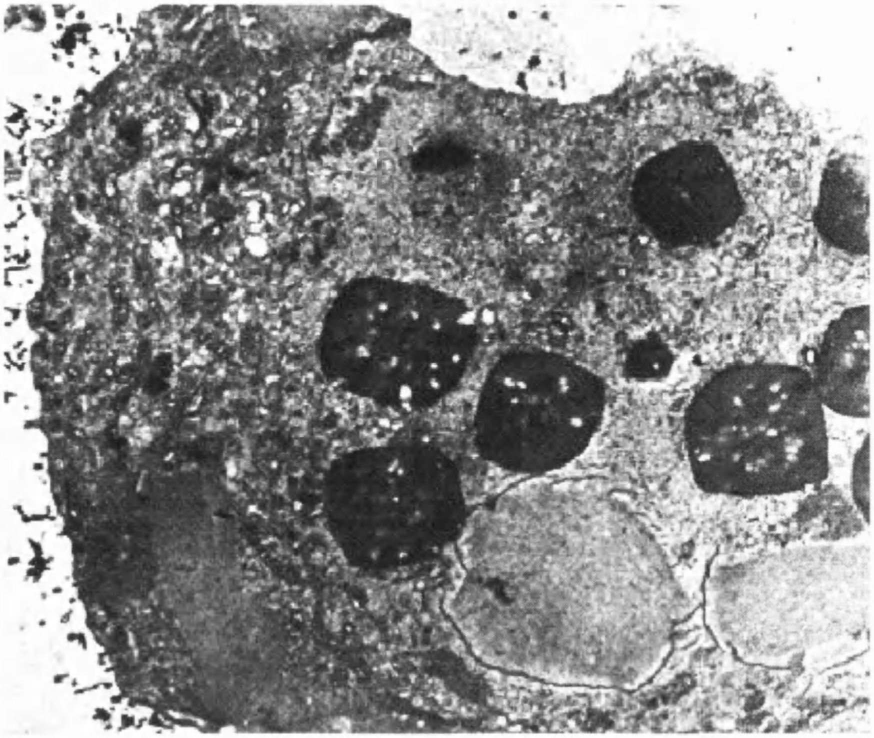


Figure 4.4: Low magnification electron micrograph showing a thin section of an insect cell infected with the baculovirus AcMNPV. The nucleus contains many polyhedra which in turn contain many occluded enveloped virions. Several different types of virus inclusion bodies are also visible in both the nucleus and cytoplasm. (Taken from http://meds-ss10.meds.queensu.ca/~carstens/).

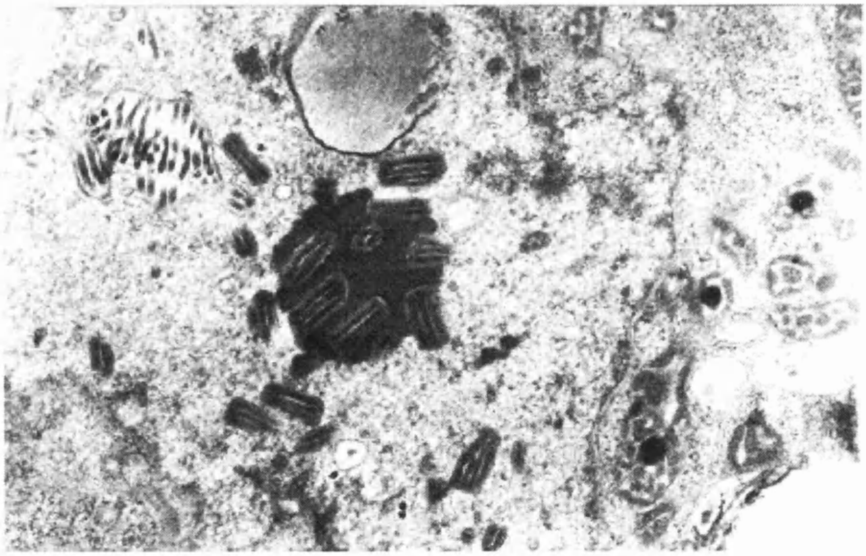


Figure 4.5: High magnification electron micrograph showing a thin section of an insect cell infected with the baculovirus AcMNPV. A portion of the nucleus containing enveloped virions in the process of being occluded into a developing polyhedron is shown. (Taken from http://meds-ss10.meds.queensu.ca/~carstens/).

(4.3.1.5) Development of the Baculovirus Expression System.

Several strong promoters of the late and very-late genes have been used to make a variety of baculovirus transfer plasmids. All these plasmids contained an *E. coli* origin of replication and an ampicillin resistance gene. This allowed the plasmids to be amplified in *E. coli* using standard techniques. Purified recombinant plasmid (containing the gene of choice under control of one of the baculovirus promoters) would then be co-transfected with linearized AcNPV DNA into insect cells. Hopefully after several days recombinant viruses would arise from homologous recombination between the plasmid and the genomic viral DNA (Figure 4.6) (King and Possee, 1992;'Gruenwald and Heitz, 1993).

Artificial deletion or insertional inactivation of the polyhedrin gene of AcNPV wild type virus would result in the production of occlusion body-negative viruses. Plaques from these viruses are distinctly different from those of occlusion body-positive wild type viruses. Modified AcNPV viruses now exist which allow colour selection to identify recombinants (e.g the Bac-N-Blue™ system, Invitrogen) or even permit positive survival selection for recombinants (e.g the BaculoGold™ system, Pharmingen) rendering occlusion body identification methods obsolete.

(4.3.1.6) Advantages of using the Baculovirus Expression System.

Overexpression of recombinant protein in insect cells is produced in an environment where the protein is properly folded, has correct disulphide bond formation, oligomerization, and most post-translational modifications. This will lead to a protein which should closely resemble its native counterpart, structurally and functionally (e.g

Ullman *et al.*, 1998; Table 4.2 and 4.3). However if the native protein functions as a heterodimer or relies on tissue or species specific modification the recombinant protein will not be functionally active, unless its binding partner or modifying enzyme is cloned into the same system and co-expressed. Recombinant baculoviruses have the ability to expand the size of their capsids to accommodate extra (larger than their genome size) DNA. In contrast to bacterial and other expression systems, the baculovirus expression system is capable of expressing the authentic protein and not a fusion protein. However in some situations fusion proteins may be desirable and can also be produced (Grunwald and Heitz, 1993).

Most post-translational modifications have been reported to occur properly in the baculovirus system. These include N- and O-linked glycosylation, phosphorylation, acylation, amidation, carboxymethylation, isoprenylation, signal peptide cleavage and proteolytic cleavage (Tables 4.2 and 4.3). The site where these modifications occur are often identical to those of the non-recombinant protein in its native cellular environment. However, the baculovirus system is designed to express the gene(s) of interest at an extremely high expression rate which may overwhelm the cellular processing apparatus. The highest expression level reported is 50% of the total cellular protein of an infected insect cell corresponding approximately to 1 g/10^9 cells (Grunwald and Heitz, 1993).

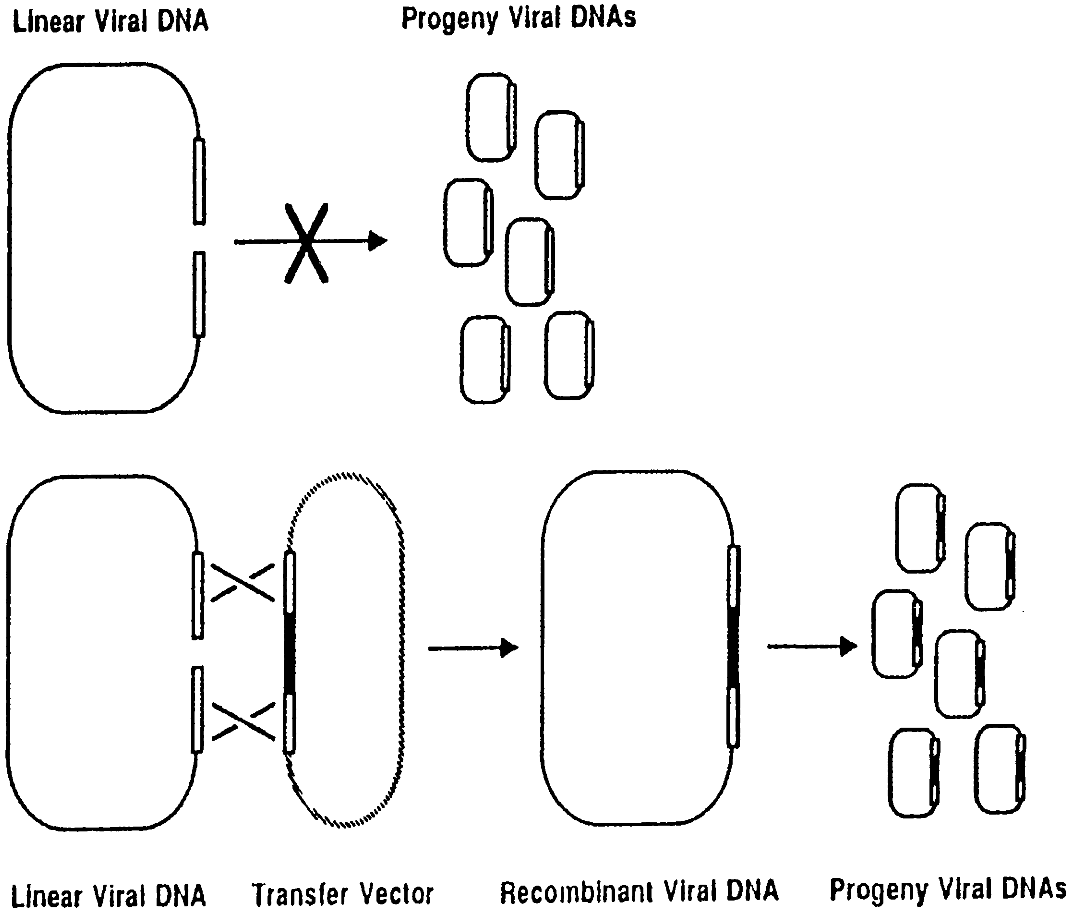


Figure 4.6: Rescue of linear virus DNA by recombination with a transfer vector. Upper panel: linear DNA cannot replicate because the replication apparatus of AcMNPV is designed to work on the native virus DNA which is circular. Lower panel: the circularity of the virus DNA can be restored by recombination with a transfer vector carrying DNA homologous to the viral sequences on either side of the break. A double cross-over generates a recombinant virus DNA molecule which, being circular, is competent for replication (Kitts *et al.*, 1990).

Protein	Species/virus	Membrane-targeted (MT) Secreted (S)	Reference
Virus examples			
env-gp85	Avian leukaemia virus	MT[1]	Noteborn *et al.* (1990)
Spike gp	Bovine corona virus	S	Yoo *et al.* (1991)
Haemagglutinin	Fowl-plague virus	MT[1,2]	Kuroda *et al.* (1986)
Surface antigen	Hepatitis B virus	S[1,2]	Kang *et al.* (1987) Takehara *et al.* (1988) Lanford *et al.* (1989)
Glycoprotein D	Herpes simplex virus (type 1)	MT[1,2]	Krishna *et al.* (1989)
gp160	Human immunodeficiency virus	MT[1,2]	Rusche *et al.* (1987)
Haemagglutinin	Human influenza virus	MT	Possee (1986) Kuroda *et al.* (1989) Kuroda *et al.* (1990)
Fusion glycoprotein	Human parainfluenza virus	MT[1,2]	Ray *et al.* (1989)
F glycoprotein	Human respiratory syncytial virus	MT[1,2]	Walthen *et al.* (1989)
Haemagglutinin	Japanese encephalitis virus	MT	Matsuura *et al.* (1989)
Peplomer gp (E2)	Murine coronavirius (JHM)	MT[1]	Yoden *et al.* (1989)
HN proteins	Parainfluenza virus (type 3)	MT	Van Wyke Coelingh *et al.* (1987)
G protein	Rabies virus	MT[1,2]	Préhaud *et al.* (1989)
Glycoprotein	Vesicular stomatitis virus	MT	Bailey *et al.* (1991)
E and NS1	Yellow fever virus	----[1,2]	Desprès *et al.* (1991a,b)
Non-virus examples			
Diuretic hormone	*Mamestra sexta* (tobacco hornworm)	S[1]	Magda (1989b)
Juvenile hormone esterase	*Heliothis virescens* (tobacco budworm)	S[1]	Hammock *et al.* (1990)
Acid ß-glucosidase	Human	MT	Grabowski *et al.* (1989)
ß-adrenergic receptor	Human		George *et al.* (1989)
CD4 receptor	Human	MT[1]	Webb *et al.* (1989)
EGF receptor	Human	MT[1]	Greenfield *et al.* (1988)
Glucocerebosidase	Human	S[1]	Bergh *et al.* (1990) Martin *et al.* (1988)
Haptoglobin	Human	S[1]	Heinderyckx *et al.* (1989)
Immunoglobulin Heavy chain (γ-1)	Human	S[1]	Hassemann and Capra (1990)

Protein	Species/virus	Membrane-targeted (MT) Secreted (S)	Reference
Immunoglobulin Light chain (91A3)	Human	S[1]	Hassemann and Capra (1990)
Insulin receptor	Human	MT	Herra *et al.* (1988) Paul *et al.* (1990)
ß-interferon	Human	S[1]	Smith *et al.* (1983b)
Myelin-associated glycoprotein	Human	S[1]	Johnson *et al.* (1989)
Plasminogen	Human	S[1]	Davidson *et al.* (1990) Whitefleet-Smith *et al.* (1989)
Chimeric plasminogen activators	Human	S	Devlin *et al.* (1989)
Poliovirus receptor	Human		Kaplan *et al.* (1990)
Tissue-type PA	Human	S[1]	Jarvis and Summers (1989)
Transferrin receptor	Human	MT[1]	Domingo and Trowbridge (1988)
Urokinase-type PA	Human	S	King *et al.* (1991b)
$GABA_A$ receptor	Bovine	MT	Atkinson *et al.* (1992) Joyce *et al.* (1993)
Phaseolin	*Phaseolus vulgaris* (French Bean)	S[1]	Bustos *et al.* (1988)

[1] Antigenic [2] Elicited neutralizing antibodies

Table 4.2: Examples of glycoproteins expressed using the baculovirus expression system. (Based on King and Possee, 1992).

Protein	Species/virus	Phosphorylated (P) Acylated (Ac)[1] Amidated (Am)	Reference
Virus examples			
E1A	Adenovirius	P	Patel *et al.* (1988)
E2 protein	Bovine papillomavirius (type 1)	P	McBride *et al.* (1989)
Core antigen	Hepatitis B virus	P	Lanford and Notvall (1990)
Surface antigen	Hepatitis B virus	Ac-M	Lanford *et al.* (1989)
$p17^{gag}$	HIV	Ac-M	Overton *et al.* (1989)
$p24^{gag}$	HIV	P	Overton *et al.* (1989)
$p40^{X}$ trans-activator	Human T-cell leukaemia virus (HTLV-1)	P	Jeang *et al.* (1987a) Nyunoya *et al.* (1988)
Nucleoprotein	Rabies virus	P	Préhaud *et al.* (1990)
Large T antigen	SV40 virus	P, Ac-P	Lanford (1988) Murphy *et al.* (1988)
gag precursor	Simian immunodeficiency virus (SIV)	Ac-M	Delchambre *et al.* (1989)
Non-virus examples			
Krüppel	*Drosophila*	P	Ollo and Maniatis (1987)
Diuretic hormone	*Mamestra sexta*	Am	Maeda (1989b)
EGF receptor	Human	P	Greenfield *et al.* (1988)
Insulin receptor	Human	P	Herrera *et al.* (1988) Paul *et al.* (1990)
c-*myc* proto oncogene	Human	P	Miyamato *et al.* (1985)
P^{210} BCR-ABL oncogene	Human	P	Pendergast *et al.* (1989)
Terminal transferase	Human	P	Chang *et al.* (1988)
Transferrin receptor	Human	Ac-P	Domingo and Trowbridge (1988)
Protein Kinase C-γ	Bovine	P	Patel *et al.* (1988, 1989)
p53	Murine	P	O'Reilly and Miller (1988)
Tyrosine hydroxylase	Rat	P	Fitzpatrick *et al.* (1990)
$pp60^{v\text{-}src}$ (rsk- α/β)	*Xenopus*	P	Vik *et al.* (1990)
Transposon Ac	*Zea mays* (corn)	P	Hauser *et al.* (1988)

[1] Ac-M = myristylation. Ac-P = palymitylation.

Table 4.3: Examples of foreign proteins that have been phosphorylated, acylated or amidated in insect cells using the baculovirus expression system. (Modified from King and Possee, 1992).

(4.3.2) Non Insect Cell Systems.

Whilst not used in this thesis, it is important to mention other eukaryotic expression systems.

(4.3.2.1) Yeast and Fungal Systems.

Various proteins including whole antibodies and Fab fragments have been expressed and secreted by the yeast *Saccharomyces cerevisiae*, generally by co-expression of heavy and light immunoglobulin genes on either one or two plasmids within the same cell (Verhoeyen and Windust, 1996). Yields of functional, reassembled antibody or antibody fragments have remained persistently low, at micrograms per litre although minor improvements have been made recently by changing the codon usage of the gene to codons more commonly used in yeast (Kotula and Curtis 1991), in contrast to the many *E. coli* expression systems for antibody fragments which have become increasingly efficient (Verhoeyen and Windust, 1996). This factor, when taken into account with other parameters such as possible plasmid instability and a tendency for hyperglycosylation of expressed proteins, means that *S. cerevisiae* is not the best system for the expression of whole antibodies or antibody fragments, and other recombinant proteins (Verhoeyen and Windust, 1996). Other types of yeast have been successfully used for the expression of recombinant proteins. *Schizosaccharomyces pombe* has been used for the successful expression of an anti-fluorescein scFv (Davis *et al.,* 1991), and methylotrophic yeasts such as *Pichia pastoris* and *Hansenula polymorpha* have also been shown to be useful for the production of recombinant proteins (Buckholz and Gleeson, 1991). These yeasts have several advantages over *Saccharomyces cerevisiae* in that they possess strong, stringently controlled promoters, high levels of expression, and a glycosylation pattern closer to mammalian cells (Verhoeyen and Windust, 1996).

Filamentous fungi have also been used for protein expression. For example *Aspergillus* spp and *Trichoderma* spp possess features which make them exceptionally useful. These include the ability to secrete large quantities (up to 25 g/l) of recombinant protein in culture. Utilising procedures used for antibiotic production it is possible to achieve large-scale production for low-cost using existing fermentation equipment (Verhoeyen and Windust, 1996; Stanbury *et al.*,1995). Signal sequences are accurately processed, protein is correctly folded and correct disulphide bond formation and, in some cases, glycosylation patterns similar to mammalian glycoproteins (Verhoeyen and Windust, 1996). The first reported expression of a dimeric recombinant protein in a filamentous fungi was the expression of an Fab from *Trichoderma reesei* (Nyyssonen *et al.*, 1993). The Fab was successfully secreted from *T. reesei* that had been co-transfected with plasmids encoding the Fd and light chain of a murine anti-2-phenyloxazolone. The yield was substantially increased when the Fd was fused to the *T. reesei* cellulase, cellobiohydrolase I (CBHI). Immunologically active Fab could be released from the fusion protein using an extracellular *T. reesei* protease, and had the same affinity as the parent immunoglobulin. The differences in yield were quite substantial, this being 1 mg/l for the Fab compared to 150 mg/l for the CBHI-Fab fusion protein. (Nyyssonen *et al.*, 1993; Verhoeyen and Windust, 1996).

(4.3.2.2) <u>Viral Expression Systems for use Within Animal Cells.</u>

Viruses are infectious agents that, in their simplest form, are composed of a viral genome (either single or double stranded DNA or RNA) surrounded by a protein coat and in some viruses a lipid envelope. Upon infection, viruses enter cells by either binding to a specific receptor on the cell surface or by fusion with the envelope to the plasma membrane. Following entry, the viral genome is uncoated and sequesters the translational and/or transcriptional machinery of the host to direct the synthesis of more virus particles, although the genome may exist in a latent, non-lytic stage. Because of the high level of expression of the viral encoded proteins from strong promoter elements within the virus genome, virus vectors are good vehicles for the expression of recombinant proteins within cultured animal cells. In addition, the advantage of mammalian proteins receiving authentic post-translational modifications is retained. Viruses that have been used for the expression of recombinant proteins include viruses that have both double stranded DNA and double and single (positive and negative sense) stranded RNA genomes.

The double stranded DNA viruses that have been engineered to express recombinant proteins include the adenoviruses, the poxviruses and the herpesviruses (Grunhaus and Horwitz, 1992; Moss, 1992; Gloriso *et al.*, 1992). Expression of foreign genes within adenoviruses may be achieved by recombination between the adenovirus genome and an adenovirus subgenomic fragment which contains the major late promoter, an inverted terminal repeat and the tripartite leader where the foreign gene is inserted (Grunhaus and Horwitz, 1992). This subgenomic fragment can be carried on a bacterial plasmid and transfected into susceptible cells. Insertion of recombinant DNA coding sequences directly into the adenovirus genome in the early gene regions E1, E3 or near E4 is also possible, however this may result in replication defective viruses. In some cases these viruses may yield higher levels of expression because of longer survival of the host cell and even integration into the host genome (Grunhaus and Horwitz, 1992). This has led to the investigation of the adenoviruses for use in gene therapy, in particular the intratracheal installation of the human cystic fibrosis transmembrane conductance regulator gene into the bronchial epithelial cells of cotton rats (Rosenfeld *et al.*, 1992). Integration is also a characteristic of the retroviruses and has led to their study as gene transfer vectors (Majors, 1992). Likewise, the herpesvirus HSV-1 (herpes simplex virus-type 1), which is a neurotrophic human virus, has also been investigated as a gene transfer expression vector because of its ability to exist extrachromosomally in a latent, non-lytic, state within neurones and direct expression from promoters (LATp1 and LATp2) within a region of the genome responsible for latency (Glorioso *et al.*, 1992). The poxviruses encode their own RNA polymerases and transcription factors which makes the expression of recombinant proteins possible by introducing a plasmid carrying a poxvirus promoter and the recombinant gene (flanked by regions of poxvirus DNA) into cells infected with poxvirus. The foreign gene can be introduced into the poxvirus genome by recombination between the genome and regions of poxvirus DNA carried on the plasmid. Insertional inactivation of the thymidine kinase gene or cotransfer of a β-galactosidase gene allows selection of recombinants (Moss, 1992). Amongst the RNA viruses, the single-stranded positive sense alphaviruses are the most versatile for the expression of recombinant proteins. Heterologous sequences can be engineered within the genome to provide recombinants that are replication and packaging competent or in place of the structural or non-structural genes to provide packaging defective or replication defective viruses respectively. Recombinant proteins expressed from these systems include human transferrin receptor, mouse dihydrofolate reductase and chicken lysozyme (Bredenbeek and Rice, 1992).

(4.3.2.3) The Use of Transgenic Animals For Heterologous Gene Expression.

Vectors described in the previous section can be used to introduce foreign DNA sequences into the embryos of animals to result in a whole animal system, known as a transgenic animal (Jänne *et al.*, 1992). The retroviruses are particularly useful in this respect because of their ability to integrate into genomic DNA, although the most widely used method is microinjection of the foreign genes into one of the two pronuclei of a fertilized oocyte (Jänne *et al.*, 1992). The first transgenic mice were described by Gordon *et al.* (1980), but larger transgenic animals such as rabbits, sheep and goats have since been developed. Expression of heterologous proteins in these animals can be placed under the control of the milk protein gene promoters (e.g. the β-casein and β-lactoglobulin gene promoters) resulting in a high-level secretion of recombinant proteins in the animal's milk. In one study, it was calculated that a transgenic goat expressing recombinant tissue plasminogen activator in its milk at a level of 3 mg/ml would produce, in a day's milk, the equivalent to a daily harvest of a 1000 litre cell culture bioreactor (Jänne *et al.*, 1992).

(4.3.2.4) Expression of Recombinant Proteins in Plant Cells.

Plant cells have been engineered for expression of recombinant proteins. Traits for herbicide tolerance and resistance to insect pests were engineered into tobacco plants in early experiments (Comai *et al.*, 1985; Hilder *et al.*, 1987). In the latter example, the *35S* RNA promoter of the cauliflower mosaic virus (CaMV) was used to direct the transcription of cloned sequence. This promoter and its enhancer sequences or a 42-bp sequence from the maize *Shrunken* 1 gene exon 1 (which is thought to improve splicing) have been employed in expression vectors for both monocotyledonous and dicotyledonous plants (Topfer *et al.*, 1993). Recently, these vectors have been utilised to express and secrete monoclonal antibodies in plant cells. The monoclonal antibodies were functional and were identical to the native proteins except for the glycosylation of the heavy chain. This has implications for cheap and almost limitless supply of monoclonal antibodies for passive immunization (Hiatt and Ma, 1993). The *35S* RNA promoter has also been successfully used in the creation of fertile transgenic soybeans highly expressing a synthetic *Bacillus thuringiensis cryIAc* toxin gene for insect resistance (Stewart *et al.*, 1996).

(4.4) Protein Engineering.

It is often desirable to alter the protein of interest. This could involve the addition of a "tag" to aid in purification (Figure 4.7), alteration of specific residues in a protein to see what they do, or could involve the removal of whole domains from the protein of interest.

(4.4.1) Engineering Proteins to Facilitate Purification.

In 1983 Uhlén and co-workers demonstrated the first example of a gene fusion for affinity purification of recombinant proteins. In the years that passed a large number of affinity-fusion systems were developed, and thousands of gene fusions have been used for affinity purification (Nygren *et al.*, 1994).

A large number of affinity tails have been developed to aid purification of recombinant proteins. These permit several alternative purification protocols. Factors such as proteolytic stability, solubility, the ability of the protein to be secreted, protein folding and the purification conditions have to be considered when choosing a suitable expression strategy. Proteins produced as inclusion bodies must be refolded, or at least solubilised, before affinity purification can take place. Kits based on several of these tails have been produced commercially, in which expression vectors are manufactured together with a corresponding affinity resin. Examples of these systems are shown in Table 4.4.

One of the best-characterized systems, staphylococcal protein A and its derivative ZZ, has been used successfully in several different hosts such as bacteria (Hammarberg *et al.*, 1990), yeast (Stirling *et al.*, 1992; Zueco and Boyd, 1992), Chinese hamster ovary (CHO) cells (Nygren, 1994) and insect cells (Andersons *et al.*, 1991). The need for a low pH (pH 3 is routinely used) for the elution of the protein from the affinity column can be circumvented by the use of a competitive elution strategy based on engineered competitive elution strategies based on engineered competitor proteins that can be removed efficiently from the eluate mix. The human polyclonal IgG used as the ligand in this system can be replaced by a recombinant Fc fragment. Only fusions using this tag placed C-terminally have been described (Nygren *et al.*, 1994)

The Strep-tag system (Schmidt and Skerra, 1993), uses a short (10 residue) peptide sequence with affinity for streptavidin which, in contrast with the natural ligand, biotin, can be eluted using mild conditions (1 mM iminobiotin). A similar system is the PinPoint™ kit which utilizes an *in vivo* biotinylation by *E. coli* of a 13 kDa sequence attached to the target protein. Purification is performed using immobilized monomeric streptavidin which enables mild elution conditions (5 mM biotin) to be used. A Lac-

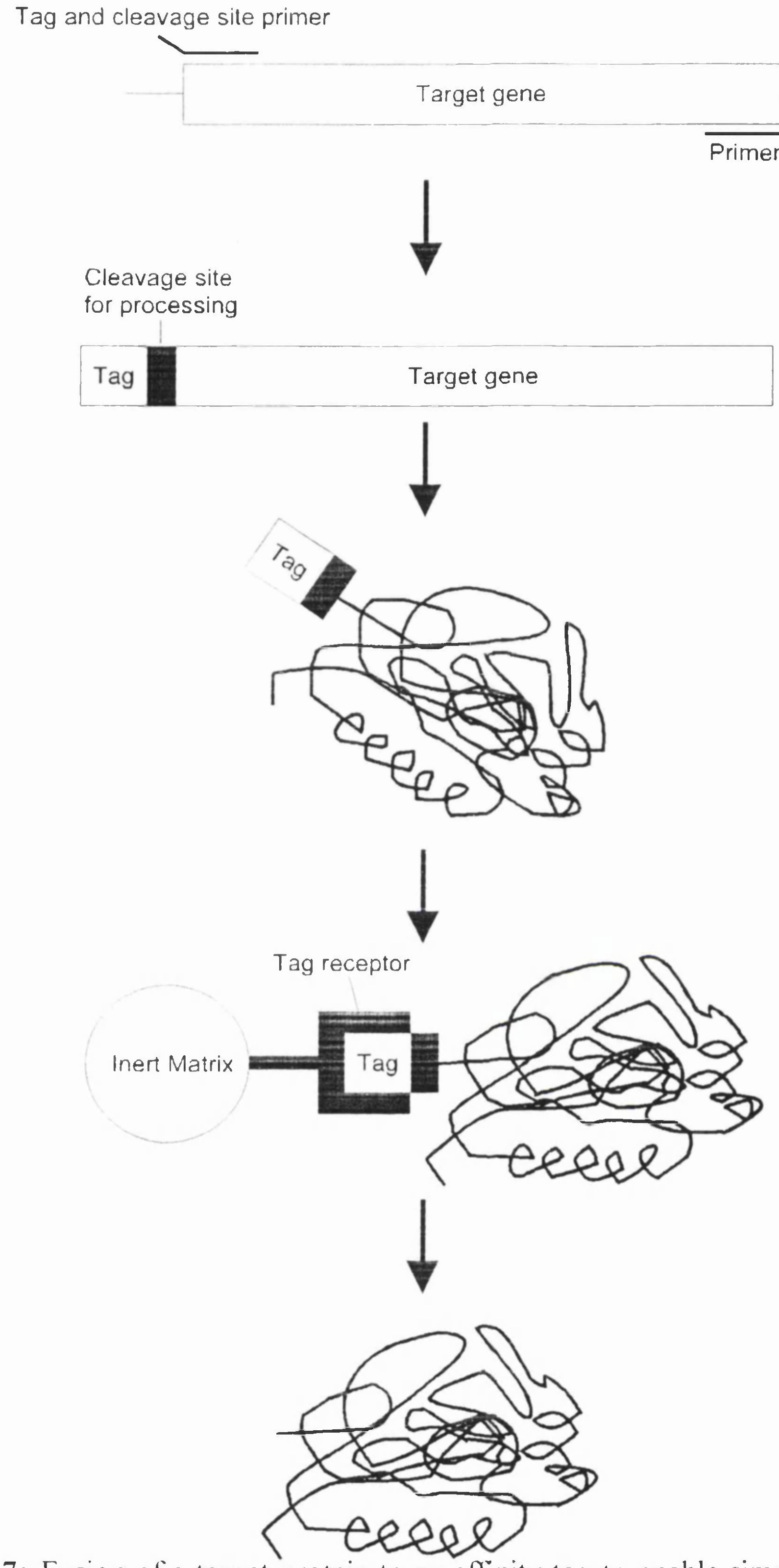


Figure 4.7: Fusion of a target protein to an affinity tag to enable simplified recovery using polymerase chain reaction (PCR). The tag sequence is introduced via a non-complementary handle sequence in one of the PCR primers used for amplification. The "tag" refers to the fusion partner and the "tag receptor" is the appropriate material for the "tag". For example if the "tag" is GST the "tag receptor" would be glutathione. (Based on Nygren *et al.*, 1994).

Fusion Partner	Size	Ligand	Elution Condition	Suppliers
ZZ	14 kDa	IgG	Low pH	Pharmacia, Sweden
His-tag	6-10 a.a	Ni^{2+}	Imidazole	Novagen, USA Invitrogen, USA Qiagen, USA
Strep-tag	10 a.a	Streptavidin	Iminobiotin	Biometra, Germany
PinPoint™	13 kDa[b]	Streptavidin[c]	Biotin	Promega, USA
MBP	40 kDa	Amylose	Maltose	New England Biolabs, USA
GST	25 kDa	Glutathione	Reducing agent	Pharmacia, Sweden
Flag™ Peptide	8 a.a	Specific mAb	Low calcium	IBI Kodak, USA

[a] Abbreviations: a.a, amino acids; His, histidine; GST, glutathione-S-transferase; MBP, maltose-binding protein.

[b] Encodes a domain biotinylated *in vivo* in *E. coli*.

[c] Monomeric streptavidin.

Table 4.4: Examples of commercial systems used in the production of recombinant proteins fused to affinity tails[a]. Based on Nygren *et al.*, 1994.

repressor-fusion-based peptide library has been developed where a 13 amino acid sequence was found to be sufficient to mimic the much larger domain normally recognized by the biotinylating enzyme (Schatz, 1993). Both N- and C-terminal fusions to the tail have been demonstrated to be functional and utilisation of a smaller tag results in a higher product:tail ratio (Nygren *et al.*, 1994).

The Flag™ system is based on the Ca^{2+} -dependent binding of a monoclonal antibody (mAb) to an eight-residue peptide containing an enterokinase recognition site fused N-terminally to the target protein. Elution using a low Ca^{2+} buffer makes this system useful for the recovery of sensitive proteins. Another gentle elution protocol system uses a fusion to maltose-binding protein (MBP) and a buffer containing maltose,. These fusions can be produced intracellularly or, alternatively, as secreted proteins (Nygren, 1994).

Another popular system involves the use of a polyhistidine affinity tag which allows the possibility of purifying the recombinant protein by immobilized metal ion affinity chromatography (IMAC) under denaturing conditions, such as 8 M urea or 6 M guanidine hydrochloride (Hochuli *et al.*, 1988; Ljungquist *et al.*, 1989). Another intracellular expression system is based on fusion of the protein of interest to glutathione-S-transferase (GST). This method utilises the affinity of GST for glutathione (usually attached to an inert support) (Smith and Johnson, 1988; Nygren *et al.*, 1994). Numerous examples have shown that high-level cytoplasmic production of heterologous proteins are frequently associated with precipitation of the product into insoluble inclusion bodies. However, if the target protein is fused to a highly soluble partner such as the GST moiety, up to 45% of the total cell protein can be produced intracellularly in *E. coli* in a soluble form (Ray *et al.*, 1993).

(4.4.2) Mutagenesis of Proteins.

Sometimes it may be desirable to elucidate what functions are performed by various parts of the protein. This could be specific to a predicted or known catalytic residue to elucidate its function, or to a change of its substrate specificity, or involve a removal of whole domains to try and elucidate its function.

Site-specific (or site-directed) mutagenesis is the *in vitro* induction of mutagenesis at a specific site in a target DNA molecule. Various methods have been used. In one method (D loop mutagenesis), small fragments of single stranded DNA (ssDNA), each corresponding to the site to be mutated, are mixed (*in vitro*) with the full-sized supercoiled double stranded DNA (dsDNA) target molecules in the presence of RecA protein (a recombination protein) and ATP. Under these conditions the ssDNA fragments invade the dsDNA to generate a single-stranded D loop which is vulnerable to the ssDNA-specific mutagen bisulphite (which converts cytosine to uracil). The bisulphite-treated DNA is then transformed into *ung*$^-$ bacteria (uracil-DNA glycosylase deficient cells, otherwise the mutation will be repaired). During a subsequent round of DNA synthesis, the uracil residues pair with adenine, leading to G·C-to-A·T transitions (Singleton and Sainsbury, 1987).

Another method involves the use of a chemically synthesised oligonucleotide containing the desired mutation(s) base sequence. The oligonucleotide can hybridize with (e.g.) a covalently closed circular (ccc) ssDNA target molecule (forming a heteroduplex) and can function as a primer, for example by the Klenow fragment. Treatment with a DNA ligase results in a ccc dsDNA molecule containing a mismatch at the site of the mutation. The molecule can then be introduced into a cell by transformation, and subsequent DNA replication will segregate the mutant and non-mutant strands. Recovery of mutants may be low owing to mismatch repair in the recipient cells (Singleton and Sainsbury, 1987). The procedure shown in Figure 4.8 uses a variation of the Deng and Nickoloff (1992), unique site-elimination mutagenesis technique. The loss of the unique restriction enzyme site (caused by the selection primer) in the mutated plasmid is the selection technique. Because the parental DNA will be linearised it will not be transformed as efficiently as the mutated DNA, by going through this selection and transformation procedure twice the majority of the parental DNA will be removed. The resulting colonies will still have to be screened to ensure the mutation has been properly incorporated. Selection can also be aided by incorporating a different novel restriction endonuclease site into the mutagenic primer and selecting colonies which cut with that particular enzyme.

Mutations can be created using the polymerase chain reaction (PCR) (Mullis and Faloona, 1987) (Figure 4.9). The procedure outlined in Figure 4.10 shows how inverse PCR can be used to introduce mutations. The selection procedure uses a restriction enzyme (*Dpn*I) which is active against methylated DNA which recognises GA↑TC (Brown, 1991). Because the PCR product is not methylated it will not be cut. Figure 4.11 shows how inverse PCR can be used to introduce a variety of mutations.

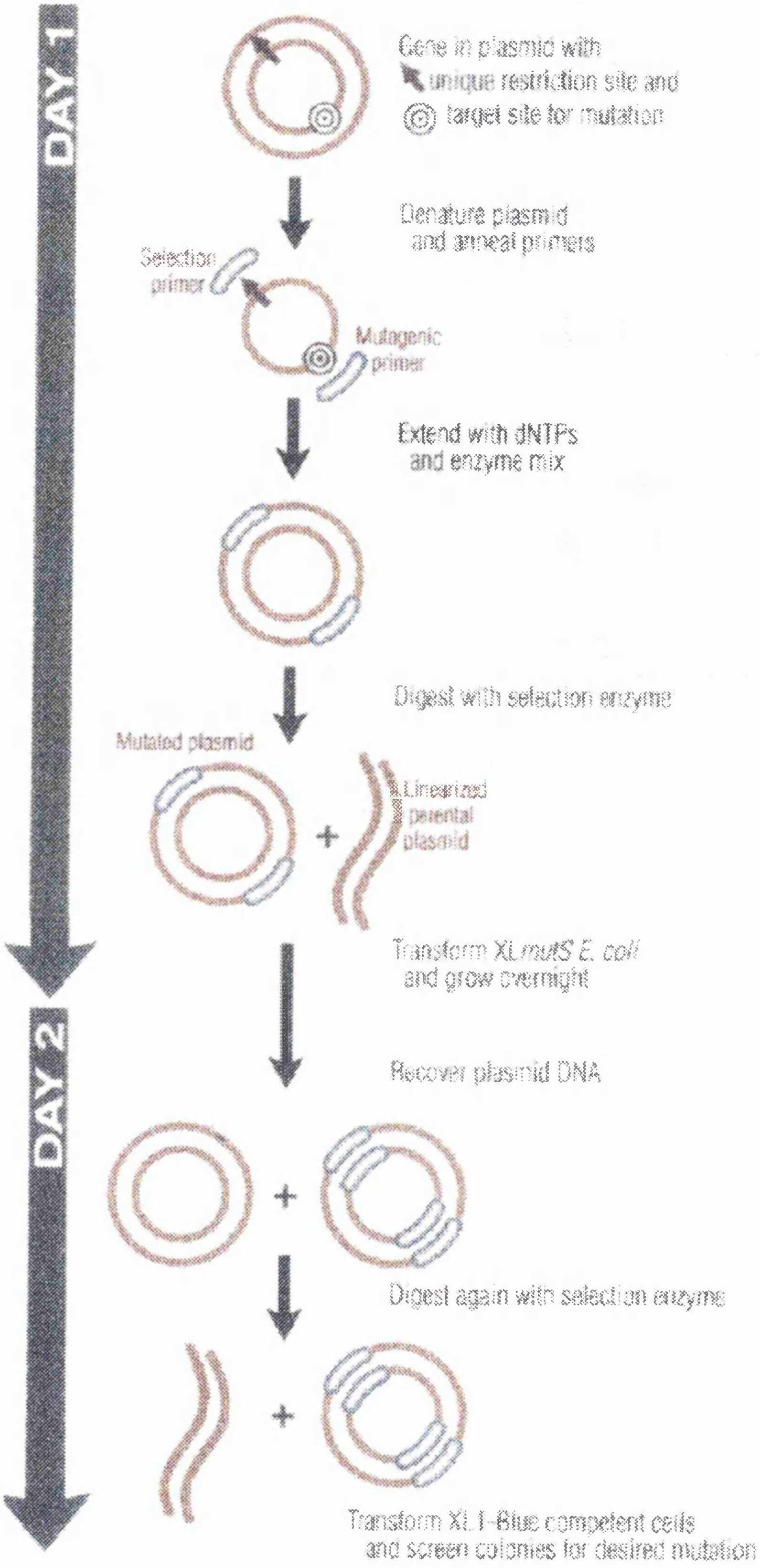


Figure 4.8: Schematic diagram demonstrating the use of mutated oligonucleotides for the introduction of mutations into a desired sequence. (Diagram modified from the 1997 Stratagene catalogue).

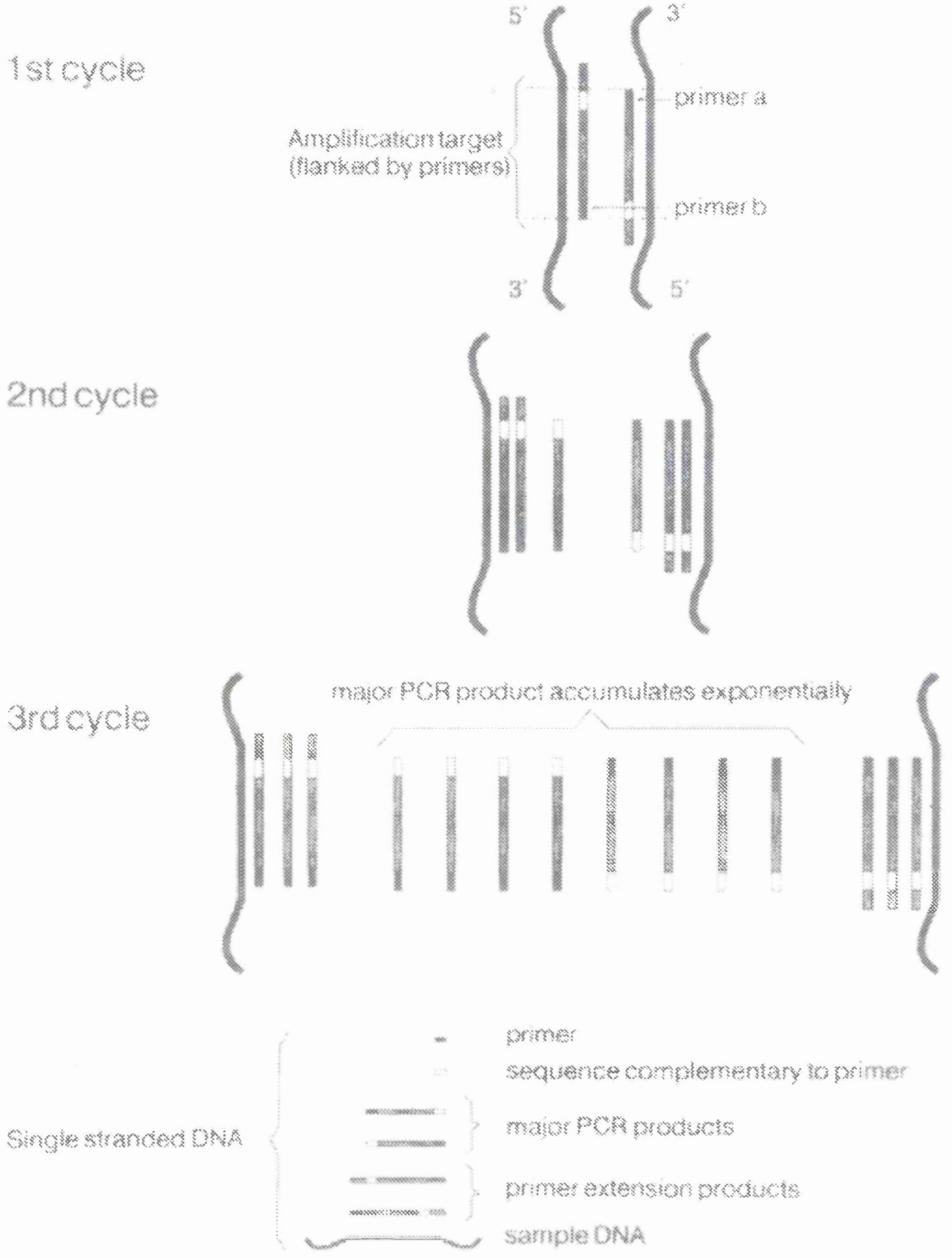


Figure 4.9: Schematic diagram of PCR. By using primer pairs a and b annealed to the complementary strands of DNA, two new strands are synthesized by primer extension. If the process is repeated, both the sample DNA and the newly synthesised strands can serve as templates, leading to an exponential increase of product which has its ends defined by the positions of the primers. Products with primer at only one end (and therefore of indeterminate length) increase at a linear rate throughout the process, and together with the starting DNA form only a small fraction of the total PCR product. For clarity, the DNA and the PCR products are shown as single stranded entities, as they would be at the denaturation stage of the reaction; the final product should be double-stranded, however, since the final step in PCR is a synthesis step. (Reproduced from Taylor, 1991).

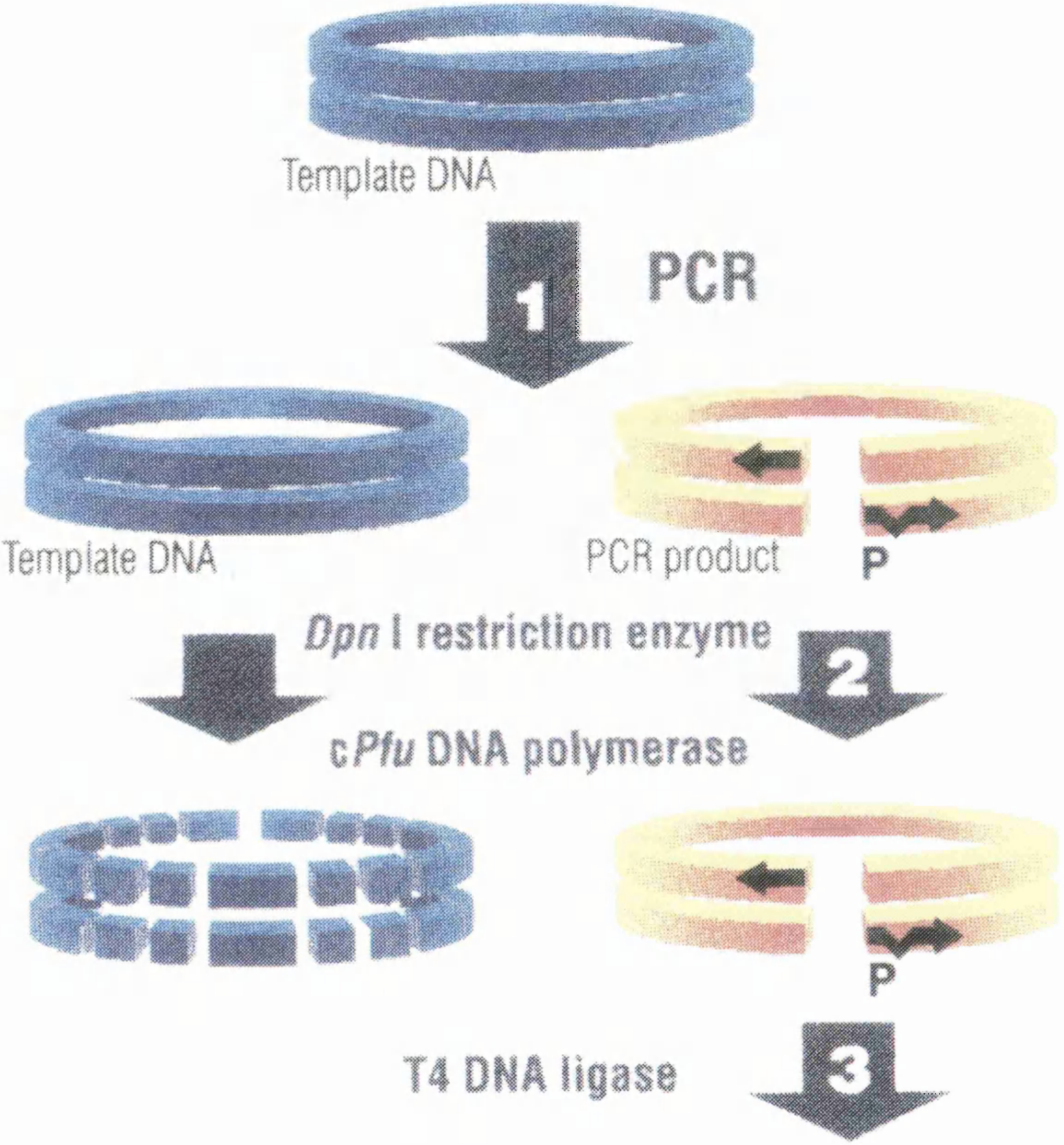


Figure 4.10: Schematic diagram demonstrating the creation of a mutation in a DNA molecule using inverse PCR. The selection step for mutants requires the use of *Dpn*I, a restriction endonuclease which is only active against *dam* methylated DNA so the PCR product is left intact. (Diagram modified from the 1997 Stratagene catalogue).

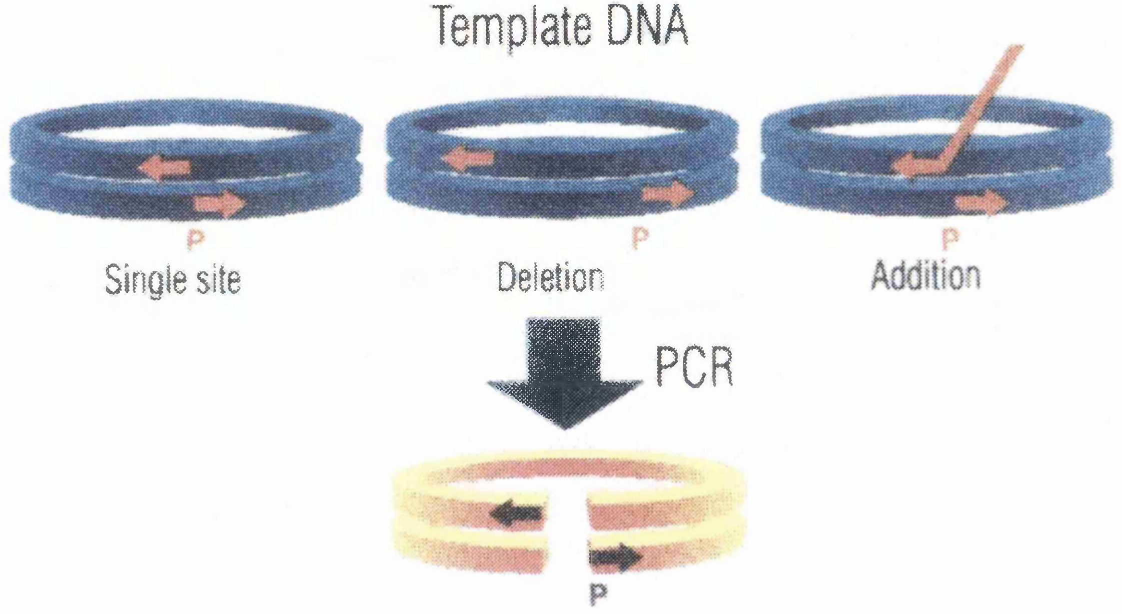


Figure 4.11: Schematic diagram demonstrating the sorts of mutations that can be created using selected primers in an inverse PCR. (Diagram modified from the 1997 Stratagene catalogue).

(4.5) Protein Purification.

To perform biophysical techniques or functional studies on a protein it is necessary to have pure starting material. There exist several techniques for the purification of proteins including engineering "tags" (as mentioned above), ammonium sulphate precipitation and various modes of chromatography.

Ammonium sulphate precipitation is a useful technique for concentrating and partially purifying proteins. In this technique the protein mixture (cell lysate, tissue culture media etc.) is mixed with a high concentration (1.0-5.0 M) of a lyotropic salt, normally ammonium sulphate. The ammonium sulphate masks ionic interactions between proteins and dramatically increases the strength of hydrophobic interactions between surface residues. When the ammonium sulphate concentration is high enough, certain proteins in the mixture will aggregate and, when the aggregates become big enough precipitate out of the solution. Proteins differ significantly in the amount of ammonium sulphate required for precipitation, with the more hydrophobic precipitating at a lower concentration. By exploring various ammonium sulphate concentrations it is possible to determine the minimum and maximum concentrations for the precipitation of the protein of interest and discard all fractions before and after the concentration range containing the required protein (Fulton and Vanderburgh, 1996).

(4.5.1) Ion Exchange Chromatography.

Ion exchange chromatography separates molecules based on differences in their accessible surface charges. The technique can be used for virtually any charged molecule that is soluble in aqueous system, and typically provides high resolving power and high binding capacity. Ion exchange is widely used in the separation of proteins because the relatively mild binding and elution conditions allow high protein recovery with intact biological activity (Fulton and Vanderburgh, 1996).

Ion exchange is based on the ionic attraction between molecules of opposite electric charge. The bonded phase of an ion exchange packing has functional groups that have either a positive charge (anion exchange), used to separate negatively charged target molecules (anions), or a negative charge (cation exchange), used to separate positively charged target molecules (cations). The electrostatic interactions between the opposite charge groups on the surface of the chromatographic packing material and on the binding molecules takes place over a relatively short distance. Because of this, in ion exchange chromatography, molecules are separated based on the number of positive or negative charges accessible on their surfaces (Willard *et al.*,1981; Fulton and Vanderburgh, 1996).

Anion or cation exchange functional groups are classified as either "weak" or "strong". Weak ion exchange groups are titratable, i.e. they gain or lose electrical charge as the pH of the mobile phase changes. It should be noted that the terms strong or weak do not refer to the strength of the binding but only to the effect of pH on the charge of the functional groups. The most common weak anion exchange group is diethylaminoethyl (DEAE). The most common weak cation exchange group is carboxymethyl (CM). Strong ion exchange groups maintain their charge independent of the pH of the mobile phase. The most common strong anion exchange groups are quaternary amines. The most common strong cation exchange groups are sulfonates (Willard *et al.*, 1981; Fulton and Vanderburgh, 1996).

There is substantial energy involved in charge-charge interactions. The laws of physics dictate that the number of positive and negative charges in any given volume must be almost exactly equal. Each charged group in any solution or on a surface has a corresponding counter ion nearby of the opposite charge. The most common counter ions are small salt or buffer molecules (Fulton and Vanderburgh, 1996).

Ion exchange binding occurs when the ionic strength of the mobile phase is reduced to the point that the ionic groups on the sample molecules begin to serve as the counter ions for the charged groups on the stationary phase. This causes the sample molecules to bind to the surface. Elution takes place when the ionic strength of the mobile phase is increased. As this happens, salt molecules displace the bound sample molecules back into the mobile phase. At the same time, salt molecules with the same charge as the bonded phase (called the co-ions) bind to the charge groups on the sample molecules. It should be noted that all the charged species in the solution (including any salts present and even the buffer ions) interact in a complex way either with the bonded phase or the sample molecules during binding and elution. The precise concentrations and chemical nature of all these species have a significant impact on the selectivity of the separation (Willard *et al.*, 1981; Fulton and Vanderburgh, 1996).

Elution methods may also include changes in pH along with (or instead of) increases in ionic strength. The pH can affect the charge of the sample molecules (isoelectric point, Figure 4.12) as well as (in the case of weak ion exchange media) the charge of the bonded phase. Changes in pH can therefore be used to weaken or eliminate charge-charge interactions and cause elution (Fulton and Vanderburgh, 1996).

Isoelectric Point (pI)

pH at which molecule has zero net charge.

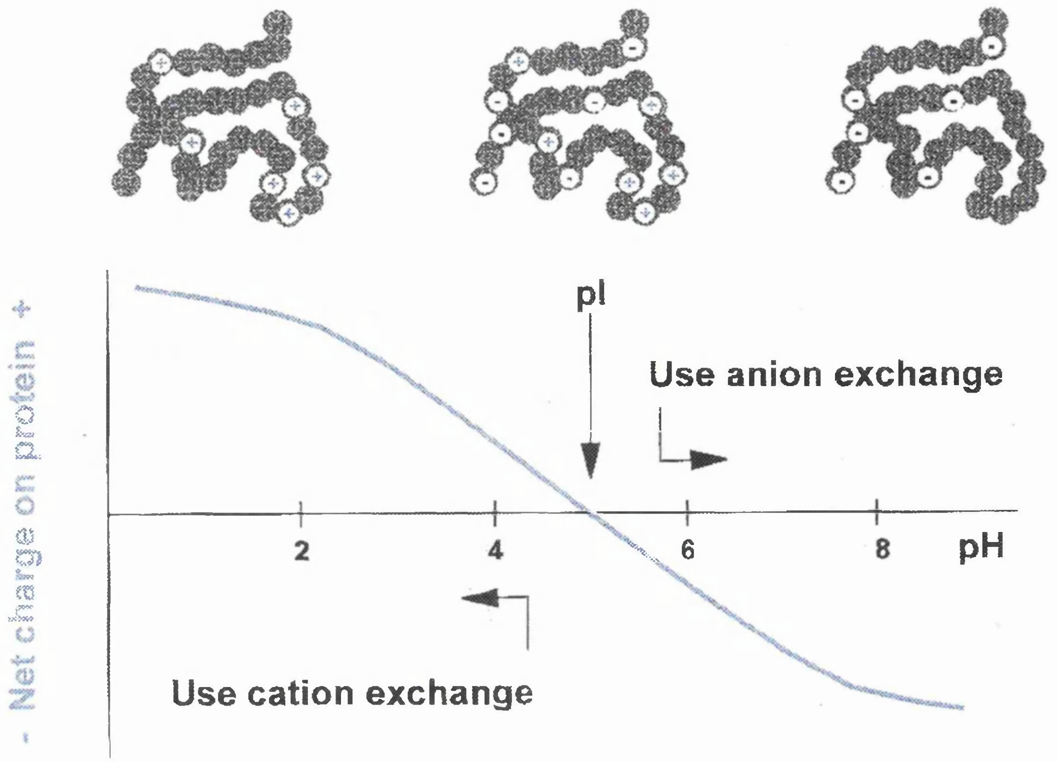


Figure 4.12: Isoelectric points in relation to the ion exchange chromatography method used. An important characteristic relating to the charge of a molecule is its isoelectric point (pI), the pH at which the total number of positive charges equals the total number of negative charges resulting in a net charge of zero. The pI is usually determined by isoelectric focusing electrophoresis. It is important to note however that pI is only of limited use in relation to ion exchange chromatography because only surface accessible residues interact with the bonded phase. (Reproduced from Fulton and Vanderburgh, 1996).

(4.5.2) Hydrophobic Interaction Chromatography.

Hydrophobic interaction chromatography (HIC) separates biomolecules based on the hydrophobic groups on their surfaces. Binding of biomolecules to the mildly hydrophobic surface of a HIC column is induced by the addition of high salt concentrations of the sample and equilibration buffer. Elution is effected by decreasing the salt concentration. Although the mechanism is somewhat different, from a purely functional point of view, HIC can be viewed as a high resolution version of ammonium sulphate precipitation. Any type of protein or large protein is a potential candidate for HIC (Fulton and Vanderburgh, 1996).

The organizational structure of the solvent molecules (water) surrounding the solutes and the binding surface is the driving force for hydrophobic interaction. When hydrophobic areas bind together (as when a protein binds to the surface of an HIC packing), water is effectively released from surrounding the hydrophobic areas, causing a thermodynamically favourable increase in entropy. Salts that induce hydrophobic interactions are those which increase the ordered structure (decrease the entropy) of water. Some salts or other agents that decrease the ordered structure of water (solvents or chaotropes) weaken hydrophobic interactions (Fulton and Vanderburgh, 1996).

Because HIC is driven by entropy, temperature can have a strong effect. Generally, increasing the temperature increases the binding strength. However, temperature can also affect protein conformation, causing the resulting binding effects to be quite complex. Similarly, the effect of pH on hydrophobicity can be complex and hard to predict (Fulton and Vanderburgh, 1996).

Ammonium sulphate, by virtue of its good lyotropic (salting out) properties, high solubility in water, ready availability and low cost, is by far the most common salt used for HIC. Table 4.5 lists both anions and cations in order of increasing lyotropic effect and decreasing chaotropic effect (Fulton and Vanderburgh, 1996).

The salt concentration required for the induced hydrophobic interaction can be modulated by changing the hydrophobicity of the bonded phase. More hydrophobic surfaces require less salt for binding, causing less risk of protein precipitation. However, if the bonded phase is made too non-polar (which is the case with virtually all reversed-phase packings), the protein may not elute when the salt is removed and require organic solvent. In some cases, the protein may actually unfold on the hydrophobic surface, lose its 3-dimensional structure and denature. Determining the proper hydrophobicity of the bonded phase surface for a particular application is quite important (Fulton and Vanderburgh, 1996).

The binding interaction induced by salt between proteins and the HIC packing surface is reversible. Therefore reducing the salt concentration will elute proteins that are bound. As with ion exchange, using a continuous gradient (in this case a decreasing salt concentration) will separate different bound molecules from each other (Fulton and Vanderburgh, 1996).

Some proteins (such as membrane proteins) are so hydrophobic that they will bind to the stationary phase in a low salt mobile phase. In these cases elution can be achieved by adding an increasing gradient of a solvent, chaotropic agent (such as ethylene

Most Lyotropic

PO_4^{3-}	NH_4^+
SO_4^{2-}	Rb^+
CH^3COO^-	K^+
Cl^-	Na^+
Br^-	Cs^+
NO_3^-	Li^+
ClO_4^-	Mg^{2+}
I^-	Ca^{2+}
SCN^-	Ba^{2+}

Most Chaotropic

Table 4.5: Properties of various ions useful in hydrophopic interaction chromatography. The ions are listed in order of increasing lyotropic effect and decreasing chaotropic effect. (Based on Fulton and Vanderburgh, 1996).

glycol, urea, guanidine-HCl or thiocyanate salts) or a detergent. Some times these agents can be used in the presence of ammonium sulphate to aid in solubility of the protein (Fulton and Vanderburgh, 1996).

(4.5.3) Reversed-Phase Chromatography.

Reversed-phase chromatography (RPC) is the most common chromatographic mode for the analysis of any type of molecule, as well as for preparative separations of small molecules, peptides and oligonucleotides. The principles are similar to HIC because (RPC) separates molecules based on differences in hydrophobicity (Fulton and Vanderburgh, 1996).

RPC is not widely used for preparative purification of proteins, except for small, robust proteins (generally below 30K MW). Because both the extremely non-polar stationary phase and the organic solvents used for elution in RPC can cause irreversible denaturation, protein separations based on hydrophobicity are carried out by HIC (Fulton and Vanderburgh, 1996).

(4.5.4) Gel Filtration Chromatography.

Gel filtration chromatography (GFC) also known as size exclusion chromatography or gel permeation chromatography, separates molecules on the basis of their size. This technique is most often used as a final polishing technique since it is the only separation method available to remove aggregated protein species without any chemical or physical change that may cause more aggregates to form. GFC has limited usefulness as a high throughput technique. The separation mechanism requires a slow flow rate, and in most cases, sample load should be only 1-5% of the column bed volume to ensure good results (Fulton and Vanderburgh, 1996).

GFC is different from other modes of chromatography in that one goes to great lengths to prevent any binding interaction at all between the sample molecules and the bonded phase. Gel filtration depends on the fact that within each particle of the stationary phase there is a distribution of pore sizes. For small enough molecules, the pores are so large that the molecules can penetrate all the internal volume of the particle. If the molecules are large enough, the pores are so small that the molecule is completely excluded from the internal volume. Molecules of intermediate sizes will have access via diffusion to a portion of the internal volume but will be excluded by the smaller pores from the rest. In GFC, molecules are separated based on size or molecular weight (Fulton and Vanderburgh, 1996). For structural studies it is important that the protein solution is free from aggregates, so all proteins used in the studies presented here will have been through a gel filtration stage before the experiments are performed.

(4.5.5) Protein Purification Using Monoclonal Antibodies.

If specific monoclonal antibodies have been generated to the protein of interest it is possible to create an affinity column by cross-linking the antibody to an inert support (e.g. cyanogen bromide sepharose). Once the antibody sepharose is packed in a column and washed with a suitable buffer, it is then a simple task of running the material containing the molecule of interest through it (after centrifugation and filtration of the supernatant). After the material has been passed through the column is then washed with the buffer to remove non-bound material.

To elute the protein from the antibody either a change in pH of the wash buffer can be used (which may denature certain proteins) or elute using a 3 M $MgCl_2$ solution in the wash buffer. After elution, the column must be thoroughly washed to remove the $MgCl_2$ otherwise binding will not occur. This technique has been successfully used for the purification of serum derived and recombinant complement factor I (Chapter 7). It may be necessary to run the supernatant through the column two or three times to ensure thorough removal of the protein of interest.

CHAPTER 5

OLIGOMERIC STRUCTURE OF AMIC

(5.1) Introduction.

Pseudomonas aeruginosa is a ubiquitous gram negative rod-shaped bacterium that is an important opportunistic pathogen in man and other animals (Singleton and Sainsbury, 1987; Jawetz *et al.*, 1987; Koch and Høiby, 1992). It can be isolated from infected burns, urinary tract infections, and the lungs of patients with cystic fibrosis. It can occasionally be pathogenic in stressed plants. *P. aeruginosa* can utilise short chain aliphatic amides such as acetamide $CH_3.CO.NH_2$ as sole carbon and nitrogen sources, and these are hydrolysed to ammonia and acetic acid. The enzyme system is induced by the presence of amides (Kelly and Clarke, 1962; Stanier *et al.* 1966). The amidase operon consists of 5 genes, namely *amiE, amiB, amiC, amiR,* and *amiS* in that order. AmiC is a soluble cytoplasmic protein that functions as an amide sensor and negative regulator of the amidase operon (Figure 5.1). AmiC controls the activity of the transcription antitermination factor AmiR, which in turn regulates expression of the amidase enzyme system. *AmiE* is the gene which corresponds to the amidase enzyme, and *amiB* and *amiS* appear to form a membrane transport system for the importation of amide into the bacteria (Drew and Wilson, 1992; Wilson *et al.*, 1995).

The combination of secondary structure predictions and fold recognition analyses indicated that, despite only 17% amino acid sequence identity, AmiC had the same protein fold as the leucine-isoleucine-valine binding protein (LivJ) of *Escherichia coli* (Sack *et al.*, 1989a; Wilson *et al.*, 1993). LivJ corresponds to the Cluster 4 subclass of periplasmic binding proteins (Tam and Saier, 1993). The prediction was confirmed by the crystal structure of AmiC bound to its substrate acetamide (Pearl *et al.*, 1994).

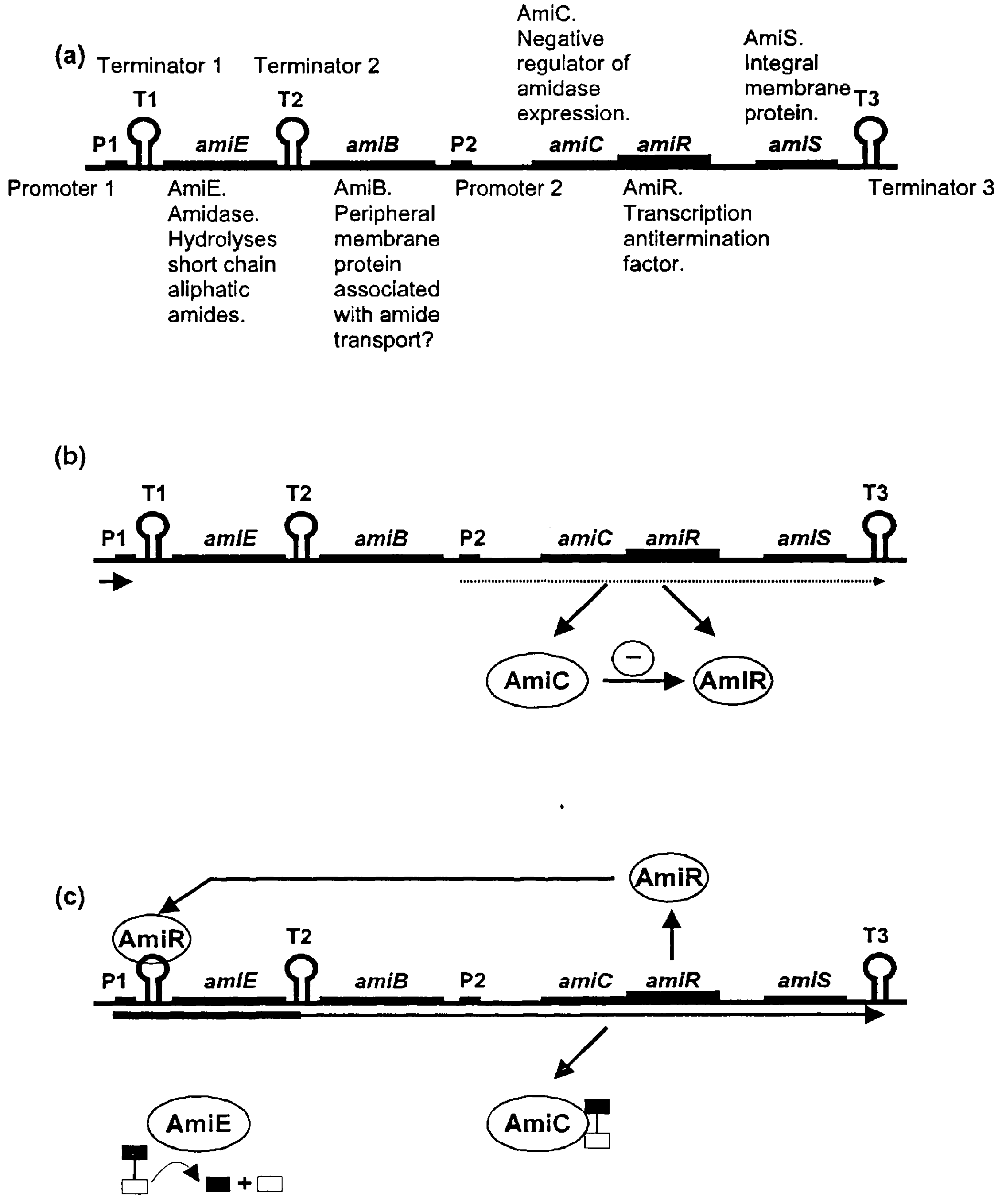


Figure 5.1: Genetic organization and control of the amidase operon. (a) Features and descriptions of elements in the amidase operon. (b) In the uninduced state, AmiC inhibits AmiR and transcription of the *amiE* gene is prematurely terminated by formation of a stem-loop structure (T1) in the nascent mRNA, between the promoter (P1) and the start of the *amiE* structural gene. (c) Binding of amide inducers (linked black and white boxes) to AmiC relives inhibition of AmiR which prevents termination at the stem-loop and allows transcription of the whole operon. A weak promoter (P2) provides a low level of transcription of *amiC*, *amiR* and *amiS* genes in the induced state (Pearl *et al.*, 1994).

The similarity of the AmiC structure to that of periplasmic binding proteins is of interest in that these proteins form a large family of related structures that are involved with the transport of small molecules into bacteria (Tam and Saier, 1993). A total of 8 different prokaryotic subclasses that bind to sugars, amino acids and anions have been identified, and crystal structures have been determined for six of these (Table 5.1). Nonetheless AmiC exhibits distinct functional properties in that it controls AmiR in response to a signal from acetamide, while the periplasmic binding proteins transport small molecules within the inner bacterial membrane. A similar relationship with LivJ has also been identified for the extracellular domain of the eukaryotic protein glutamate receptor, which is involved in neurotransmitter activity (O'Hara *et al.,* 1993; Stern-Bach *et al.,* 1994). AmiC is constructed from two nonequivalent α-helix/β-sheet domains joined by 3 polypeptide links which flank a ligand-binding site in a large cleft between them (Figure 5.2). Interestingly, the binding site cleft in LivJ is opened by a domain movement of approximately 35° compared to that in AmiC (Figure 5.2). In the AmiC-acetamide crystal structure, the cleft is substantially closed. The AmiC amide binding site is extremely specific for acetamide with a dissociation constant of 3.7 μM. Butyramide $CH_3.CH_2.CH_2.CO.NH_2$ is an anti-inducer of AmiC and has a 100-fold bigger dissociation constant. It is possible that AmiC-acetamide and AmiC-butyramide may possess alternative conformations.

Small angle X-ray and neutron scattering are powerful low resolution methods for studies of the arrangement of domains in multidomain proteins and their degree of oligomerisation (Perkins, 1988). They have advantages in that the data are obtained in solution. The utility of the methods has been much improved by the development of

Periplasmic binding protein	Brookhaven code	R_G (nm)	Reference
Cluster 1 (molecular weight 40,600)			
Maltodextrin binding protein	(1omp)	2.48	Sharff *et al.* (1992)
Maltodextrin binding protein + maltose	(2mbp)	2.39	Spurlino *et al.* (1991)
Maltodextrin binding protein + β-cyclodextrin	(1dmb)	2.41	Sharff *et al.* (1992)
Maltodextrin binding protein mutant + maltose	(1mpb)	2.43	
Maltodextrin binding protein mutant + maltose	(1mbc)	2.43	
Maltodextrin binding protein mutant + maltose	(1mbd)	2.34	
Maltodextrin binding protein mutant + maltose	(1mdp)	2.31, 2.30	Sharff *et al.* (1995)
Maltodextrin binding protein mutant + maltose	(1mdq)	2.31	Sharff *et al.* (1995)
Cluster 2 (molecular weight 33,000)			
Arabinose binding protein + arabinose	(1abe)	2.26	
Arabinose binding protein mutant + arabinose	(6abp)	2.24	
Arabinose binding protein mutant + arabinose	(1bap)	2.24	Vermersh *et al.* (1990)
Arabinose binding protein + fucose	(1abf)	2.24	Quicho *et al.* (1989)
Arabinose binding protein mutant + fucose	(1apb)	2.23	Vermersh *et al.* (1990)
Arabinose binding protein mutant + fucose	(7abp)	2.25	
Arabinose binding protein + galactose	(5abp)	2.26	Quicho *et al.* (1989)
Arabinose binding protein mutant + galactose	(9abp)	2.23	Vermersh *et al.* (1990)
Arabinose binding protein mutant + galactose	(8abp)	2.24	
Galactose binding protein	(2gbp)	2.31	
Galactose binding protein + glucose	(3gbp)	2.28	
Galactose binding protein + galactose	(1glg)	2.30	
Cluster 3 (molecular weight 26,100)			
Histidine binding protein + histidine	(1hpb)	2.02	Oh *et al.* (1994)
Histidine binding protein + histidine	(1hsl)	1.99, 1.99	Yao *et al.* (1994)
Lys-Arg-Orn binding protein	(2lao)	2.12	Oh *et al.* (1993)
Lys-Arg-Orn binding protein + lysine	(1lst)	1.99	Oh *et al.* (1993)
Lys-Arg-Orn binding protein + arginine	(1laf)	1.99	Oh *et al.* (1993)
Lys-Arg-Orn binding protein + ornithine	(1lah)	1.97	Oh *et al.* (1993)
Lys-Arg-Orn binding protein + histidine	(1lag)	1.98	Oh *et al.* (1993)
Cluster 4 (molecular weight 36,800 and 41,200)			
Leu-Ile-Val binding protein LivJ	(2liv)	2.43	Sack *et al.* (1989a)
Leucine binding protein	(2lbp)	2.44	Sack *et al.* (1989b)
Acetamide binding protein AmiC + acetamide	(1pea)	2.23	Pearl *et al.* (1994)
Acetamide binding protein AmiC + butyramide	( )	2.25	O'Hara *et al.* (1998)
Cluster 5 (molecular weight 59,100)			
Oligopeptide binding protein + tetrapeptide	(1ola)	2.54	Tame *et al.* (1994)
Oligopeptide binding protein + tripeptide	(1olb)	2.56	Tame *et al.* (1994)
Cluster 6 (molecular weight 34,300)			
Phosphate binding protein + phosphate	(1abh)	2.23	Luecke and Quiocho (1990)
Phosphate binding protein mutant + phosphate	(1pbp)	2.23	
Sulphate binding protein + sulphate	(1sbp)	2.15	

Table 5.1: Radius of gyration analysis of 36 representative crystal structures of periplasmic binding proteins with and without ligands from the Brookhaven Protein Database. Cluster classification from Tam and Saier (1993).

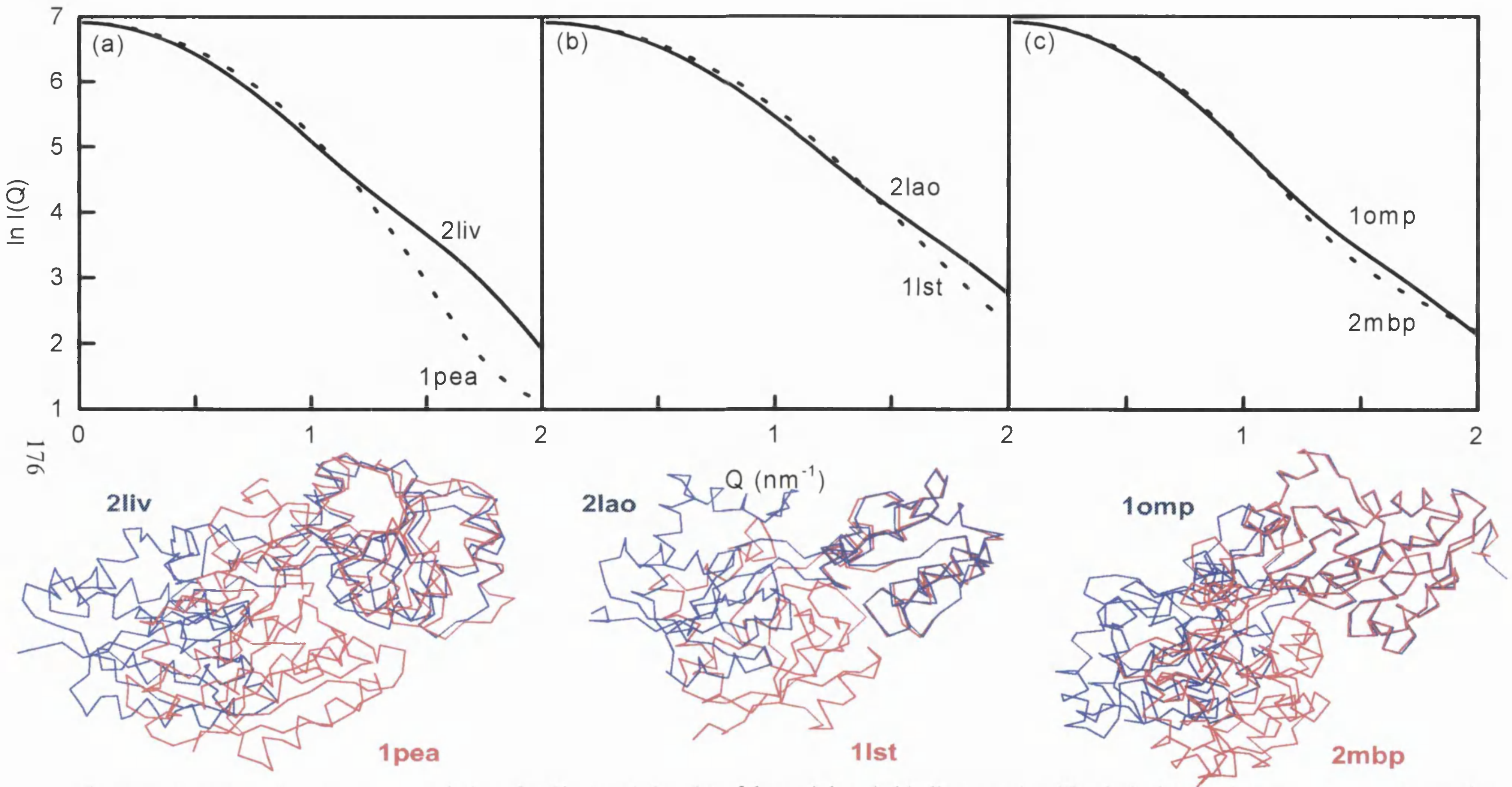


Figure 5.2: X-ray scattering curve simulations for Clusters 4, 3 and 1 of the periplasmic binding proteins. The dashed scattering curves correspond to the closed conformations. The α-carbon coordinates of the two proteins are shown with the C-terminal domains superimposed, and the closed conformations are represented in bold outline. The views of theα-carbon traces are shown to maximise the domain movement seen between the open and closed forms. (a) AmiC (1pea) and LivJ (2liv) in Cluster 4 were represented by 623 and 559 spheres respectively of radius 0.220 nm. (b) Lysine-argine-ornithine binding protein (1lst and 2lao) in Cluster 3 was represented by 398 spheres of radius 0.220 nm. (c) Maltodextrin binding protein (1omp and 2mbp) in Cluster 1 were represented by 94 and 96 spheres respectively of radius 0.410 and 0.405 nm.

calibrated procedures for the calculation of scattering curves from known crystal structures (Smith *et al.*, 1990; Perkins *et al.*, 1993). Automated scattering curve fit procedures constrained by known atomic structures can now be used to assess the unknown structure of a multidomain protein (Mayans *et al.* 1995; Beavil *et al.*, 1995; Boehm *et al.*, 1996). The previous application of X-ray scattering to periplasmic binding proteins showed that L-arabinose binding protein was monomeric, and that on the addition of L-arabinose its R_G value of 2.12 nm decreased by 0.094 ± 0.033 nm (Newcomer *et al.*, 1981). This decrease corresponded to a rotation of the two domains closer to one another by $18^\circ \pm 4^\circ$. Other periplasmic binding proteins showed a 52° domain rotation between ligated and unligated lysine-arginine-ornithine binding protein (Oh *et al.*, 1993), and a 35° domain rotation between ligated and unligated maltodextrin binding protein (Sharff *et al.*, 1992) (Figure 5.2). Here, X-ray and neutron scattering methods are applied to determine the solution structure of AmiC. Unlike the classical periplasmic binding proteins which are monomeric, we show that AmiC exists in monomeric and trimeric forms, the proportions of which depend on the presence of acetamide or butyramide. We assess whether a ligand-dependent conformational change may occur, and describe how automated curve fit methods can be applied to interpret the scattering curves in terms of a structure for trimeric AmiC.

(5.2) Materials and Methods.

(5.2.1) Expression and Purification of AmiC for Solution Scattering.

The expression system consisted of a 1.3×10^3 base pair fragment of the amidase system containing the *amiC* open reading frame, cloned into a broad host range vector pMMB66HE, and transformed into a *P. aeruginosa* amidase deletion strain as described and characterized by Wilson and Drew (1991). The bacteria were fermented in 8 l of modified Oxoid no.2 broth and protein expression was started by the addition of 3mM isopropyl-ß-D-thiogalactopyranoside when $O.D_{670nm}$ >1.0. Growth continued until $O.D_{670nm}$ >4.0. Cells were harvested by low speed centrifugation (5,000 g), and lysed immediately by sonication in AmiC buffer (20 mM Tris HCl, pH 8.0, 1 mM dithiothreitol, 1 mM EDTA, 1 mM phenylmethylsulphonylfluoride) (Wilson *et al.,* 1991). The supernatant after sonication was clarified by centrifugation at 25,000 g for 30 min, and AmiC was precipitated using 40-60% saturated $(NH_4)_2SO_4$ (Chapter 4) The AmiC fractions were pooled, re-suspended in AmiC buffer and dialysed overnight to remove $(NH_4)_2SO_4$. AmiC was purified further by ion exchange (Chapter 4) (2.6 × 10 cm Q Sepharose, Pharmacia) when it was eluted in the range 450-550 mM NaCl using a 0-1M NaCl gradient (flow rate 6 ml/min; 6 ml fractions). These fractions were pooled, made up to 1.2 M $(NH_4)_2SO_4$, loaded onto a phenyl-Sepharose hydrophobic interaction column (Chapter 4) (1.6 × 10 cm phenyl-Sepharose, Pharmacia) and eluted using a 0-1 M $(NH_4)_2SO_4$ gradient (flow rate 4 ml/min; 2 ml fractions). The pooled AmiC fractions were dialysed for several days against AmiC buffer containing 10 mM butyramide to remove acetamide. AmiC was then concentrated in an Amicon pressure cell using a YM10 membrane and purified by gel filtration (Chapter 4) (1.0 × 30 cm Superdex 200, Pharmacia) as a single peak to give concentrations of up to 17 mg/ml (flow rate 0.8 ml/min; 1.25 ml fractions) (Figures 5.3 and 5.4). The absorption coefficient of AmiC (1%, 1 cm, 280 nm) was calculated as 13.6 from its amino acid composition (Perkins, 1986).

AmiC samples were stored at 6°C and used within a few days for scattering experiments. For X-ray scattering and neutron scattering in H_2O buffers, the AmiC-

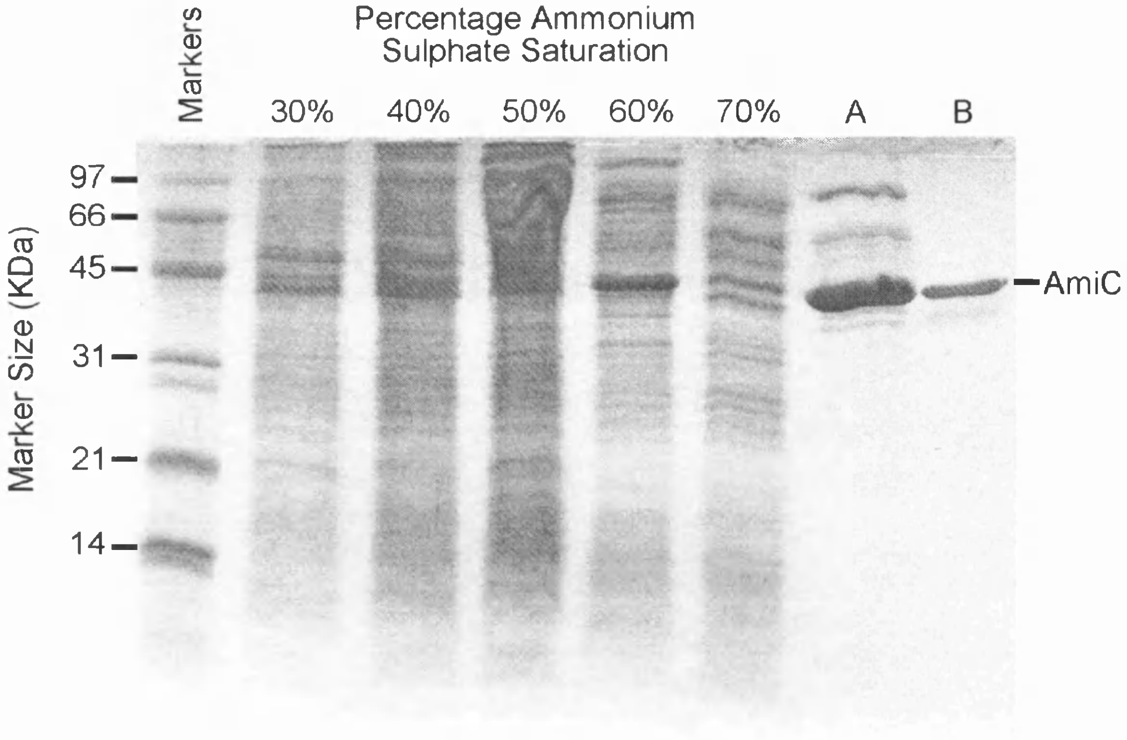


Figure 5.3: 10% reducing polyacrylamide gel showing the purification of AmiC. The 40-60% ammonium sulphate precipitate fractions were pooled, dialysed, and then loaded on an ion exchange column. Ion exchange purified protein (lane A) was then further purified with hydrophobic interaction chromatography and by gel filtration (lane B).

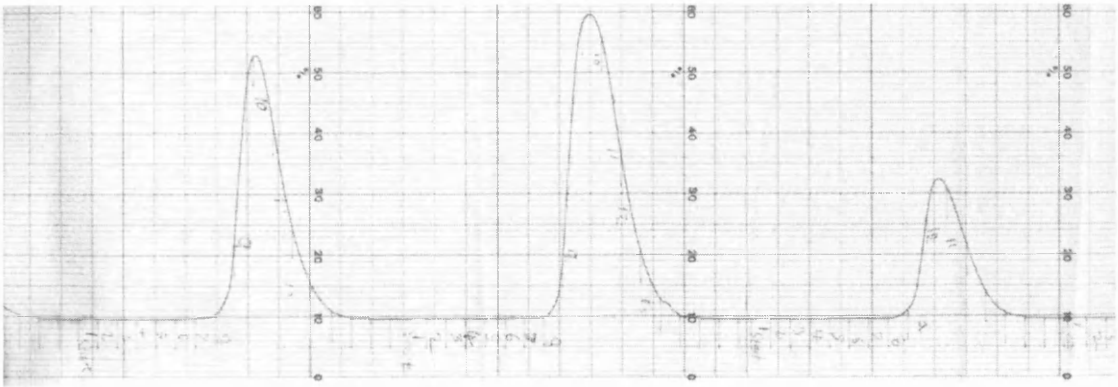


Figure 5.4: Gel filtration profiles of the purification of AmiC. In three separate runs, AmiC eluted as a single symmetric peak at the same elution volume. Top of chart = 1 absorbance unit at 280 nm.

butyramide samples were used as prepared above, and AmiC-acetamide was generated by adding 10 mM acetamide immediately prior to data collection to displace butyramide. For neutron scattering in ^2H_2O buffers, the AmiC-butyramide samples were dialysed for 36 h with four buffer changes into AmiC buffer prepared in ^2H_2O and containing either 10 mM acetamide or butyramide. Alternatively, AmiC-butyramide was dialysed into its ^2H_2O buffer containing 10 mM butyramide, and AmiC-acetamide was generated by adding 10 mM acetamide in AmiC buffer in ^2H_2O immediately prior to data collection.

(5.2.2) X-ray and Neutron Scattering Data Collection.

X-ray scattering curves were obtained in two beam sessions using a camera with a quadrant detector at Station 2.1 at the Synchrotron Radiation Source, Daresbury (Towns-Andrews *et al.*, 1989; Worgan *et al.*, 1990). Sample-detector distances of 3.14 m or 3.17 m were used, with beam currents of 122-173 mA and a storage ring energy of 2.0 GeV. This resulted in a usable Q range of 0.1 to 2.3 nm^{-1} ($Q = 4 \pi \sin \theta / \lambda$; scattering angle = 2θ ; wavelength = λ). Data acquisition times were 10 min, obtained as 10 time frames of 1 min each as a control for radiation damage. Other details of data collection and analyses are described elsewhere (e.g. Beavil *et al.*, 1995; Chapter 2). Sample temperatures were set at 15°C.

Neutron scattering data were obtained in one session on Instrument D11 at the Institut Laue-Langevin, Grenoble (Lindner *et al.*, 1992). Sample-detector distances of 2.00 m and 5.00 m were used. With the monochromator set for λ of 1.00 nm, and using a 64 × 64 cm detector, the two detector settings resulted in a usable Q range of 0.06 to 1.1 nm^{-1}. Using a rectangular beam aperture of 7 × 10 mm, data acquisition times were typically 5 min in ^2H_2O buffers and 30 min in H_2O buffers. Samples were measured at 15°C in rectangular quartz Hellma cuvettes of path length 2 mm for samples in ^2H_2O buffers and 1 mm for samples in H_2O buffers, and absorbances at 280 nm for AmiC concentrations were measured directly in the same cells. Sample and buffer transmissions were measured relative to an empty cell transmission for use in data reduction. Data were processed using standard Grenoble software (RNILS, SPOLLY, RGUIM and RPLOT; Ghosh, 1989). A cadmium run for electronic and neutron background was first subtracted from each scattering curve. The buffer background run was subtracted from that of the sample run, and the result was normalized for the detector response by using a water run from which an empty cell background, corrected for the transmission of water, had been subtracted.

Neutron scattering data were also obtained in one session on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, U.K. (Heenan and King, 1993). The moderated pulsed neutron beam was derived from a tantalum target after proton bombardment at 50 Hz (proton beam current of 171 μA). Based on a fixed sample-detector distance of 4.3 m, the usable Q range was 0.1 to 2.0 nm^{-1}. The data acquisition time was 1 h at a sample temperature of 15°C. Other details of data collection and analyses are as described in Chapter 2.

(5.2.3) Guinier and Distance Distribution Function Analyses of Reduced Scattering 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 and 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 weights M_r of the proteins when referenced against a suitable standard (Kratky, 1963; Jacrot and Zaccai, 1981; Wignall and Bates, 1987). 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 in reciprocal space I(Q) into that in real space P(r) was carried out using GNOM (Svergun *et al.*, 1988; Semenyuk and Svergun, 1991; Svergun, 1992).

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

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. For this, the X-ray I(Q) curve in the range between 0.3-2.0 nm^{-1} contained 345 data points, which were reduced to 255 points by GNOM for the transformation. The LOQ neutron I(Q) contained 76 data points in the Q range between 0.1-2.1 nm^{-1}. GNOM employs a regularisation procedure with an automatic choice of the transformation parameter α to stabilise the P(r) calculation

(Svergun, 1992). The P(r) curve contains 61 points. A range of maximum assumed dimensions D_{max} 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.

(5.2.4) Automated Procedure for Debye Sphere Modelling of AmiC.

The X-ray and neutron scattering curves were modelled using small single-density spheres to represent the AmiC 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 (Perkins and Weiss, 1983):

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

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

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. The method has been calibrated with known crystal coordinates (Smith *et al.,* 1990; Perkins *et al.,* 1993; Ashton *et al.*, 1997; Chapter 3).

The monomeric AmiC-acetamide coordinates (Brookhaven code: 1pea) formed the asymmetric unit in space group $P4_22_12$, and were used for all calculations. The coordinates were converted to spheres by placing all residue atoms within a three-dimensional grid of cubes of side 0.457 nm. A cube was included in the model if it contained sufficient atoms above a specified cut-off such that the total volume of the 580 cubes equalled that of the dry protein of 55.0 nm^3 calculated from the sequence (accession code: AMIC_PEASE; P27017) (Chothia, 1975; Perkins, 1986). As AmiC contains 384 residues, while only 369 residues were visible in the crystal structure for reason of crystallographic disorder at the N- and C-termini, this procedure compensated for the 4% smaller volume present in the crystal structure. X-ray curve fits were based on a rescaled hydrated model, whose volume is the sum of the dry model and that of a hydration shell of 0.3 g H_2O/g AmiC. The latter corresponds to an electrostricted volume of 0.0245 nm^3 per bound H_2O (Perkins, 1986). The rescaled cube coordinates have sides of 0.496 nm and correspond to spheres of radius 0.308 nm. The sphere sizes are much less than the nominal resolution of $2\pi/Q_{max}$ of the scattering curves. No corrections were applied for X-ray wavelength spread or beam divergence as these are considered to be negligible. For both LOQ and D11 data, a 16% spread in λ for a nominal λ of 1.0 nm and a beam divergence of 0.016 radians were used to correct the calculated neutron scattering curve for the reasons discussed in Mayans *et al.* (1995). Neutron curve fits were used after X-ray curve fitting to confirm that possible solute-solvent contrast effects were not significant. The R_G value of the model was calculated from the Guinier fit of the calculated curve in the same Q range used for experimental data. The quality of the curve fit was defined using an R-factor $R_{2.0}$ to measure the agreement between the experimental and calculated X-ray curves in the Q range between 0.1 to 2.0 nm^{-1} (Smith *et al.*, 1990; Beavil *et al.*, 1995). For a given set of models and curve fits, the R_G and $R_{2.0}$ values were imported into a spreadsheet for filtering and sorting to identify the best fit. Models for oligomers were not retained if they contained less than 95% of the required total of spheres in order to exclude models with significant steric overlap between the monomers.

In application to the comparative simulations of Figure 5.2, two changes were made: (I) The Brookhaven database files themselves were used directly in the simulations without correction for residues not observed in the electron density maps. (ii) As only α-carbon coordinates were reported in the 2mbp structure, only the α-carbon coordinates in the 1omp structure were used for reason of consistency.

In application to automated X-ray curve-fitting analyses, INSIGHT II 95.0 (Biosym/MSI, San Diego, USA) was used for all manipulations. Three approaches were developed: (I) A symmetric trimer was considered by orientating arbitrarily three monomers parallel to each other along their long axes such that they were related by 120° rotations about a three-fold Z-axis of symmetry. Starting from a model in which the centres of the three monomers were close to the central three-fold axis of symmetry and the monomers were sterically overlapped, further models for curve calculations were generated using INSIGHT II macros by moving the monomers outwards from the central Z-axis in 0.2 nm steps in a total range of 4 nm. (ii) Mixtures of monomeric, dimeric, trimeric and tetrameric AmiC were considered by calculating the scattering curves for each of the crystallographic monomer, dimer, trimer and tetramer. The putative dimer was generated using the symmetry-related transformation x, y, z to y, x, -z by application of the crystallographic dyad at x, x, 1/2 (Pearl *et al.*, 1994). A putative tetramer was then generated by application of the crystallographic dyad at 1/2, 1/2, z to this dimer. The putative trimer was formed by deleting any one of the four monomers in the tetramer. All combinations of these four curves in 1% steps were summed for fits with the experimental data. (iii) A putative asymmetric trimer model was considered using the crystallographic dimer and monomer. These were aligned manually so that their centres were close to each other without steric overlap. Cartesian axes were defined by reference to the centre of mass of the AmiC monomer. The monomer was then moved -6 nm along its major translational Z-axis and -3 nm along the X- and Y-axes. The two structures were then translated in +0.2 nm steps relative to each other in the X, Y and Z directions for distances of up to 12 nm using INSIGHT II macros, and the scattering curve was then calculated from each model for comparison with experimental data.

(5.2.5) Sedimentation Equilibrium Analyses of AmiC.

Preliminary experiments using sedimentation equilibrium were performed to determine, the apparent weight-average molecular mass $M_{w,app}$ of AmiC at 1, 2 and 4 mg/ml in three AmiC buffer regimes, starting with 10 mM butyramide, to which was added 150 mM $(NH_4)_2SO_4$, and then 10 mM acetamide. Buffer densities at 15°C were determined to be 1.000179 g/ml, 1.010513 g/ml and 1.011979 g/ml respectively, using an Anton Paar DMA OC2 precision densimeter. Runs at 15°C using a Beckman XL-A analytical ultracentrifuge were performed at three rotor speeds (7,000, 11,000 and 15,000 r.p.m.) using an An60Ti rotor. Samples were housed in multichannel centrepieces (column height of 1.5 mm) to permit the multiplexing of 9 samples in one equilibrium experiment. The equilibrium solute distribution was observed with the scanning absorbance optics set to 300 nm, and equilibrium was determined by the perfect overlay of traces acquired 2 h apart. After the last equilibrium had been reached at 15,000 r.p.m., the rotor was accelerated to 40,000 r.p.m. in order to obtain a true optical baseline E free from solute. Data analyses using the Marquardt-Levenberg algorithm were performed assuming a single ideal solute to obtain the whole-cell $M_{w,app}$ using the Beckman IDEAL1 model (Ralston, 1993; McRorie and Voelker, 1993):

$$A_x = \exp [\ \ln A_o + M_{w,app} H (r_x^2 - r_o^2)\ ] + E$$

where A_x is the absorbance as a function of the radius r_x in cm, A_o is the absorbance at a reference point r_o, E is the optical baseline, and H is given by $(\omega^2/2RT)(1 - \bar{v}\rho)$ where ω is the rotor speed in radians/sec, R is the gas constant, T is the absolute temperature, $\bar{v}$ is the partial specific volume of AmiC (0.732 ml/g from Perkins, 1986), and ρ is the density of the solution (g/ml). The quality of the fit was evaluated from the residuals, obtained by subtracting the calculated best fit from the experimental data. To analyse the self-association of AmiC, data analyses were performed using the four-exponent Beckman ASSOC4 model (McRorie and Voelker, 1993):

$$A_x = \exp [\ \ln A_o + M_{w,app} H (r_x^2 - r_o^2)\ ] + \exp [\ n_2 \ln A_o + \ln K_{a2} + n_2 M_{w,app} H (r_x^2 - r_o^2)\ ]$$
$$+ \exp\ [\ n_3 \ln A_o + \ln K_{a3} + n_3 M_{w,app} H (r_x^2 - r_o^2)\ ] + \exp [n_4 \ln A_o + \ln K_{a4} + n_4 M_{w,app}$$
$$H (r_x^2 - r_o^2)\ ] + E$$

where n_2, n_3 and n_4 were set as 2, 3 and 4 for dimeric, trimeric and tetrameric AmiC, and K_{a2}, K_{a3} and K_{a4} correspond to the association constants for the monomer- n-mer equilibrium defined by $K_a = c_{n\text{-}mer}/(c_{monomer})^n$. Note that the association constants are defined on an absorbance concentration scale and not on a molar concentration scale, and conversion is required (equation 11 of McRorie and Voelker, 1993) : $K_{1\text{-}3,\mathrm{conc}}=\mathrm{c}_3/\mathrm{c}_1{}^3=K_{1\text{-}3,\mathrm{abs}}\varepsilon^2 l^2/3$).

(5.3) Results and Discussion.

(5.3.1) AmiC Oligomers by Synchrotron X-ray Scattering.

X-ray scattering data for AmiC in AmiC buffer containing 10 mM butyramide or 10 mM acetamide as appropriate (Methods) were obtained in the concentration range between 0.4 -16.4 mg/ml. These are denoted as AmiC-buytramide and AmiC-acetamide respectively. Figure 5.5(a) shows that linear Guinier plots in satisfactory $Q.R_G$ ranges were obtained. Guinier analyses of the 10 time-frames used during data acquisition indicated the absence of radiation damage effects that are commonly seen with other proteins. However pronounced concentration effects were observed at above 10 mg/ml, when the Guinier plots exhibited diminished intensities at the lowest Q values. These are typical of interparticle interference effects when each protein molecule senses the presence of its neighbours (Guinier and Fournet, 1955). At these higher concentrations, a reduced $Q.R_G$ range of fit corresponding to 0.35-0.5 nm^{-1} was required in order to obtain linear Guinier analyses.

The Guinier analyses showed that, at concentrations below 5 mg/ml, both the R_G and I(0)/c values decreased with decrease in AmiC concentration (Figures 5.6a and 5.6b). This is typical of the dissociation of an oligomeric protein. The R_G and I(0)/c values were consistently higher for AmiC-acetamide when compared with AmiC-butyramide, in particular at AmiC concentrations below 2 mg/ml, and again at above 10 mg/ml. This suggested that the presence of acetamide induced a higher degree of oligomer formation in AmiC compared to butyramide. In contrast, L-arabinose-binding protein behaved as a monomeric protein in the concentration range of 6-36 mg/ml protein (Newcomer *et al.*, 1981).

To determine whether a conformational change could be detected in AmiC when the ligand was changed from butyramide to acetamide, the R_G values for the two forms were compared for curves with identical I(0)/c values (i.e. similar degrees of oligomerisation). For an I(0)/c value of 9.3, the full dashed lines in Figure 5.6(a) showed that the R_G values of AmiC-butyramide at 8-13 mg/ml were the same at 3.35 nm (within a range of 0.05 nm) to those for AmiC-acetamide at 3-5 mg/ml. This showed that no large conformational change had occurred within these limits. The errors in R_G values are larger at low concentrations, and it was not possible to consider this question for AmiC below 2 mg/ml. If the binding of acetamide had induced a rotational closure of the cleft in AmiC, the R_G value would be expected to be smaller by about 0.1 nm by analogy with L-arabinose-binding protein (Newcomer *et al.*, 1981).

Changes in AmiC oligomerisation were also visible in the full scattering curves I(Q) out to Q = 2 nm^{-1} (Figure 5.7a), in which a submaximum at Q of 1.2 nm^{-1} at high AmiC concentration disappeared at low AmiC concentration. Satisfactory distance distribution functions P(r) were calculated from I(Q) curves on the basis of a presumed maximum dimension D_{max} of 12 nm (Figure 5.7b). The P(r) curves offered an alternative determination of R_G and I(0)/c and these corroborated the Guinier analyses of Figures 5.6(a) and 5.6(b) (data not shown). The P(r) curves also demonstrated a concentration dependence which is larger for AmiC-butyramide. At all concentrations, the maximum dimension L of AmiC is close to 9 nm, and shows that the different oligomers are similar

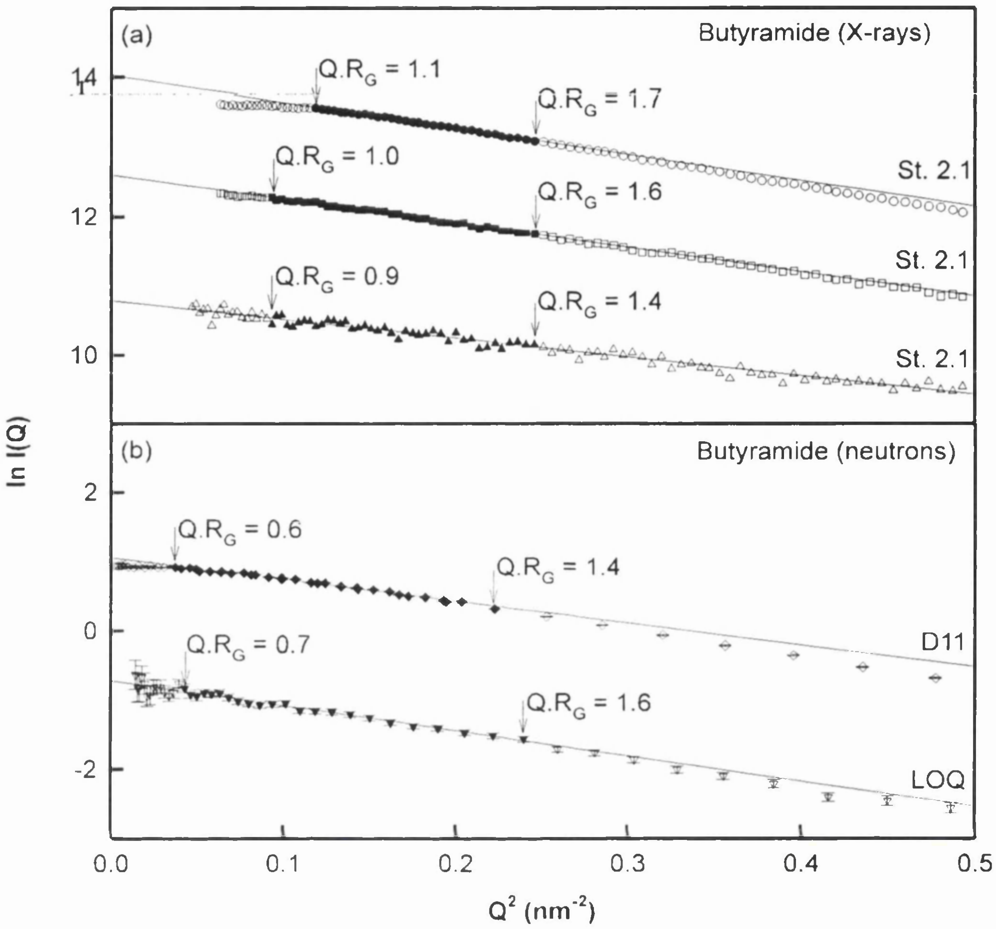


Figure 5.5: Guinier R_G plots for AmiC-butyramide. The filled symbols between the Q.R_G values as arrowed denote the range used to determine I(0) and R_G. Statistical error bars are only shown when these are large enough to be visible. (a) Dilution series studied by synchrotron X-ray scattering for AmiC-butyramide at concentrations of 13.7 mg/ml (○), 4.1 mg/ml (□), and 1.0 mg/ml (Δ). The Q range for Guinier fits was 0.3 - 0.5 nm^{-1} for the 1 mg/ml curve and 0.35-0.5 nm^{-1} for the 13.7 mg/ml curve. (b) Neutron scattering data for AmiC-butyramide concentrations of 5.1 mg/ml used for D11 (◇) and 2.6 mg/ml used for LOQ (∇).

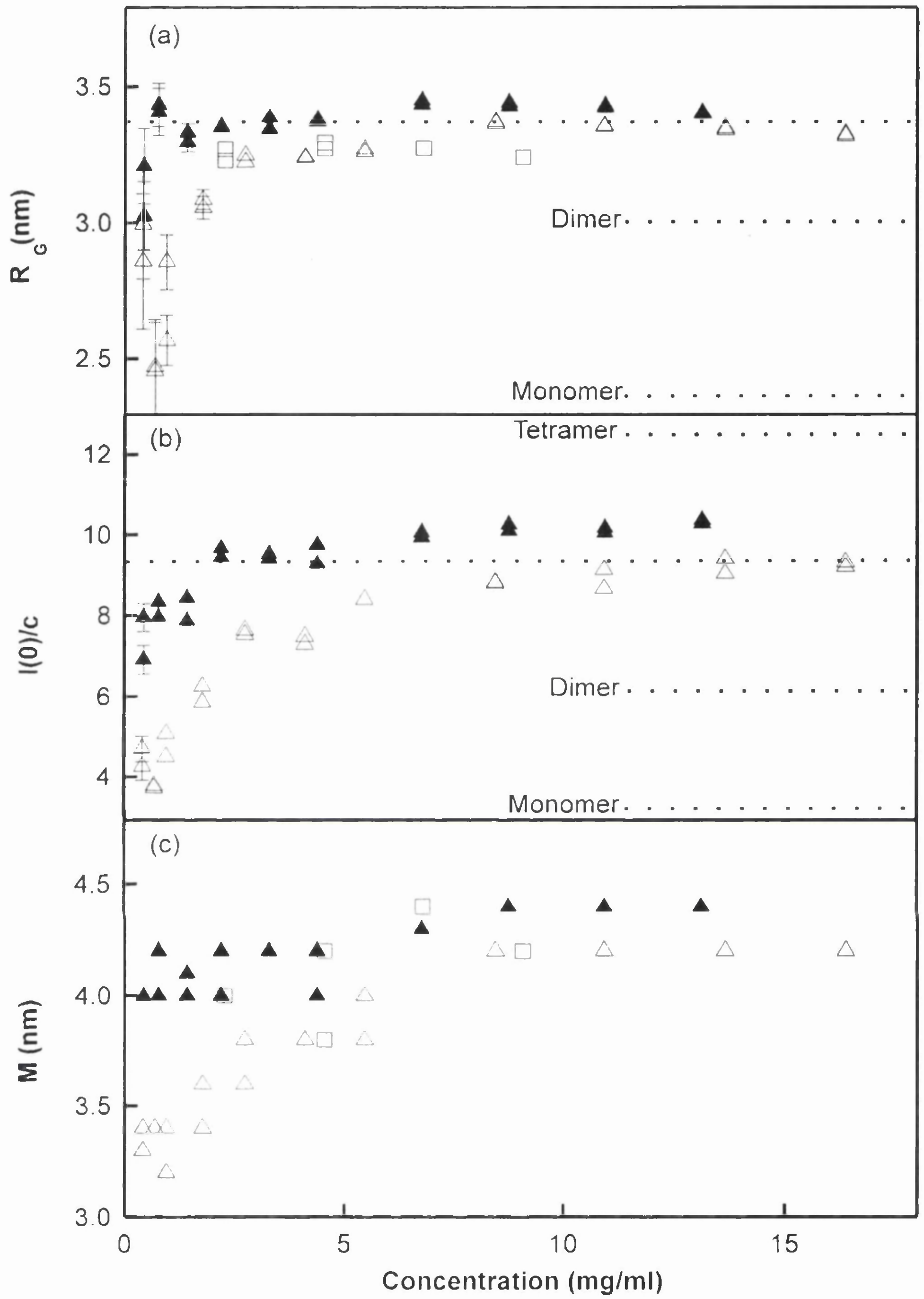


Figure 5.6: Concentration dependence of the AmiC X-ray Guinier and P(r) parameters. Statistical error bars are only shown when these are large enough to be visible. (a) Those for the R_G values for AmiC-butyramide in two different beam-time sessions (□ and Δ) and for AmiC-acetamide (▲) is summarised. The dashed lines denote the R_G values of 1, 2 and 3 subunits of AmiC from Table 2 for comparison. (b) The corresponding I(0)/c values for AmiC-butyramide and AmiC-acetamide are shown, using the same symbols as in (a). The dashed lines denote the I(0)/c values corresponding to the molecular weights for 1, 2, 3 and 4 subunits of AmiC. (c) The most frequently occurring distance M in P(r) curves for AmiC-butyramide and AmiC-acetamide are shown, also using the same symbols as in (a).

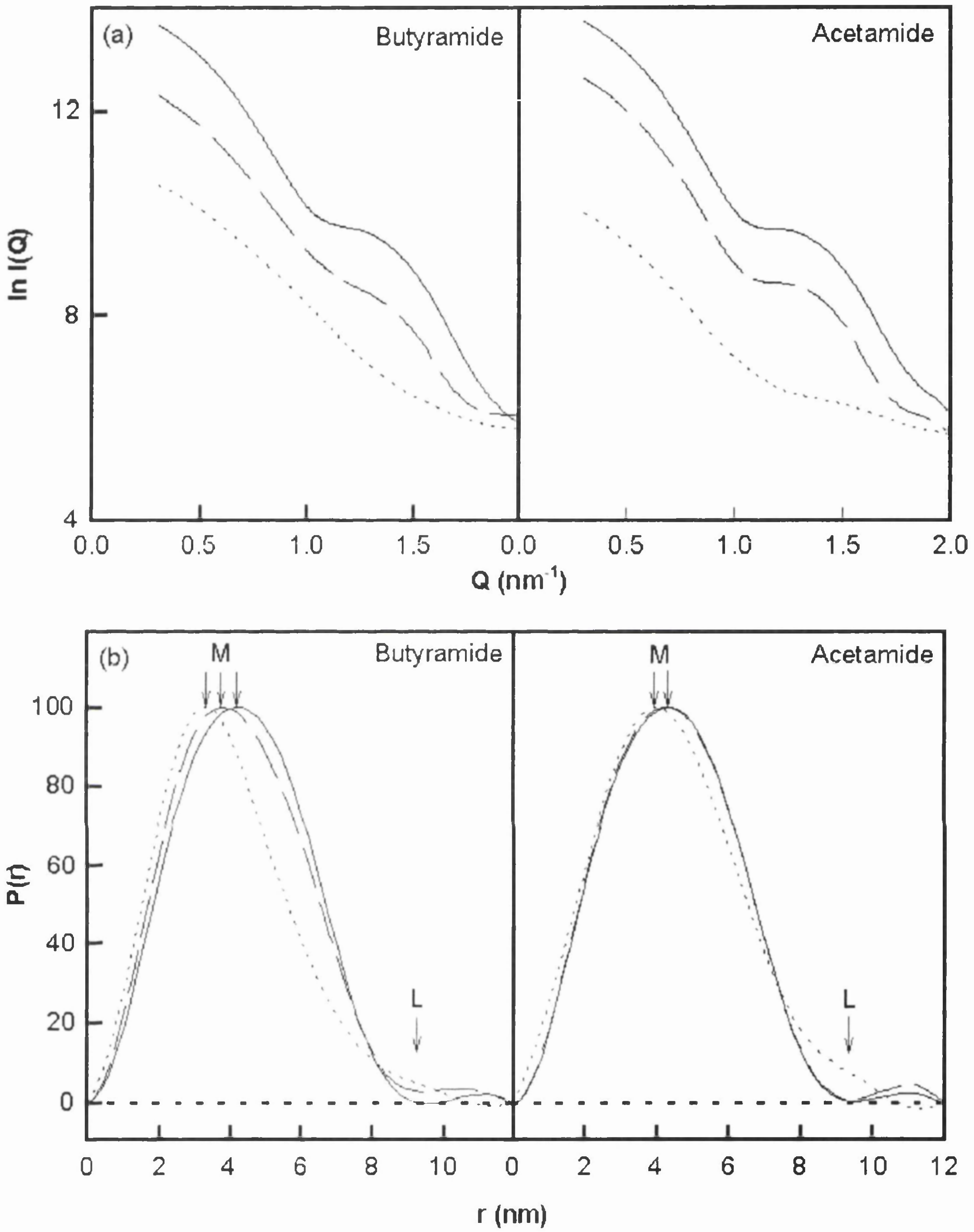


Figure 5.7: Comparison of the X-ray scattering curves I(Q) and distance distribution functions P(r) for AmiC-butyramide and AmiC-acetamide. (a) Concentration dependence of the X-ray I(Q) curves for AmiC-butyramide and AmiC-acetamide. For clarity, the I(Q) curves were smoothened using GNOM. Continuous: 13.7 mg/ml; Dashed: 4.1 mg/ml; Dotted: 1.0 mg/ml. (b) Corresponding distance distribution functions P(r) for AmiC-butyramide and AmiC-acetamide. Each concentration is denoted as in (a).

in overall length. The peak maximum M of the P(r) curves corresponds to the most commonly occurring distance in AmiC. The concentration dependence of M in Figure 5.6(c) exhibited similar trends to those already observed with the R_G and I(0)/c values, in that its position demonstrated a greater concentration dependence with AmiC-butyramide, and ranged from 4.2 nm to 3.2 nm.

(5.3.2) <u>Identification of AmiC Trimers by Neutron Scattering.</u>

Neutron scattering for AmiC in ^2H_2O buffers provided molecular weights as well as acting as a control for the absence of X-ray radiation damage effects and internal scattering density inhomogeneity effects in different solvents. AmiC is now visualised in a high negative solute-solvent contrast in place of the high positive contrast seen by X-rays. Linear Guinier R_G plots were obtained from both the neutron cameras D11 and LOQ (Figure 5.5b). The higher neutron flux on D11 permitted a dilution series of AmiC-butyramide and AmiC-acetamide to be measured between 0.8-5.1 mg/ml (data not shown). The neutron Guinier I(0)/c values confirmed the X-ray concentration dependence in Figure 5.6(b). The mean D11 R_G value for AmiC-butyramide was 3.26 ± 0.08 nm (5 determinations between 1.2-5.1 mg/ml), and that for AmiC-acetamide was 3.30 ± 0.06 nm (4 determinations between 0.8-1.9 mg/ml). The corresponding LOQ R_G values were in agreement at 3.30 ± 0.05 nm for AmiC-butyramide (1 determination at 2.6 mg/ml) and 3.26 ± 0.06 nm for AmiC-acetamide (1 determination at 2.8 mg/ml). The neutron R_G values were close to but slightly less than the X-ray value of 3.35 nm as expected (Table 5.2). The small decrease of up to 0.1 nm in the neutron R_G values is attributable to the surface location of hydrophilic amino acids and the core location of hydrophobic amino acids in AmiC, since hydrophilic residues have a higher scattering density than hydrophobic residues (Perkins, 1986, 1988).

M_r calculations were performed from the neutron I(0)/c values, as I(0)/c is measured relative to known standards. For D11 data, I(0)/c for AmiC-butyramide at 3.9 mg/ml in H_2O buffer was determined to be 0.072 ± 0.005 relative to the incoherent scattering of H_2O at a wavelength of 1.0 nm, and this gave an M_r of 127,000 ± 10,000. Since monomeric AmiC has an M_r of 42,600, this is equivalent to 3.0 ± 0.2 subunits. For LOQ data, the mean I(0)/c value of 0.176 observed for AmiC-butyramide and AmiC-acetamide referenced to a known polymer standard and other I(0)/c values determined for five proteins of known M_r between 51,000-144,000 (Mayans *et al.*, 1995; Ashton *et al.* 1995; Beavil *et al.*, 1995) gave an M_r of 150,000 ± 25,000, which corresponds to 3.6 ± 0.6 subunits. The full X-ray I(0)/c concentration series in Figure 5.6(b) shows that AmiC is predominantly trimeric between 5-10 mg/ml, and undergoes significant dissociation at concentrations below 5 mg/ml. As an I(0)/c value of 9.3 can be assigned to 3 AmiC subunits in Figure 5.6(b), an I(0)/c value of 3.1 will correspond to monomeric AmiC. Figure 5.6(a) shows that the R_G of the AmiC monomer is less than 2.5 nm, and that AmiC dissociates into monomers at low concentrations.

(5.3.3) X-ray Scattering Curve Simulations for Three Periplasmic Binding Proteins.

Curve simulations were performed using known crystal structures for free and complexed forms of the periplasmic binding proteins in order to assess whether solution scattering will monitor domain movements between their open and closed conformations.

(I) The periplasmic binding proteins from six clusters (Tam and Saier, 1993) exhibited

R_G values between 1.99-2.56 nm in a M_r range between 26,100-59,100 (Table 5.1). The Cluster 4 proteins AmiC and LivJ showed a decrease of 0.21 nm in R_G values on going from the unbound to the complexed form. The Cluster 3 proteins gave a smaller decrease of 0.13 nm, and that for the Cluster 1 proteins gave a decrease of 0.08 nm (Table 5.1).

(ii) Corresponding changes were seen in the full scattering curves out to Q of 2.0 nm^{-1} for these three groups of proteins (Figure 5.2). The scattering curve at low Q exhibited small changes to correspond to the changes in the Guinier region. More noticeable intensity changes between the free and complexed forms were visible in the Q range beyond 1 nm^{-1}.

Figure 5.2 also indicates the domain movements between the free and complexed forms of these proteins when the C-terminal domains were superimposed upon each other. While large domain movements of the order of 30-40° are observed and are detectable by solution scattering, Figure 5.2 and Table 5.1 show that accurate measurements will be required. In the case of trimeric AmiC, no domain movements could be detected within a precision in R_G values of 0.05 nm.

(5.3.4) X-ray Scattering Curve Simulations for Trimeric AmiC

To extend the data interpretation, scattering curve simulations were performed in three different analyses for trimeric AmiC, starting from the crystal structure for AmiC-acetamide. The trimer will have a three-fold axis of symmetry, as observed crystallographically for proteins such as tumour necrosis factor α, deoxyUTPase and chloramphenicol transferase. A structure based on the asymmetric association of a

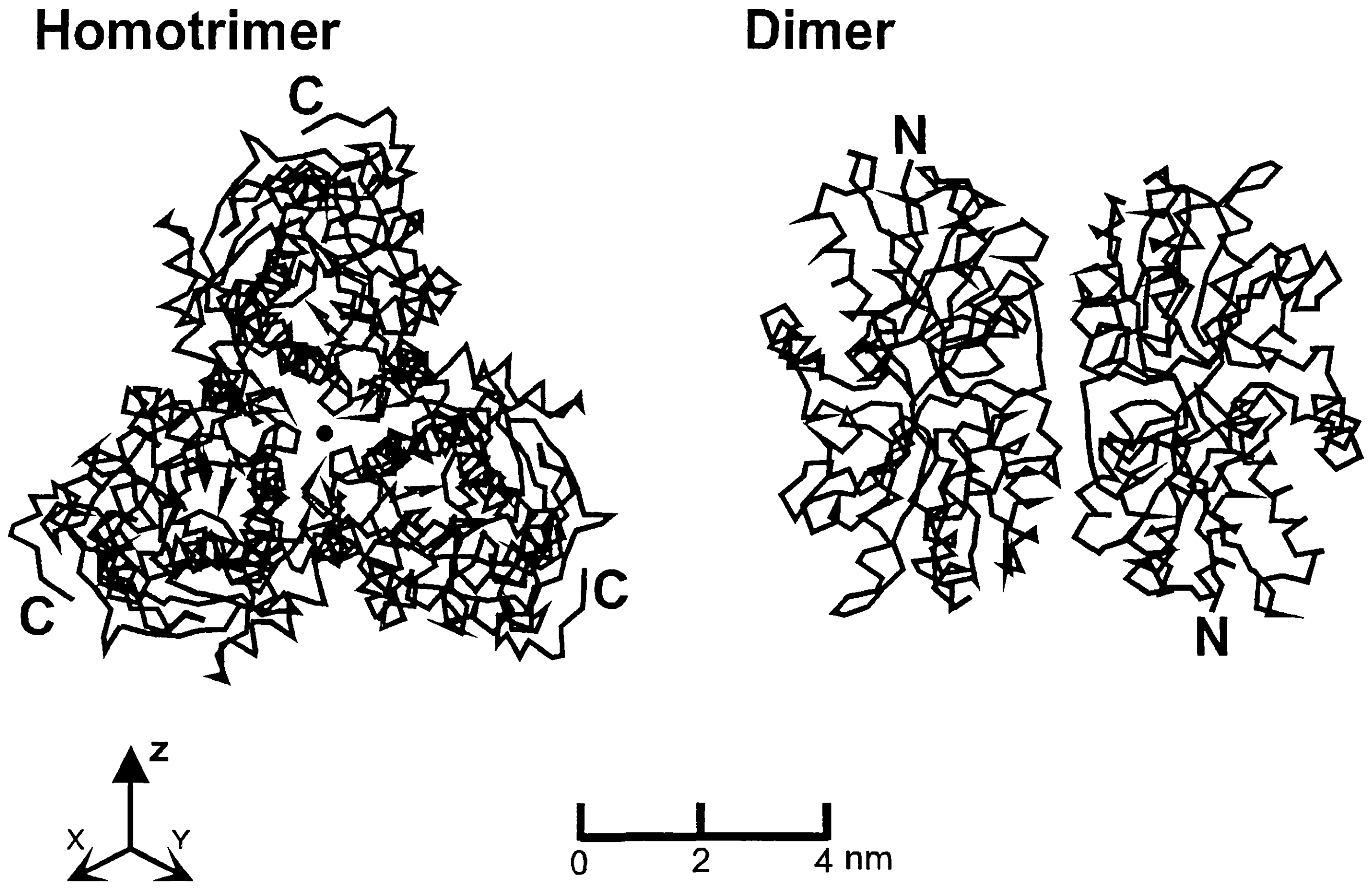


Figure 5.8: α-Carbon outlines of the homotrimer and dimer models for AmiC, based on the AmiC-acetamide crystal structure. The homotrimer of AmiC is viewed down the Z-axis which is indicated by the dot at the centre of the structure. The C-termini in this model are denoted by C. In the final model, the centre of mass of the monomer is 2.2 nm from the centre of mass of the trimer on its three-fold axis of symmetry. The putative crystallographic dimer is depicted as an antiparallel association of two monomers, in which the N-termini are denoted by N. In the putative crystallographic tetramer, the second dimer is rotated clockwise by 135° which is then positioned in front of the first dimer as shown.

monomer with a dimer is most unlikely on the grounds of symmetry. If a monomer is bound to one face of a dimer in such a trimer, a symmetry-related site for a second monomer will exist on the other side of the dimer, and AmiC would be tetrameric.

A symmetric AmiC homotrimer was constructed from three monomers whose longest axis were aligned parallel to each other with their ligand clefts arbitrarily set to face outwards and with a three-fold axis of symmetry between them along the Z-axis (Figure 5.8). Based on the scattering curve for AmiC-butyramide at 6.8 mg/ml, a one-parameter translational search explored the effect of varying the separation between the monomers in the XY-plane while retaining three-fold symmetry. The best model by this approach in Figure 5.8 gave a good curve fit 3s in Figure 5.9 with a low $R_{2.0}$ value of 4.7%. This model also resulted in a good fit (not shown) to the experimental curve at 1 mg/ml in Figure 5.7(a) with a satisfactory $R_{2.0}$ value of 7.3%, using a scattering curve constructed from 40% monomer and 60% homotrimer (curves 1 and 3s in Figure 5.9). This monomer:trimer ratio resulted in an estimated association constant K_{a3} of 2×10^{10} M^{-2}, where $K_{a3} = c_{trimer}/(c_{monomer})^3$ (McRorie and Voelker, 1993).

A second analysis was based on the monomer in the crystal structure of AmiC-acetamide, together with the putative dimer, trimer and tetramer (curves 1, 2, 3 and 4 in Figure 5.9: Methods). The curves changed in the Q range between 0.0-0.5 nm^{-1} to correspond to the increase in R_G with oligomerisation (Table 5.2). The dimer model (Figure 5.8) gave a reasonable curve fit for AmiC-butyramide at 1.0 mg/ml with a $R_{2.0}$ value of 8.0%, but this fit was visibly not as good as that for the monomer-homotrimer mixture above. In terms of the AmiC-butyramide scattering curve at 6.8 mg/ml, the four models gave poor curve fits with $R_{2.0}$ values of 15.3%, 12.1%, 9.7% and 39.3% respectively. In particular, the experimental curve at 6.8 mg/ml showed a subminimum at Q = 1.12 nm^{-1} that is different from that at Q = 0.98 nm^{-1} calculated from the tetramer model (curve 4). It was postulated that the observed curve may represent a combination of the four curves. Analysis of 5151 combinations of any three curves stepped in 1% increments gave a best fit with 0% monomer, 51% dimer and 49% tetramer. Analysis of all 176,851 combinations of four curves gave a best fit with 0% monomer, 51% dimer, 0% trimer and 49% tetramer (curve 2 + 4 in Figure 5.9). Although the $R_{2.0}$ value of 6.3% for this fit is reasonable, the curve fit is seen to deviate at Q values above 0.8 nm^{-1}. The limited success of these fits showed that the putative tetramer does not exist, and this supports the modelling based on a monomer-trimer equilibrium.

A third curve fit search assumed that AmiC at 6.8 mg/ml corresponded to an asymmetric trimer formed from the crystallographic monomer and dimer. In three-parameter X-, Y- and Z-axis translational searches, the long axis of the monomer was set perpendicular (curve 3:1a) or parallel (curve 3:2a) to that of the dimer. A clear minimum in the $R_{2.0}$ values was obtained for each of the X-, Y- and Z-axes in trial translational searches, and showed that a global minimum could be defined. The full systematic search was based on 31 × 31 × 41 steps of 0.2 nm along the X-, Y- and Z-axes and gave 39,401 coordinate models. Models were selected if they contained at least 1650 of the expected total of 1752 spheres (i.e. to retain only those models without monomer-dimer steric overlap) and had a calculated R_G value between 3.0-3.5 nm. From these searches, the best curve fits 3:1a and 3:2a had similar $R_{2.0}$ values of 3.9% and 4.1% (Figure 5.9 and Table 5.2). While these $R_{2.0}$ values are now better than those above, the importance of

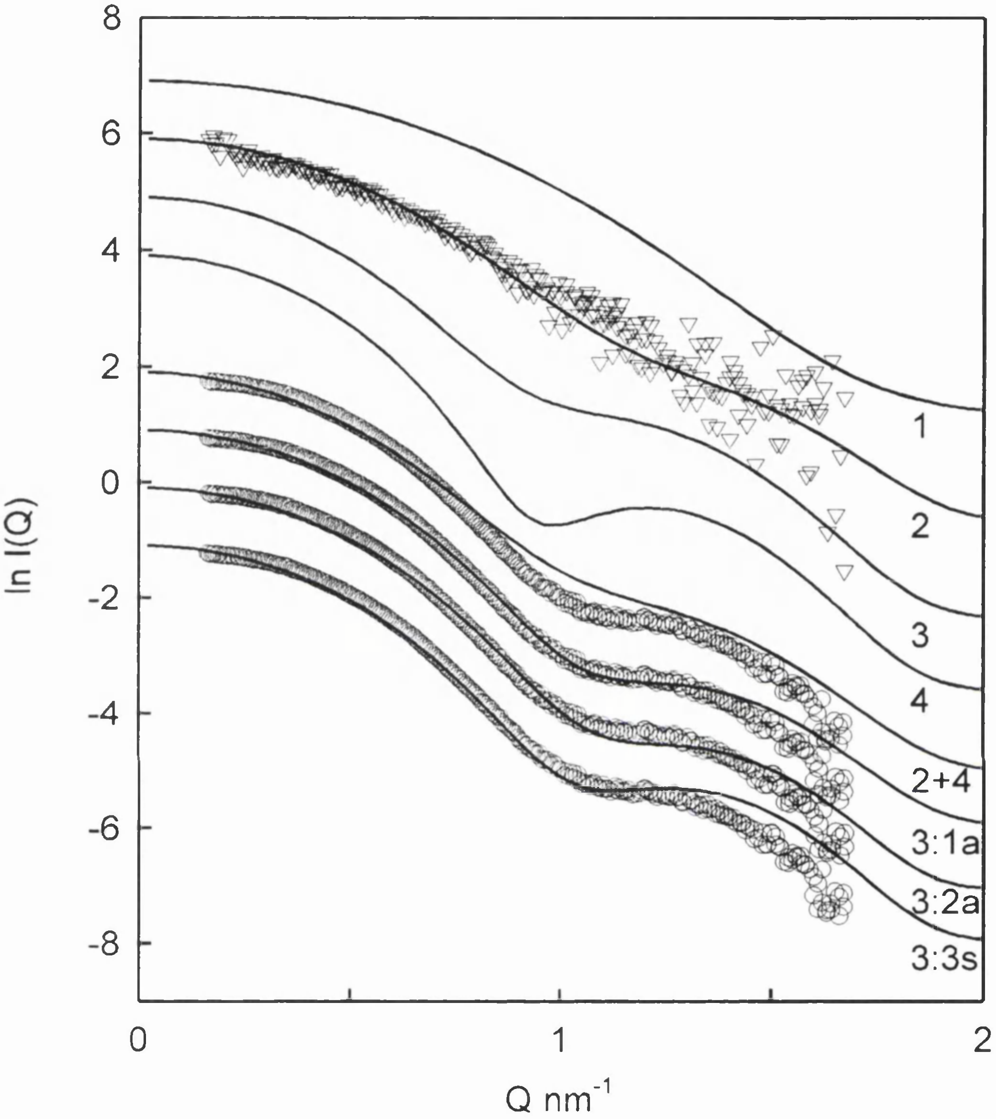


Figure 5.9: Curve fits of experimental scattering curves for AmiC-butyramide. Symmetric homotrimer, 3s; Monomer, 1; Dimer, 2; Trimer, 3; Tetramer, 4; Asymmetric trimers, 3:1a and 3:2a (Table 5.2). The curves were compared with X-ray data for AmiC-butyramide at 6.8 mg/ml. Trimer 1 corresponds to an AmiC monomer with its long axis perpendicular to that of the dimer (Figure 5.8). The monomer coordinates are superimposed on one of the two monomers in the dimer by translations of X = 1.7 nm, Y = 2.7 nm, Z = -1.5 nm, and rotations of X = 95°, Y = 12°, Z = -90°. Trimer 2 was generated from a 90° reorientation of the AmiC monomer such that the long axes of the monomer and dimer are parallel. The monomer coordinates are superimposed on one of the two monomers in the dimer by translations of X = -0.02 nm, Y = -3.0 nm, Z = -2.0 nm, and rotations of X = -14°, Y = -20°, Z = 195°.

Technique (Instrument)	Protein	Concentration (mg/ml)	Experimental R_G (nm)
Synchrotron X-ray (St 2.1)	AmiC-butyramide	7 to 16	3.35 ± 0.05
	AmiC-acetamide[1]	2 to 5	3.35 ± 0.05
	AmiC-acetamide[1]	0.4	3.12 ± 0.13
Neutron (D11)	AmiC-butyramide	1.2 to 5.1	3.26 ± 0.08
	AmiC-acetamide[1]	0.8 to 1.9	3.30 ± 0.06
Neutron (LOQ)	AmiC-butyramide	2.6	3.30 ± 0.05
	AmiC-acetamide[1]	2.8	3.26 ± 0.06

AmiC Models	Concentration (mg/ml)	R-factor (2.0 nm^{-1})	Modelled R_G (nm)
AmiC scattering (symmetric trimer)	6.8[2]	4.7	3.39
AmiC crystallographic monomer	-	-	2.34
AmiC crystallographic dimer	-	-	2.97
AmiC crystallographic trimer[3]	-	-	3.53
AmiC crystallographic tetramer	-	-	3.66
AmiC scattering (dimer + tetramer)	6.8[2]	6.3	3.38
AmiC scattering (asymmetric trimer 1)	6.8[2]	4.1	3.34
AmiC scattering (asymmetric trimer 2)	6.8[2]	3.9	3.32

[1] Not shown in Figure 5.5.

[2] Shown in Figures 5.8 and 5.9.

[3] Generated by deleting any one of the four AmiC structures in the tetramer.

Table 5.2: X-ray and neutron scattering parameters for AmiC samples and models.

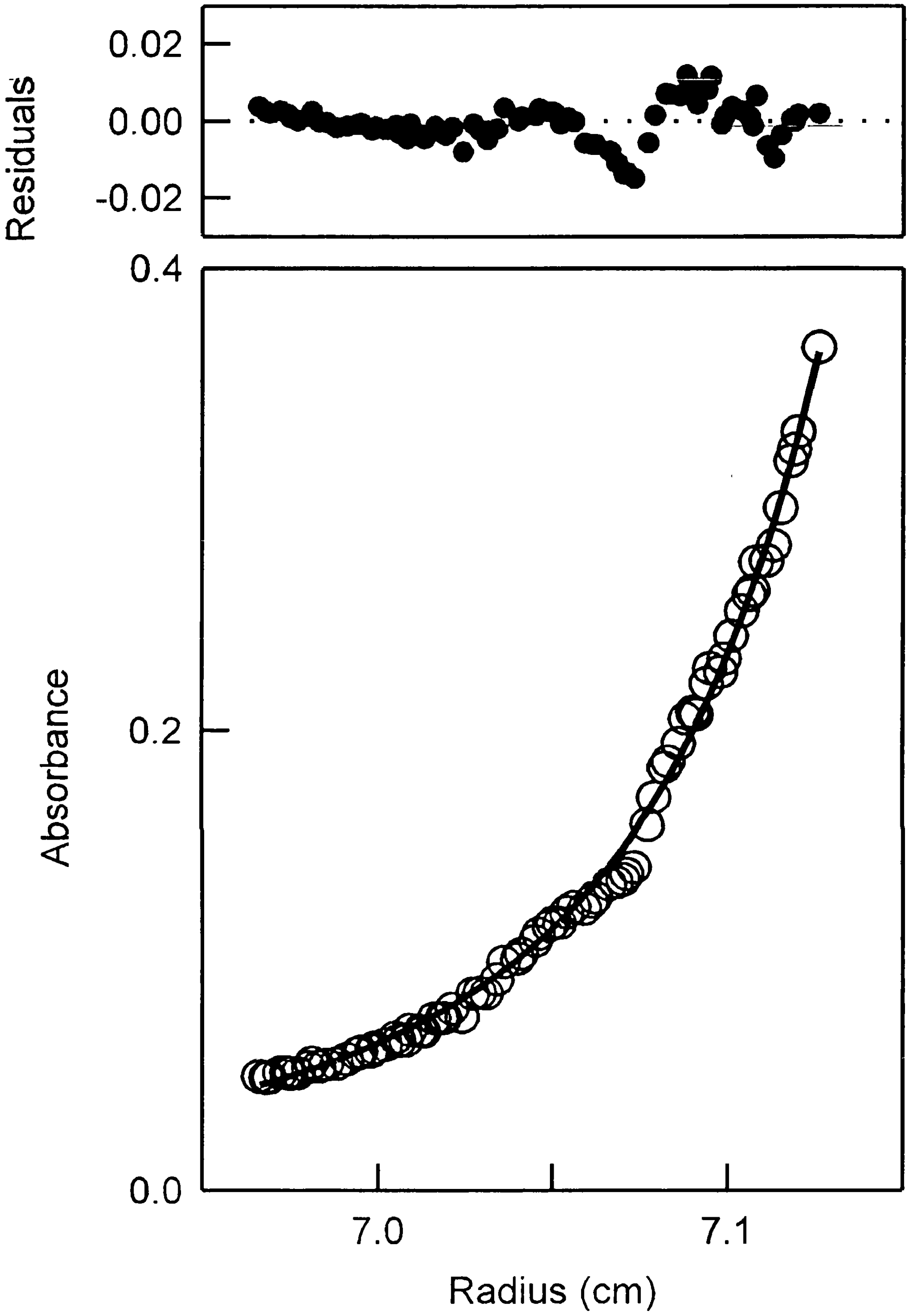


Figure 5.10: Sedimentation equilibrium data for AmiC. Absorbance values are shown as a radial distribution at equilibrium for a loading concentration of 1 mg AmiC/ml at 15°C at a rotor speed of 15,000 r.p.m. The data are fitted with a monomer-trimer model which is shown as a line through the experimental points. The corresponding distribution of the residuals is shown above the plot.

this search is that it showed that a good fit can be obtained from a model with incorrect symmetry. As these models gave the best fits, they were also used for neutron curve fits as a check of consistency. The fits (not shown) gave good R-factors in both H_2O and ^2H_2O buffers, and the calculated curve deviated slightly from the two observed curves in opposite directions at larger Q values as expected from the two opposite solute-solvent contrasts in use.

(5.3.5) Sedimentation equilibrium analyses of AmiC.

Preliminary sedimentation equilibrium experiments were performed to determine the molecular weights of AmiC using three buffer systems. For 1 to 4 mg/ml AmiC in AmiC buffer with 10 mM butyramide, the assumption of a single ideal species in the IDEAL1 model (Methods) resulted in weight averaged $M_{w,app}$ values that increased from 77,000 to 117,000 (1.8 to 2.8 monomers). This showed a concentration-dependent trimerisation in agreement with Figure 5.10. For 1 to 4 mg/ml AmiC, unchanged $M_{w,app}$ values corresponding to 2.8 ± 0.4 monomers were observed on the addition of 150 mM $(NH_4)_2SO_4$ to the first buffer, and to 3.3 ± 0.2 monomers on the further addition of 10 mM acetamide to the buffer. For the AmiC-butyramide data, final fits were performed in terms of a reversible monomer-trimer equilibrium using the ASSOC4 model (Methods). The AmiC-butyramide data at 15,000 r.p.m. were satisfactorily fitted to a monomer-trimer model with K_{a3} determined to be 10^9 to 10^{10} M^{-2}, where at this speed nonreversibly associated species will be cleared from the sample. Good exponential curve fits were obtained, and small residuals were obtained that were distributed irregularly as a function of radius (Figure 5.10). These experiments provided further evidence for the existence of trimeric AmiC at higher concentrations, and showed that trimer formation was promoted by the addition of $(NH_4)_2SO_4$ or $(NH_4)_2SO_4$ /acetamide. To improve the quality of the data, experiments were attempted using longer column lengths but unfortunately the system would not equilibrate in an acceptable period of time.

(5.4) Conclusions.

The classical view of periplasmic binding proteins is that they are monomeric with two domains that undergo large conformational change during ligand binding. While AmiC is most closely related to the Cluster 4 protein LivJ in sequence and structure, AmiC is a cytoplasmic protein that controls the activity of AmiR which is directly involved with *ami* gene expression, while LivJ is a perisplasmic protein that binds aliphatic amino acids. Both exhibit similar interactions with membrane-bound proteins (Wilson *et al.*, 1995). Unexpectedly AmiC exhibits oligomeric properties at high concentrations. Other periplasmic binding proteins are generally monomeric, although the galactose binding protein from *S. typhimurium* and *E. coli* and the maltose binding protein from *E. coli* are dimeric, and the addition of ligand causes them to become monomeric (Mowbray and Petsko, 1983; Richarme, 1982). The crystal structures of histidine binding protein (1hsl) and a maltodextrin binding protein mutant (1mdp) reveals dimeric structures, however these are attributable to repeated lattice contacts between the same domain in a pair of monomers. In distinction to these examples, AmiC forms antiparallel dimers in its crystal structure (Figure 5.9), most probably the consequence of the crystal lattice, and it participates in a monomer-trimer association in solution. These results show that AmiC behaves differently from the classical periplasmic binding proteins.

Above 2 mg/ml, AmiC is predominantly trimeric. That trimer formation is promoted in the presence of its ligand acetamide rather than the anti-inducer butyramide may be of biological interest. The concentration dependence of the scattering curves is clear from Figure 5.7, and the I(0)/c graphs rise with c to a value corresponding to trimers in Figure 5.6. Initial analytical ultracentrifugation data by sedimentation equilibrium methods using a Beckman XLA ultracentrifuge also showed this concentration dependence, and yielded a comparable estimate of the association constant K_{a3} (Ralston, 1993; McRorie and Voelker, 1993). Molecular modelling based on the AmiC-acetamide crystal structure showed that trimeric structures gave R_G values that corresponded to the observed values. The possibility of an imposter model based on a mixture of dimers and tetramers was also considered. This was discounted for two reasons: (I) A single peak and not a double peak was routinely observed during AmiC purifications by gel filtration (Methods; Figure 5.4); (ii) Modelling of the X-ray data showed that it was not possible to optimise a curve fit based on a mixture of dimer and tetramer that was equivalent to or better than one based on a trimer (Figure 5.9).

Unlike periplasmic binding proteins in general, scattering showed that no conformational changes between AmiC-acetamide and AmiC-butyramide trimers could be detected within a precision of 0.05 nm in R_G value. Butyramide contains an extra pair of $CH_2.CH_2$ carbon atoms. As butyramide binds 100-fold more weakly to AmiC than acetamide (Wilson *et al.*, 1993), this weaker binding may reflect the energy required to accommodate butyramide within its binding site in AmiC without a large cleft opening. Trimer formation in AmiC would involve extensive contacts between the monomers, and may hinder the free movement of the cleft on change of ligand. The question of whether this domain movement occurs or not on ligand binding will require data collection on monomeric AmiC or on the complex between AmiC and AmiR. Nonetheless the larger size of the butyramide ligand has clearly reduced the stability of the AmiC trimer.

Solution scattering will monitor domain movements in the periplasmic binding proteins, even though the changes are small (Figure 5.2 and Table 5.1). Here, even though L-arabinose binding protein was monomeric and well-behaved between 4-36 mg/ml concentrations (Newcomer *et al.*, 1981), AmiC' demonstrated interparticle interference effects above 10 mg/ml as well as oligomer formation in its scattering curves. Dilution series and absolute M_r calculations were key controls of the present scattering data. Data analyses were enhanced by the use of automated constrained curve fitting procedures. Hypotheses could be stated which could then be tested in detail with relatively little effort, although expensive in terms of CPU time, and enabling a choice to be made between a monomer-trimer or a monomer-dimer-tetramer association. Automated curve fits had previously been used to assess domain structures in single large multidomain proteins (Mayans *et al.*, 1995; Beavil *et al.*, 1995; Boehm *et al.*, 1996). The present study shows that the method is applicable to the study of protein-protein complexes. It should be noted that a good curve fit is only a test of consistency, and will not constitute a unique low resolution structure determination. The good curve fit obtained for the symmetry-forbidden asymmetric trimer of AmiC is an illustration of this limitation.

Following the scattering studies, thermal denaturation experiments (B. O'Hara, G. Siligardi, S. A. Wilson, L. H. Pearl and R. E. Drew, 1998, manuscript in preparation) have shown that AmiC-butyramide is less stable than AmiC-acetamide. The crystal structure of AmiC-butyramide has been determined and shows that no major conformational changes in AmiC were observed. The root-mean-square difference between the α-carbon coordinates of both forms of AmiC was 0.040 nm. This is supported by circular dichroism spectroscopy of AmiC-butyramide and AmiC-acetamide which also suggested no major conformational changes. These results are consistent with the present scattering data in the sense that the structure of the monomers remain essentially unchanged, but that subtle interactions take place that influence the ability of AmiC to form trimers.

CHAPTER 6

EXPRESSION OF RECOMBINANT COMPLEMENT FACTOR I

(6.1) Introduction.

Factor I is an essential multidomain serine protease which plays an important role in the regulation of the complement cascade. It circulates as an active protease and exhibits an unusual mechanism of action in that it has no known natural protease inhibitors and a very restricted substrate specificity. In the alternative pathway of complement activation, complexes involving C3b and the Bb fragment of factor B form the C3 convertase, whereas in the classical pathway complexes involving C4b and C2a form the C3 convertase. The convertases cleave C3 into C3b, which generates more C3b.Bb in an amplification loop. Factor I cleaves C3b into iC3b and C3f in the presence of a cofactor, namely factor H or membrane cofactor protein (MCP; CD46) or complement receptor type one (CR1; CD35) (Figure 6.1). Factor I also cleaves C4b into C4c and C4d with C4b binding protein (C4BP) or CR1 or MCP as cofactor. Therefore, factor I is required for regulating the extent of C3 cleavage, the formation of C3b and ultimately the level of C3 convertase (Law and Reid, 1995; Sim *et al.*, 1993). In factor I-deficient patients, symptoms are displayed that are characteristic of C3 deficiencies with susceptibilities to recurrent pyogenic infections (Vyse *et al.*, 1996).

The human factor I gene is located in chromosome 4 and the protein is synthesized as a pre-proenzyme that undergoes post-translational glycosylation and proteolytic processing before secretion (Goldberger *et al.*, 1984, 1987). The full sequence contains 583 amino acids with an 18-residue N-terminal signal peptide, a 318-residue heavy chain, a 4-residue linker peptide RRKR and a 243-residue light chain. Both the signal and linker peptides are removed during processing (Goldberger *et al.*, 1984; Catterall *et al.*, 1987). The six N-linked glycosylation sites on factor I result in a

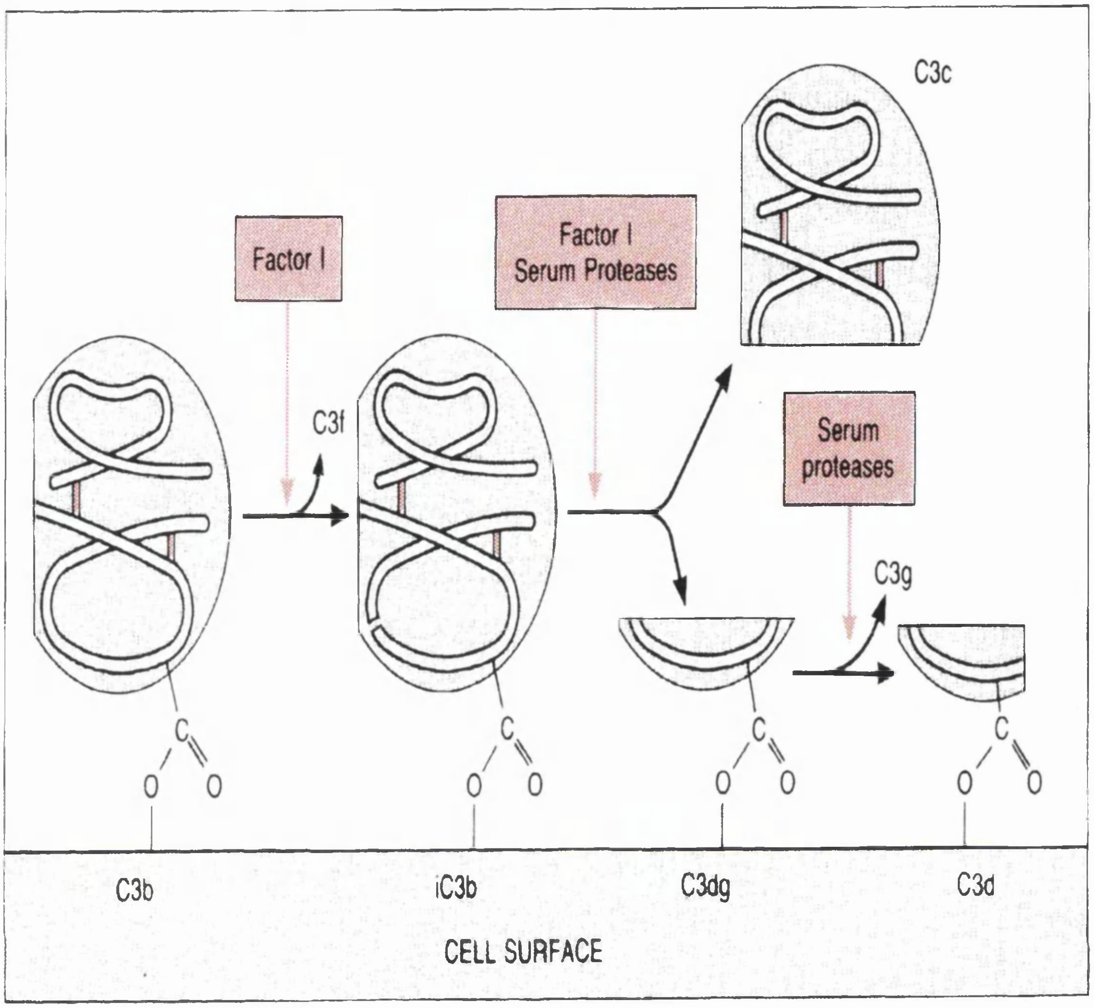


Figure 6.1: Breakdown of C3b. (1) Surface-bound C3b with an α' chain of 103 kDa and a β chain of 75 kDa is cleaved by factor I, in the presence of cofactors, at two places on the α'chain to release C3f (3 kDa), leaving the α' chain in 2 fragments of 60 kDa (N-terminal) and 40 kDa (C-terminal). (2) Serum proteases or factor I in association with CR1 cleave the 60 kDa fragment into two fragments of 23 kDa (N-terminal) and 37 kDa (C3dg). The released C3c then contains three chains including an intact β chain and two fragments derived from the α' chain of 23 kDa and 40 kDa. (3) Addition of exogenous proteases to surface-bound C3dg results in the release of C3g (5 kDa) leaving C3d (32 kDa) on the cell surface. Interchain disulphide bonds are indicated in red. Based on Law and Reid, 1995

carbohydrate content of 15%-26% that causes factor I to migrate anomalously on SDS-PAGE with an apparent molecular weight of 88,000 instead of an expected value between 75,000-85,000 (Perkins *et al.*, 1993a). Factor I contains five domains. The heavy chain corresponds to a factor I/membrane attack complex (FIMAC) domain, a CD5 domain (also known as the scavenger receptor cysteine-rich domain) and two low density lipoprotein receptor (LDLr-1 and LDLr-2) domains, while the light chain corresponds to a serine protease (SP) domain (Law and Reid, 1995; Sim *et al.*, 1993; Kunnath-Muglia *et al.*, 1993; Minta *et al.*, 1996).

Even though factor I is present in plasma at a high level (35 μg/ml, 0.4 μM), an expression system is required to understand the molecular structure and function of factor I. Factor I expression systems have been reported in a cell-free system in *Xenopus* oocytes programmed with mRNA from human liver, in three hepatoma-derived cell lines, and in COS-1 and CHO-K1 cells (Goldberger *et al.*, 1984; Wong *et al.*, 1995). These expression systems resulted in the secretion of a variable quantity of unprocessed pro-factor I and loss of activity, and factor I-deficient serum was required in the growth media. We report here the expression, purification and characterisation of human factor I using a recombinant baculovirus system, and assess the consequence of an altered glycosylation on its activity. Recombinant factor I is expressed at a higher level than in COS-1 cells using its native secretion signal. Production can be easily scaled up, serum-free medium is used, and post-translational processing yields a functionally active product.

The availability of rFI will make possible new mutagenesis approaches to determine the molecular basis of its interactions with factor H and C3b. A rational strategy for this requires structural information. Secondary structures for both serum-derived and recombinant factor I (sFI and rFI) were determined using circular dichroism (CD) and Fourier transform infrared (FTIR) spectroscopy. Relevant crystal structures have become available for the FIMAC, LDLr and SP domains (Hohenester *et al.*, 1997; Fass *et al.*, 1997; Spraggon *et al.*, 1995), and their superfamilies have been characterised (Ullman and Perkins, 1997; Ullman *et al.*, 1995; Perkins and Smith, 1993). While an atomic structure is unavailable for the CD5 domain at present, its superfamily and its secondary structure and disulphide bridge connections have been analysed (Chamberlain *et al.*, 1998; Resnick *et al.*, 1994, 1996). The comparison of these spectroscopic studies with structural data for the domains in factor I shows that this is predominantly a β-sheet protein.

(6.2) Materials and Methods.

(6.2.1) Cloning and Expression of Human Factor I into a Baculovirus System.

Using the restriction enzyme EcoRI, the factor I coding region was excised from a pBluescript plasmid (Stratagene) containing the factor I sequence (Catterall *et al.*, 1987). The EcoRI fragment was subcloned into the EcoRI site of pVL1393 (Invitrogen). The correct orientation of the insert was determined by digestion with the restriction enzyme PstI as there is a single site in both the plasmid and the factor I sequence (Figure 6.2). The recombinant plasmid containing the factor I sequence (pVLFI), was purified and used to transfect 2 × 10^6 *Spodoptera frugiperda* Sf21 cells (Invitrogen) grown overnight at 28°C in two 35 mm diameter dishes. The transfection mixture comprised 3 μg pVLFI, 500 ng Baculovirus Gold DNA (Pharmingen), 1 ml TC100 (Gibco BRL) containing 100 U/ml of penicillin, 100 μg/ml streptomycin and 20 μl Insectin liposomes (Invitrogen) according to the manufacturer's instructions. Recombinant baculoviruses were purified by plaque assays using Sf21 cells (King and Possee, 1992; Figure 6.3). The pure recombinant virus was then propagated to a high titre stock in Sf9 cells adapted for growth in suspension in Sf900 II Serum Free Media (Gibco) containing 100 U/ml of penicillin and 100 μg/ml streptomycin. Expression of recombinant factor I (rFI) was achieved by infecting *Trichoplusia ni* (High Five) cells (Invitrogen), adapted for suspension culture growth in Excell 401 Serum Free Media (JRH Biosciences) supplemented with 100 U/ml of penicillin and 100 μg/ml streptomycin at a multiplicity of infection of 3. These cells were grown at 28°C for 72 hours in a Gallenkamp orbital incubator with gentle shaking at 70 r.p.m.. The supernatant was clarified by centrifugation at 2000 *g* for 5 minutes at 4°C, and was stored at 4°C following the addition of 5 mM EDTA and 1 mM Pefabloc-SC (Pentapharm).

(6.2.2) Purification of rFI and sFI.

rFI was purified from the culture supernatant by affinity chromatography using the mouse monoclonal antibody MRC-OX21, which was purified from the cell culture supernatant of the hybridoma cell line MRC-OX21 (ECACC) and immobilised on Sepharose 4B (Pharmacia) (Sim *et al.*, 1993; Hsiung *et al.*, 1982). The pH of the cell culture supernatant containing 5 mM EDTA and 1 mM Pefabloc-SC was adjusted to 7.0 using Tris base. The supernatant was passed through the MRC-OX21 Sepharose 4B column equilibrated in 25 mM Tris-HCl, 140 mM NaCl, 0.5 mM EDTA, pH 7.4 (Buffer A). The column was washed with Buffer A and the bound protein was eluted with 3 M $MgCl_2$, pH 6.8 (pH adjusted with Tris base) (Sim *et al.*, 1993). rFI was dialysed into

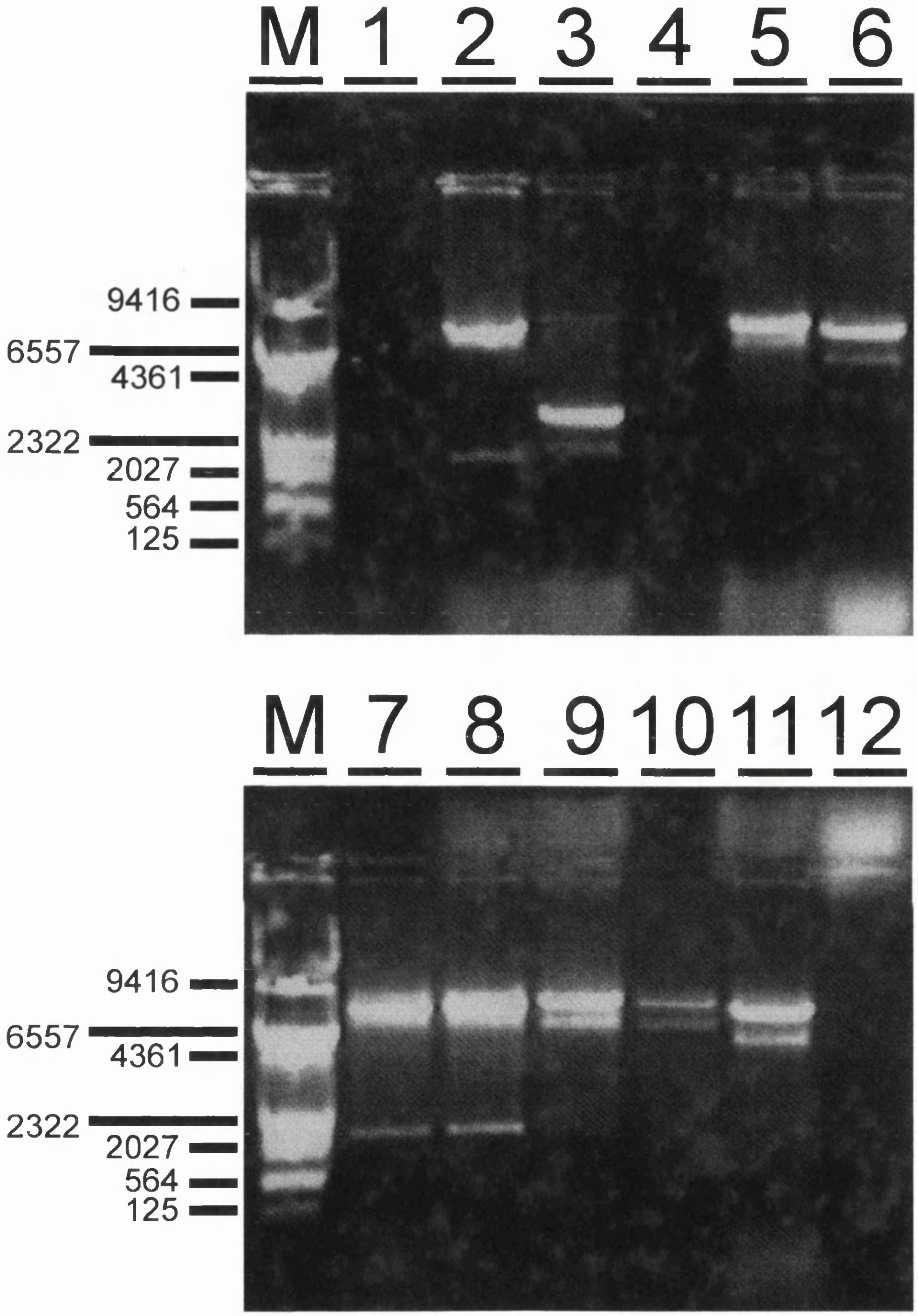


Figure 6.2: 1.2 % agarose gel of *Pst*I restriction digests to check the orientation of the factor I cDNA insert in the baculovirus transfer vector. Lanes 2, 7 and 8 contain bands of the correct size for the correct orientation of the insert. Lanes 3, 5, 6, 9, 10 and 11 did not contain digestion products of the correct size.

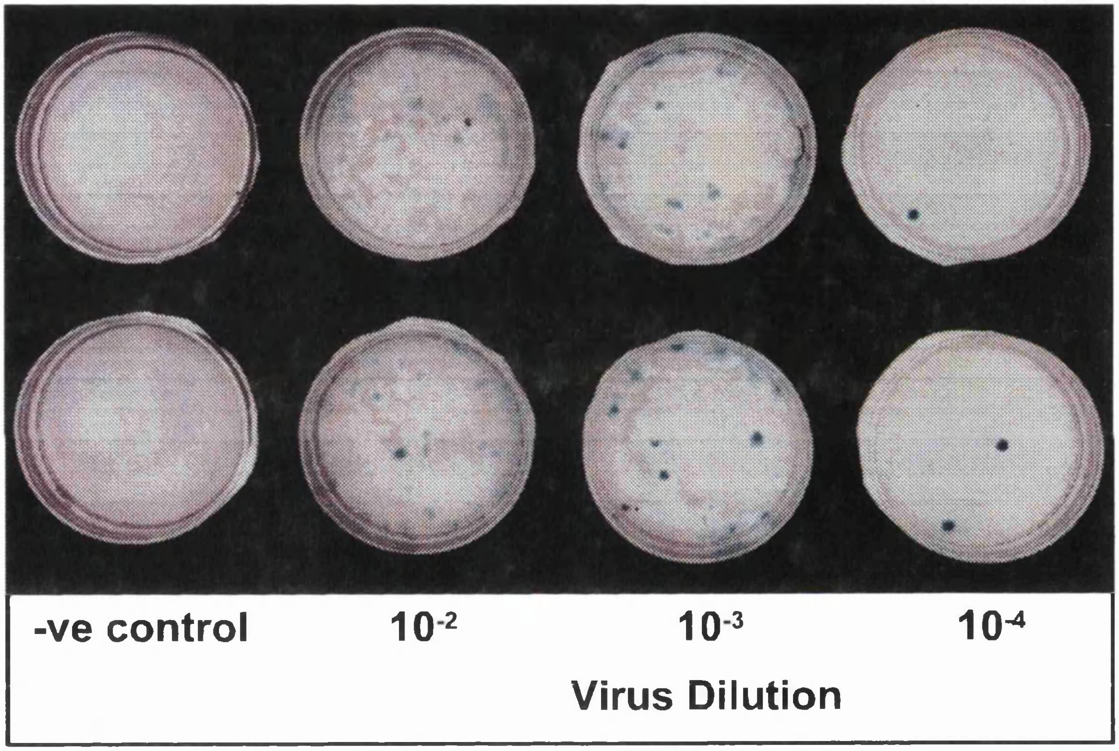


Figure 6.3: Plaque assay of AcMNPV.*lacZ* positive control virus after transfection of Sf21 cells. AcMNPV.FI is not shown because its plaques are clear (this recombinant virus is *lacZ* negative) and therefore are not very easy to see in a photograph.

Buffer A, then into Dulbecco's phosphate buffered saline (PBS) with 0.5 mM EDTA for further purification. The protein was concentrated under N_2 pressure using a YM10 membrane (Amicon) and passed through a 1.6 × 60 cm column of Superdex 200 preparation grade gel filtration medium (Pharmacia) to remove aggregated material and contaminants.

Serum-derived FI (sFI) was used as an experimental control. Using separate apparatus to avoid cross-contamination with rFI, this was purified from 0.6 l outdated serum by MRC-OX21 affinity chromatography according to Sim *et al.* (1993), then by ion exchange through a MonoQ column (Pharmacia) using a 0-250 mM NaCl gradient in 20 mM Tris, 0.5 mM EDTA, pH 9.0, followed by gel filtration through a 1.6 × 60 cm Superdex 200 preparation grade column equilibrated in PBS and 0.5 mM EDTA. sFI was concentrated as above.

(6.2.3) <u>Analyses of post-translational modifications in rFI.</u>

rFI was analysed by 8% polyacrylamide SDS-PAGE under alkylating (non-reducing) and reducing conditions in 4 M urea, 0.1 M Tris-HCl, 0.1% SDS, pH 8.0 containing either 20 mM iodoacetamide or 10 mM dithiothreitol, respectively. For sequencing, the samples were run reduced on SDS-PAGE and then electroblotted onto ProBlott membrane (Perkin-Elmer, Applied Biosystems Division). This was stained with Coomassie Brilliant Blue and the band of interest was excised and sequenced using an Applied Biosystems 494A Procise sequencer. N-terminal sequencing was performed by Mr A. C. Willis (MRC Immunochemistry Unit, Oxford).

In order to confirm glycosylation, 1 μg of reduced rFI and sFI were electrophoresed in an 8% acrylamide SDS-PAGE gel and were then blotted onto a PVDF membrane (Biorad). A bacterially expressed glutathione S-transferase fusion protein with the LDLr-1/2 domain pair of factor I was used as a negative control (Ullman *et al.*, 1995; Ullman, 1994). The membrane was then blocked with 7% fish gelatin (Sigma) in PBS at 4 °C overnight and washed once with Buffer B (50 mM Tris-HCl, 140 mM NaCl, 10 mM $CaCl_2$, 10 mM $MgCl_2$, 0.5 % Triton X-100, pH 7.4). 10 μg/ml of concanavalin A-biotin conjugate (Sigma) dissolved in Buffer B was added to the membrane and incubated at room temperature for 1 h. The membrane was washed five times with Buffer B followed by the addition of 12.5 μg/ml of streptavidin-(alkaline phosphatase enzyme polymer) (Sigma) dissolved in Buffer B. After 1 h at room temperature, the membrane was washed five times using a 5 min incubation with 10 ml of Buffer B each time and developed using 5-bromo-4-chloro-3-indolyl phosphate and nitro blue tetrazolium (Sigma).

(6.2.4) Western Blot Analyses.

rFI was analysed by Western blots following 8% acrylamide SDS-PAGE and transfer to BioBond-NC nitrocellulose membrane (Whatman). sFI was analysed for comparison. Reactivity with the following antibodies (used as purified IgG fractions) was tested: polyclonal rabbit antibodies against bacterial fusion proteins of FIMAC, the LDLr-1/2 domains and the LDLr-2 domain of factor I (Ullman *et al.*, 1995; Ullman, 1994); rabbit polyclonal antibodies against whole sFI (a gift from Dr. Teisner, University of Odense, Denmark) and the mouse monoclonal antibodies MRC-OX21 (Sim *et al.*, 1993) and MRC-OX24 (Sim *et al.*, 1993). Approximately 1 μg of rFI was loaded in each lane. Uninfected cell culture supernatant was used as a negative control. Polyclonal antibodies were incubated with the blot at final concentrations of 170-200 μg/ml, and monoclonals at 3 μg/ml, all in PBS with 1% (w/v) low fat milk powder. Following incubation with either an anti-rabbit or anti-mouse alkaline phosphatase conjugate (Sigma, Biorad) which was diluted 1/6,500 using PBS with 1% (w/v) low fat milk powder, the blots were washed with 0.15 M Tris-acetate pH 9.6 and developed using 5-bromo-4-chloro-3-indolyl phosphate and nitro blue tetrazolium in 0.15 M Tris-acetate, pH 9.6.

(6.2.5) Activity of rFI and sFI.

rFI was tested for activity in cleaving amidated C3 ($C3(NH_3)$) which was used as an equivalent of C3b, essentially as described in Sim and Sim (1983). 240 μg of C3 purified from plasma (Dodds, 1993) was converted to $C3(NH_3)$ by dialysis into 100 mM ammonium bicarbonate at pH 8.0 and incubation at 37°C for 1 h. $C3(NH_3)$ was then labelled with 0.5 mCi ^{125}I (Amersham) using Iodogen as a catalyst (Fraker and Speck, 1978). 25 μl assays were set up using 180 ng of ^{125}I-$C3(NH_3)$, 750 ng of factor H (purified from plasma; Sim *et al.*, 1993) and 21 ng or 22 ng (~0.25 nM) of rFI or sFI respectively, together with 2 μg soya bean trypsin inhibitor (Sigma), 2 mM Pefabloc-SC in 10 mM potassium phosphate, 0.5 mM EDTA, pH 7.0. The samples were incubated for 0, 5, 15, 30, 60 and 120 min and analysed by SDS-PAGE in 7% acrylamide under reducing conditions, followed by drying and exposure to autoradiographic film. The level of activity of both rFI and sFI was assessed from the ratio of the band densities corresponding to the intact α' chain of $C3(NH_3)$ at a molecular weight (M_r) of 108,000 and the cleavage product at M_r 43,000 to provide the percentage of $C3(NH_3)$ cleaved by factor I. Band densities were quantified using a Shimazdu CS9001PC densitometer.

(6.2.6) Circular Dichroism and Fourier Transform Infrared Spectroscopy.

Circular dichroism (CD) spectroscopy was performed at 20°C using a Jobin-Yvon CD6 spectropolarimeter with quartz cells of path lengths 0.2 mm and 0.5 mm. The instrument was calibrated with an aqueous solution of recrystallised D10-camphorsulphonic acid ($\theta^{0.1\%, 1\,\mathrm{cm}} = 0.308$ at 290 nm). rFI and sFI were each dialysed into 5 mM potassium phosphate, pH 7.0. Protein concentrations were calculated from absorbances measured at 280 nm, using an absorption coefficient (1%, 1 cm) of 14.0 for rFI calculated on the basis of six high mannose type oligosaccharides, and 12.0 for sFI calculated on the basis of six complex type oligosaccharides (Perkins, 1986). Spectral quantification to obtain secondary structures was performed using CONTIN (Provencher, 1982).

Fourier transform infrared (FTIR) spectroscopy was performed at 25°C using a 1750 Perkin-Elmer FT-IR spectrometer continuously purged with dry air to reduce water vapour absorption in the spectral region of interest. The limited solubility of factor I meant that it was not possible to obtain FTIR spectra in the high background of H_2O buffer, however data collection in ^2H_2O buffer was possible. Samples and buffer were measured using a CaF_2 cell fitted with a 50 µm pathlength Teflon spacer. 200 scans were signal averaged at a resolution of 4 cm^{-1}. The absorbance spectrum was obtained by digital subtraction of the ^2H_2O buffer spectrum from the sample spectrum (Haris *et al.*, 1986). Detailed analysis of the amide I band was carried out using a second derivative procedure provided by the GRAMS software (Perkin-Elmer) with a 13 data point Savitzy-Golay smoothing window.

(6.3) Results and Discussion.

(6.3.1) Expression and Purification of Recombinant Factor I.

To express factor I, the 1.9 kb gene fragment containing the coding region of factor I and a 5' untranslated region was subcloned from a pBluescript vector into the baculovirus transfer vector pVL1393 (Methods). This coding region contained the 18-residue leader and 4-residue linker peptides, both of which are enzymatically removed during processing of prepro-factor I in human cells. Recombinant protein was detected both in the culture supernatant and in the cellular extract 24 h after infection with recombinant virus. By 48 h post-infection, rFI was secreted into the culture supernatant, and optimal expression was noted after 72 h by Western blot analysis (Figure 6.4). The yield was estimated to be in a range between 2-10 mg rFI/litre of culture by ELISA assays and Western blot analyses.

As the insect cells and recombinant virus were grown in serum-free media, rFI was purified directly from the culture supernatant by affinity chromatography using immobilised MRC-OX21, which is a monoclonal antibody specific for factor I. This step removed cellular proteins that had been released into the supernatant by virus-induced lysis. A gel filtration step then removed minor contaminants to give pure rFI. From 2 l of cell culture, it was possible to obtain 2.5 mg rFI. Analysis with SDS-PAGE demonstrated a single protein species under non-reducing conditions, and this migrated as two bands under reducing conditions. Figure 6.5 thus demonstrated the processing of the 4-residue linker between the heavy and light chains of rFI.

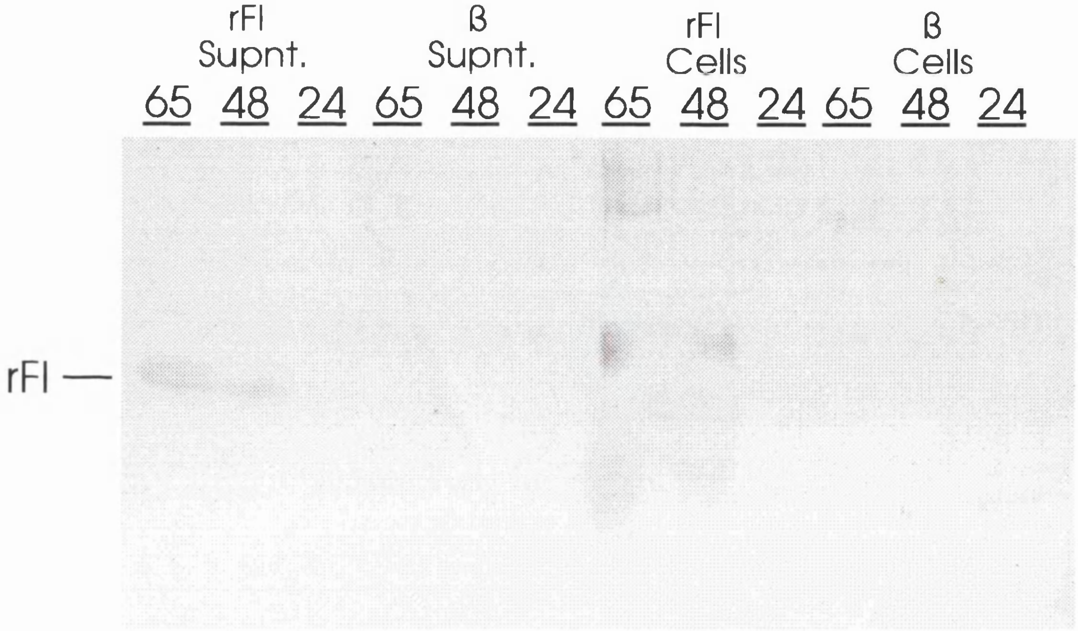


Figure 6.4: Western blot time course of recombinant factor I expression in *Sf*21 cells in serum-free media with three time points over 65h blotted with rabbit anti-factor I polyclonal antisera. Supnt. (supernatant) lanes contain cell media only, cell lanes contain cell lysates only. β refers to a negative control baculovirus expressing HLA β chain (courtesy of Dr V. C. Emery). Figure courtesy of Dr C. G. Ullman.

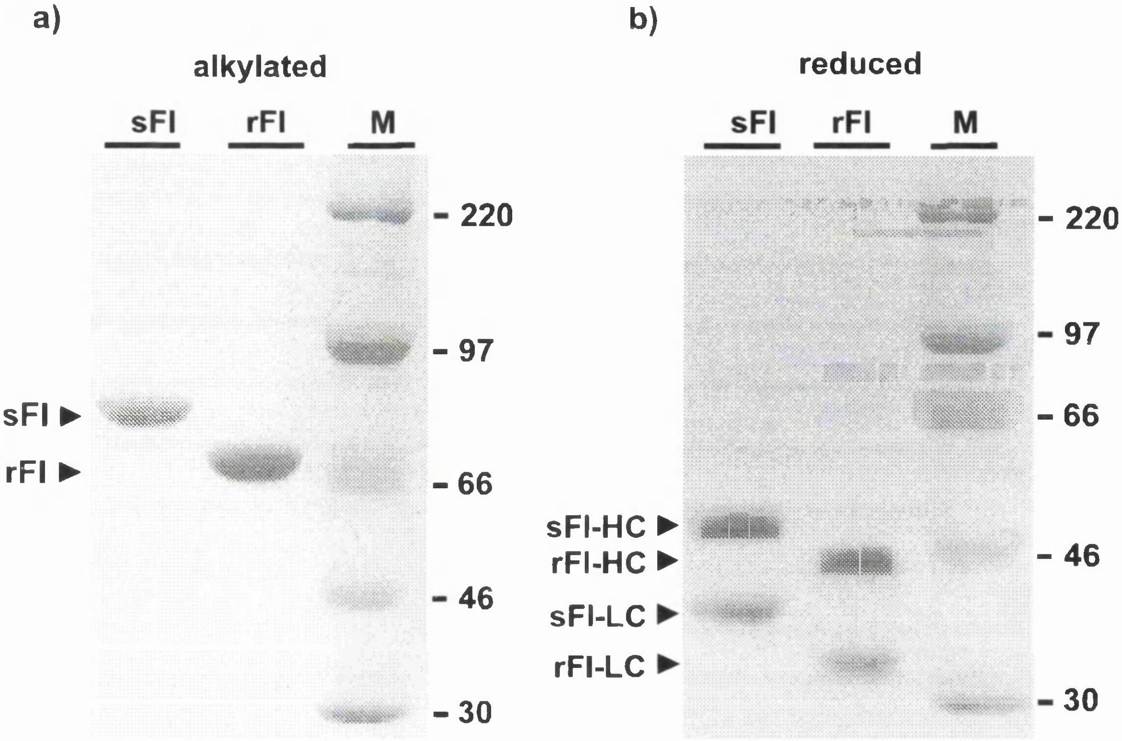


Figure 6.5: 8% SDS-PAGE analysis of factor I samples. (a) The three lanes correspond to purified sFI and rFI samples which are alkylated (non-reduced), and compared with non-reduced molecular weight markers (M) with their masses shown in kDa. (b) The three lanes correspond to purified, reduced sFI and rFI samples to show their heavy and light chains (HC and LC respectively), together with molecular weight markers. A trace amount of uncleaved rFI is observed after reduction. Figure courtesy of Dr C. G. Ullman.

(6.3.2) Post-translational Modification of Recombinant Factor I.

The SDS-PAGE analysis of Figure 6.5 showed that rFI is processed by insect cells into heavy and light chains with M_r of approximately 48,000 and 36,000 respectively, with only a trace amount of unprocessed rFI appearing at approximately 84,000. N-terminal sequence analyses showed that rFI was expressed as expected, i.e. the N-terminal sequences of the heavy and light chains were respectively KVTYT and IVGGK as in the case of sFI. The heavy and light chains of native sFI migrated with apparent M_r of 50,000 and 42,000 respectively. These values yield apparent M_r of 84,000 for rFI and 92,000 for sFI, in comparison to a calculated value of 63,300 from the polypeptide chain alone. The difference in M_r between rFI and sFI is attributable to differences in their glycosylation. Factor I contains three N-glycosylation sites on the SP domain, two on the FIMAC domain and one on the CD5 domain (Catterall *et al.*, 1987). If all six sites contain high-mannose oligosaccharide chains, each of structure $Man_9GlcNAc_2$, the predicted M_r for factor I would be 74,500 (15.0% carbohydrate w/w). If all six sites were tetra-antennary complex-type oligosaccharides, the predicted M_r for factor I would be 85,300 (25.7% carbohydrate w/w). As high glycosylation leads to overestimated M_r values by SDS-PAGE (Perkins *et al.*, 1993a), this accounts for the observed values of 84,000 and 92,000 in Figure 6.5.

Glycosylation was accordingly investigated using the lectin concanavalin A (Figure 6.6). Concanavalin A bound to both the heavy and light chains of rFI to demonstrate that both contained terminal α-D-mannosyl and/or α-D-glucosyl residues. As controls, under the same conditions, concanavalin A bound to both the heavy and light chains of sFI, but not to the bacterial fusion protein of glutathione S-transferase

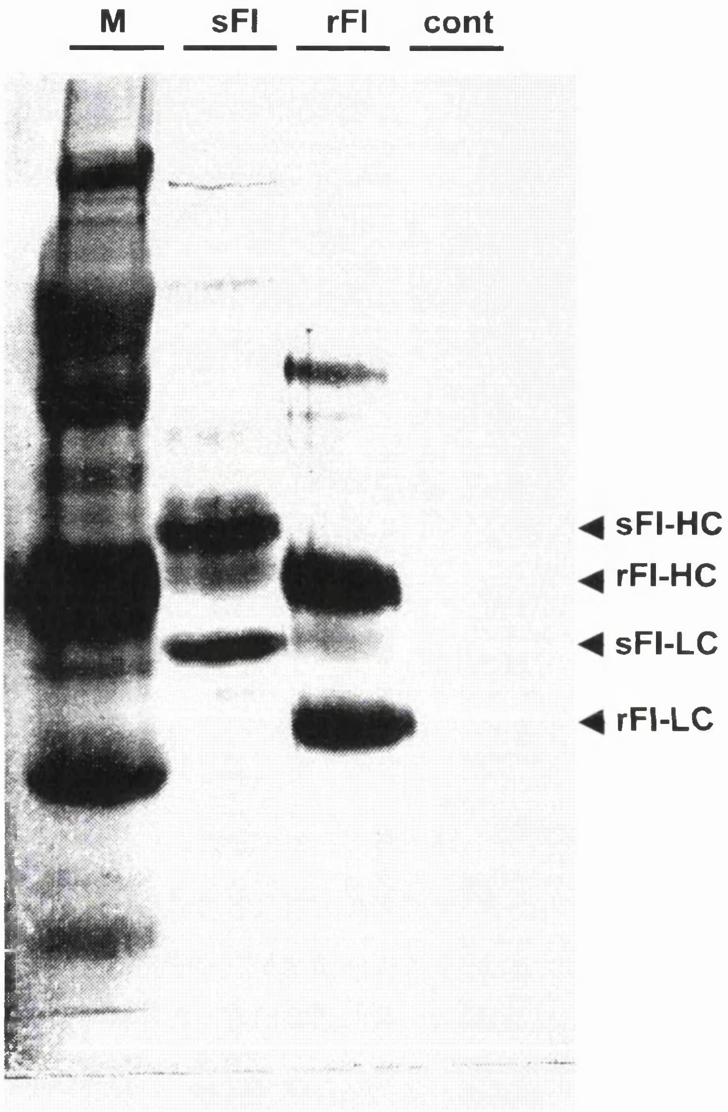


Figure 6.6: Glycosylation of reduced sFI and rFI using lectin blots. Biotinylated concanavalin A and a streptavidin-alkaline phosphatase conjugate were used for detection. Both the heavy and light chains (HC and LC respectively) of sFI and rFI bind concanavalin A, while the carbohydrate-free control (cont) based on bacterially expressed glutathione S-transferase fusion protein of the LDLr-1/2 domains in factor I does not. A small percentage of uncleaved rFI is observed. The lane with molecular weight markers with masses shown in kDa is denoted by M. Figure courtesy of Dr C. G. Ullman.

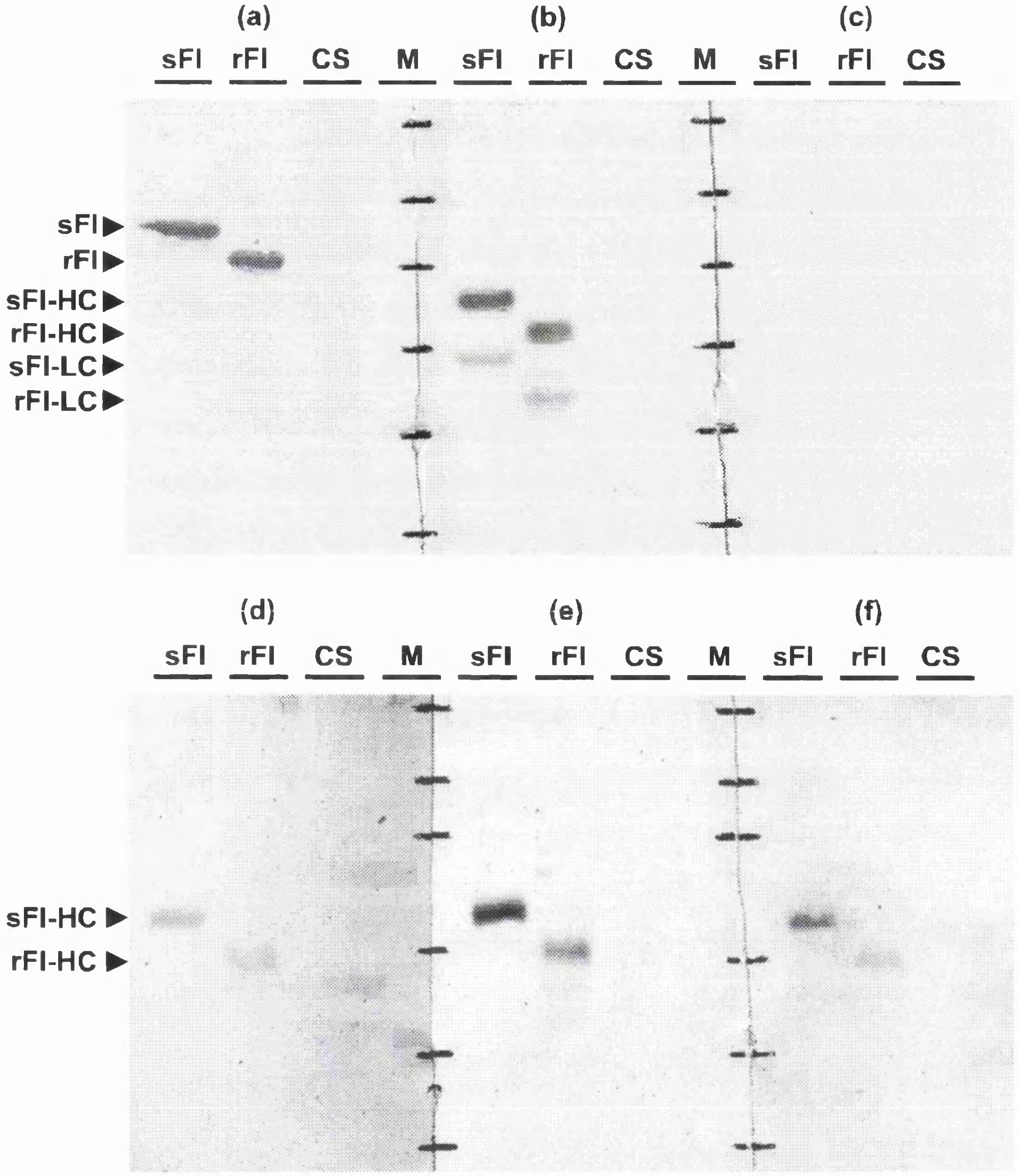


Figure 6.7: Western blot analyses of factor I samples. sFI and rFI samples were compared with the uninfected cell culture supernatant of insect cells (CS) as a negative control under both reducing and non-reducing conditions using five anti-factor I antibodies and one anti-factor H antibody. 1 μg of sFI and rFI were loaded in each lane. The sample was alkylated (non-reduced) in (a), and reduced in (b), (c), (d), (e) and (f). The antibodies were as follows: (a) MRC-OX21, an anti-factor I monoclonal antibody, (b) anti-factor I rabbit polyclonal sera, (c) MRC-OX24, an anti-factor H mouse monoclonal antibody (negative control), (d) anti-FIMAC rabbit polyclonal sera, (e) anti-LDLr-1/2 rabbit polyclonal sera, (f) anti-LDLr-2 rabbit polyclonal sera. Figure courtesy of Dr C. G. Ullman.

with the LDLr-1/2 domains of factor I. Differences in the glycosylation of rFI are expected since proteins secreted from *Trichoplusia ni* insect cells contain high mannose type N-linked oligosaccharides in which the outer-chain galactose and sialic acid residues are absent, while mannose, fructose and probably *N*-acetylglucosamine are present (Jarvis and Finn, 1995). Similar size differences in native and insect cell-expressed C3, factor H and C9 have been noted previously (Lao *et al.*, 1994; Sharma and Pangburn, 1994; Taylor *et al.*, 1994).

Western blot analysis was used to authenticate the expression of rFI. The anti-factor I monoclonal antibody MRC-OX21 recognised both rFI and sFI, but not any products in the uninfected cell culture supernatant used as a control (Figure 6.7a). The proteins were reduced in order to analyse the heavy and light chains. Use of polyclonal anti-factor I antiserum identified both the heavy and light chains of sFI and rFI (Figure 6.7b), while the monoclonal antibody MRC-OX24 (which is specific to factor H) used as a control under the same conditions showed no reaction as expected (Figure 6.7c). The heavy chain was tested by the use of polyclonal antisera raised to three bacterial fusion proteins containing the FIMAC, LDLr-1/2 and LDLr-2 domains of factor I (Ullman *et al.*, 1995; Ullman, 1994). Figures 6.7(d), 6.7(e) and 6.7(f) showed that these antibodies recognised the heavy chains of both rFI and sFI, but not their light chains.

(6.3.3) Activity of Recombinant Factor I.

The physiological reaction of factor I is to cleave C3b in the presence of factor H into the product iC3b. The activities of rFI and sFI were measured by incubation with factor H and C3(NH_3), which was generated by the thiolester cleavage of C3 using

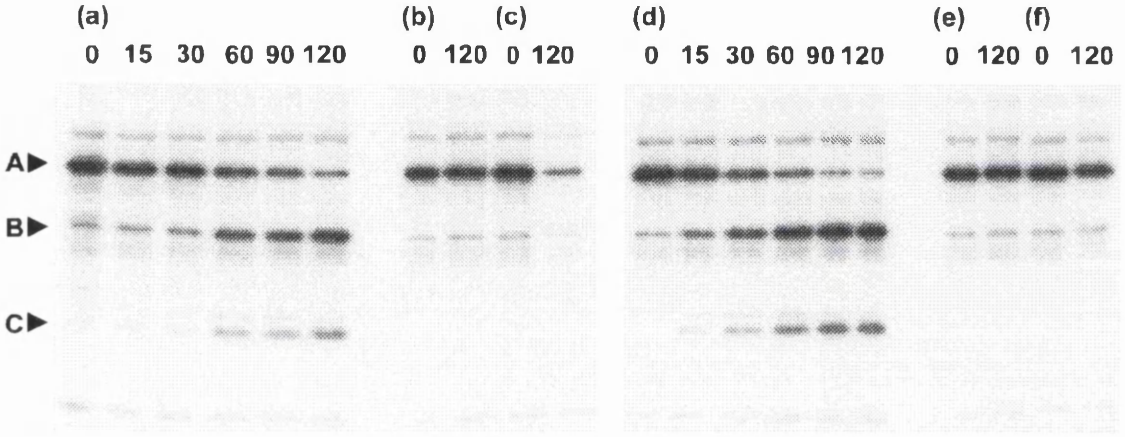


Figure 6.8: Activity assays of sFI and rFI. Gels (a) and (d) show the cleavage of $C3(NH_3)$ by rFI and sFI in the presence of factor H (FH) when monitored during 2 h using 7% SDS-PAGE gels. The band at M_r 108,000 corresponds to the α' chain of $C3(NH_3)$ is denoted by A, and decreases in intensity with time. The two bands at M_r 68,000 and 43,000 correspond to factor I-cleaved $C3(NH_3)$ and are denoted by B and C respectively. The band at 68,000 coruns with that for the β-chain of $C3(NH_3)$. The band running above band A is an irrelevant contaminant caeruloplasmin. Gels (b) and (e) are controls which contain rFI and sFI respectively with $C3(NH_3)$ but no factor H, while gel (c) is a control with only $C3(NH_3)$, and gel (f) is a control with only factor H and $C3(NH_3)$. Figure courtesy of Dr C. G. Ullman.

ammonia (Methods). The rate of breakdown of the α' chain of C3(NH_3) with an M_r of 108,000 into two fragments of M_r 68,000 and 43,000 was monitored (Sim and Sim, 1983). The two negative controls were factor H and C3(NH_3) without factor I, and C3(NH_3) alone. The autoradiographs of Figures 6.8(a) and 6.8(d) showed that, while both rFI and sFI generated cleaved fragments, rFI was slower than sFI. The use of 10 measurements in two time-course experiments showed that the activity of rFI was 55% of that for sFI. While rFI demonstrated biological activity and was therefore properly folded, the most likely cause of its reduced activity is its altered glycosylation which may affect its binding to factor H or C3(NH_3). It is known that modifications to the glycosylation of plasma proteins has variable effects upon activity, such as the increased activity of tissue-type plasminogen activator after the removal or alteration of its N-linked oligosaccharide chains (Varki, 1993).

(6.3.4) CD and FTIR Spectroscopy.

Quantitative CD measurements of the secondary structures of rFI and sFI were performed in order to compare the folding of both proteins and assess their secondary structure contents (Figure 6.9). Circular dichroism is a sensitive monitor of conformational changes in proteins (Drake, 1994). Both rFI and sFI showed CD spectra that were dominated by β-sheet structures with a strong absorption at approximately 210 nm. Their visual appearance resembled typical β-sheet rich proteins such as those of the serine proteases α-chymotrypsin and elastase with 10% α-helix and 35% β-sheet (Drake, 1994). The absence of strong signals in the region close to 220 nm suggests that there is very little α-helical conformation in either rFI or sFI. Assuming that carbohydrate makes no contribution to the CD spectrum, quantification of the CD spectrum of sFI indicated 3 ± 1% α-helix, 50 ± 9% β-sheet and 47 ± 9% coil (Table 6.1). This is in agreement with that for rFI which yielded 2% α-helix, 58% β-sheet and 40% coil, and shows that rFI has the same protein conformation as sFI. The β-sheet content in sFI and rFI is higher than those measured for α-chymotrypsin and elastase. This difference is attributable to the high disulphide content (7.1%; 40 Cys residues) of factor I when this is compared to a typical globular protein with a Cys content of 1.7%, since disulphide bridges as well as peptide bonds give rise to CD bands.

The CD analysis was supported by independent FTIR measurements of rFI and sFI in ^2H_2O buffers. After buffer subtraction, both proteins showed a broad amide I absorption band at 1640 cm^{-1} and 1639 cm^{-1} respectively (Figures 6.10a and 6.10c). These frequencies are characteristic of proteins that are predominantly β-sheet (Haris and Chapman, 1994; Haris *et al.*, 1986; Lee *et al.*, 1990). The improved resolution of the second derivative spectra revealed the fine structure of the amide I band which were similar for both rFI and sFI (Figures 6.10b and 6.10d). Thus rFI exhibited a large peak at 1636 cm^{-1} (assigned to β-sheet) and a minor peak at 1652 cm^{-1} (assigned to a small amount of α-helix), together with others at 1665 cm^{-1} (turns and bends) and 1683 cm^{-1} (assigned to the higher frequency vibration arising from antiparallel β-sheet, and possibly also from turns and bends). These band positions were very similar in sFI where absorptions were visible at 1635 cm^{-1} (β-sheet), 1652 cm^{-1} (α-helix), 1664 cm^{-1} and 1683 cm^{-1}. The dominant β-sheet features are consistent with the CD spectra of Figure 6.9. The minor peaks at 1614 cm^{-1}, 1588 cm^{-1} and 1558 cm^{-1} are assigned to bands from sidechains. Even though both FTIR spectra were similar, a peak is seen in the second derivative spectrum of sFI at 1606 cm^{-1} that is absent in rFI. As an FTIR study of sialic

	α-helix	β-sheet	Other	
Experimental values from CD spectroscopy				
rFI	2%	58%	40%	
sFI [a]	3 ± 1%	50 ± 9%	47 ± 9%	
Homologous crystal structures				Source [b]
FIMAC	9	19	39	1bmo
CD5	0	20	83	Predicted
LDLr-1	0	4	36	1ajj
LDLr-2	0	4	34	1ajj
SP	12	77	154	1lmw
Percentage total	4%	25%	70%	

a) Mean ± SD using three different preparations.

b) Brookhaven database codes for three crystal structures are given. These were analysed using DSSP software (Kabsch and Sander, 1983) to give the number of residues in each secondary structure type. The CD5 structure was derived from the mean of five secondary structure predictions (Chamberlain *et al.*, 1998; Chapter 7). The total corresponds to 491 of the 561 residues in factor I.

Table 6.1: Secondary structure content of the domains in factor I.

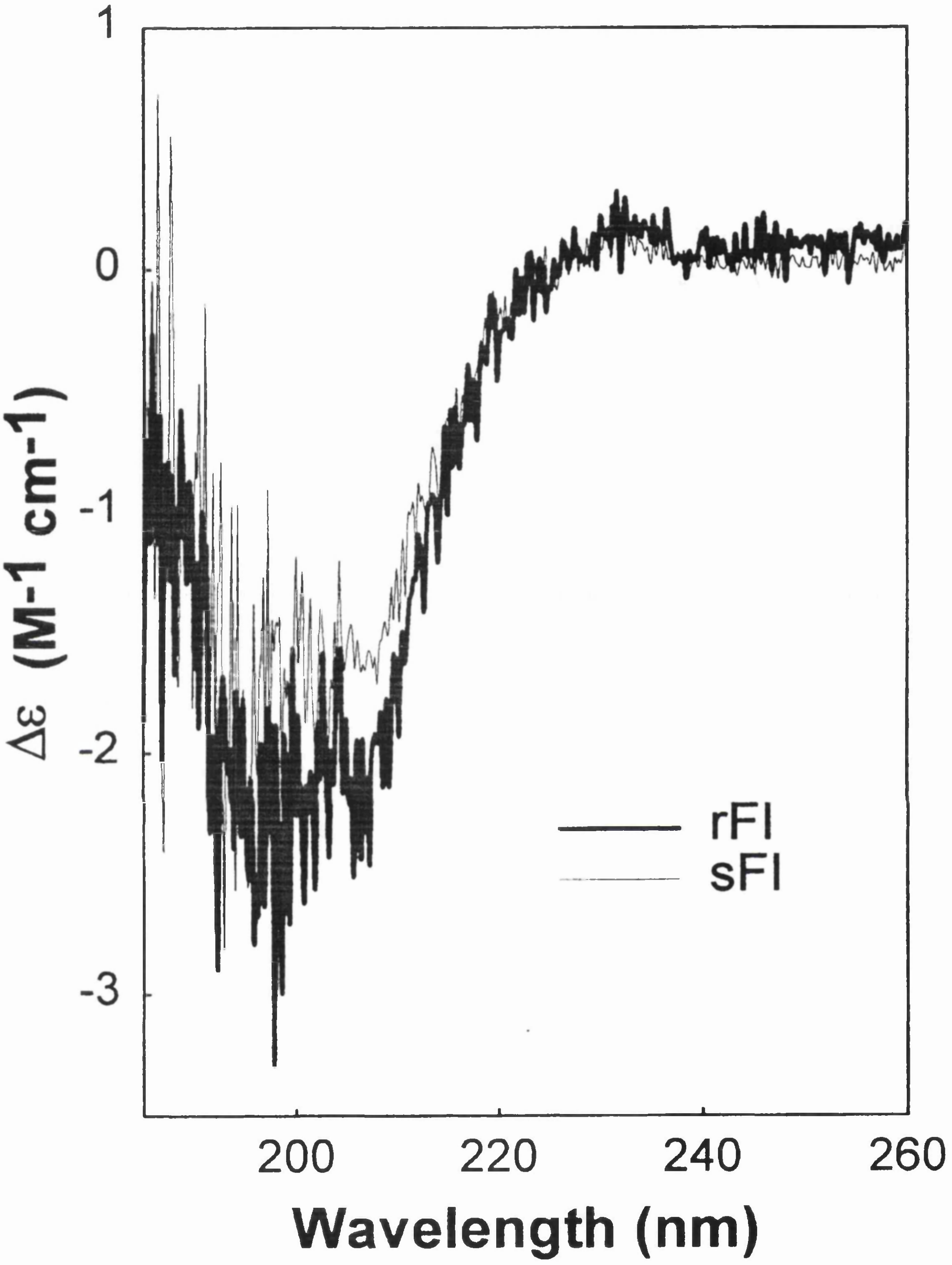


Figure 6.9: CD spectroscopy of rFI and sFI. The spectra of rFI and sFI are denoted by thick and thin lines respectively and are the average of 4 scans each at 20°C, using a buffer of 5 mM potassium phosphate, pH 7.0. The concentration of rFI was 0.24 mg/ml and that of sFI was 0.39 mg/ml. Figure courtesy of Dr C. G. Ullman.

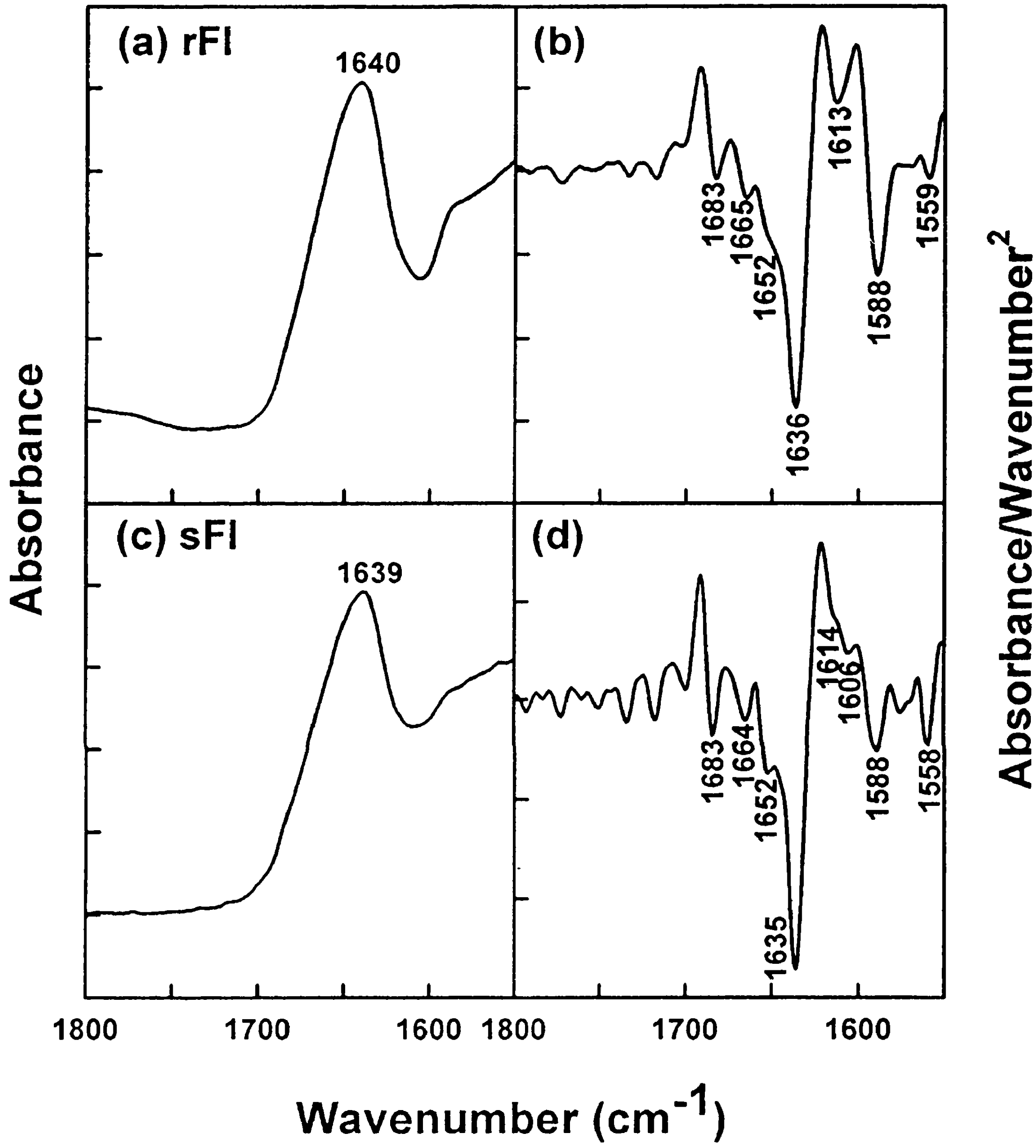


Figure 6.10: FTIR spectroscopy of the amide I band of rFI and sFI. (a,b) rFI and (c,d) sFI was studied in phosphate buffered saline in ^2H_2O. The absorbance spectra are shown in (a,c) and the second derivative spectra are shown in (b,d). Peak positions that are discussed in the text are identified by their wave numbers. The concentration of rFI was 2.0 mg/ml and that of sFI was 3.3 mg/ml. Figure courtesy of Dr C. G. Ullman.

acid showed an absorbance band at 1608 cm^{-1} (Knörle *et al.*, 1994), this peak can be attributed to the presence of sialic acid in the complex-type oligosaccharides of sFI that was absent from the high mannose oligosaccharides in rFI.

The CD and FT-IR spectroscopic data can be compared with secondary structures for the five domains in factor I. These are well understood for the FIMAC, LDLr and SP domains, where multiple sequence alignment analyses resulted in β-sheet secondary structure predictions (Ullman and Perkins, 1997; Ullman *et al.*, 1995; Perkins and Smith, 1993) that agreed well with those found in crystal structures for the follistatin, LDLr and SP domains (Table 6.1; Kabsch and Sander, 1983a; Hohenester *et al.*, 1997; Fass *et al.*, 1997; Spraggon *et al.*, 1995). Since averaged secondary structures from multiple sequence alignments can be predicted with accuracies as high as 77-81% (Edwards and Perkins, 1996; Rost and Sander, 1996), the β-sheet prediction for the CD5 domain can be included (Chamberlain *et al.*, 1998; Chapter 7). The summation of the five individual secondary structures gave a content of 4% α-helix and 25% β-sheet (Table 6.1) that confirms the dominance of β-sheet structures in factor I in accordance with spectroscopy, although the agreement is qualitative in keeping with general experience of CD (Drake, 1994).

(6.4) Conclusions.

Factor I is essential for complement regulation, in which it has a limited substrate specificity to C3b and C4b, and its full activity requires cofactor binding. A recombinant expression system is needed to understand the molecular basis for these interactions. Here, we have shown that a baculovirus expression system is able to produce adequate amounts of active factor I and overcomes difficulties encountered with earlier ones (Goldberger *et al.*, 1987; Wong *et al.*, 1995). Thus, in contrast to rFI synthesised in mammalian cells, rFI in the insect cell system used here is processed into two chains. In comparison with similar insect cell expression systems for recombinant factor H, C3 and C9 which are also secreted, glycosylated and processed into multichain proteins (Lao *et al.*, 1994; Sharma and Pangburn, 1994; Taylor *et al.*, 1994), rFI is expressed at a comparable level into a two-chain active form with only trace amounts of uncleaved factor I, whereas a certain amount of pro-C3 remained in the C3 system. Its glycosylation is different from the native protein, yet it retains native function at the level of 55% of that for sFI. This parallels similar results for recombinant factor H, C3 and C9, however all three recombinant proteins showed 90-100% biological activity. As protein sequencing and CD and FT-IR spectroscopy showed no difference in the protein structure of sFI and rFI, the reduced activity of rFI is attributable to differences in glycosylation. For example, the absence of sialic acid in rFI may affect activity since the interactions between factor I and factor H and between factor I and C3(NH_3) are mainly ionic (Soames and Sim, 1997). Alternatively the change in glycosylation between sFI and rFI may affect the separation between the heavy and light chains in the three-dimensional structure of factor I to modify its activity (Varki, 1993; Chamberlain *et al.*, 1998; Chapter 7).

The routine preparation of rFI from 2 l cell cultures has provided sufficient material for CD and FTIR spectroscopy and this showed that rFI contained a dominantly β-sheet secondary structure that is expected from related crystal structure and structure prediction analyses (Table 6.1). Likewise adequate amounts of rFI were successfully purified for X-ray and neutron scattering to permit study of the domain arrangement in rFI and sFI (Chamberlain *et al.*, 1998). Mutagenesis of recombinant factor H, C3 and C9 have resulted in an improved understanding of the immunological function of these proteins (Taylor *et al.*, 1994; Sharma and Pangburn, 1996, 1997; Lambris *et al.*, 1996). Mutants of factor I can now be prepared for similar work, and this work can be rationally planned on the basis of the related crystal structures listed in Table 6.1.

CHAPTER 7

MOLECULAR MODELLING OF THE DOMAIN STRUCTURE OF COMPLEMENT FACTOR I

(7.1) Introduction.

Factor I of the complement system of immune defence is a five-domain serine protease which is involved in the regulation of the C3 convertase of the classical or alternative pathways of activation (Sim *et al.*, 1993; Law and Reid, 1995; Chapter 6). It specifically cleaves the α' chains of C3b and C4b into smaller fragments in the presence of the cofactor proteins factor H or C4b-binding protein respectively. Other cofactors for factor I-mediated cleavage of C3b and C4b include complement receptor type 1 (CR1, CD35) and the membrane cofactor protein (MCP, CD45). Unusually factor I is not inhibited by any known plasma protease inhibitors, and is specific only for C3b and C4b. The importance of factor I is demonstrated in deficiencies that lead to the excessive consumption of C3 and recurrent pyogenic infections (Vyse *et al.*, 1996).

Human factor I contains a heavy chain with 317 residues (including 27 Cys residues), and a catalytic light chain with 244 residues (including 11 Cys residues) (Catterall *et al.*, 1987; Goldberger *et al.*, 1987). An 18-residue signal sequence and a 4-residue RRKR linker between the heavy and light chains are removed during processing. Three glycosylation sites are present on each of the heavy and light chains (Figure 7.1), giving a total molecular weight of 85,000 and a glycosylation of 26% (w/w) (Perkins *et al.*, 1993). The four domains in the heavy chain (Figure 7.1) are the factor I/membrane attack complex (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. Additional 24-residue and 32-residue sequences are present at the N-terminus and C-terminus of the heavy chain, the latter of which is strongly species dependent (Kunnath-Muglia *et al.*, 1993; Minta *et al.*, 1996). The light chain contains a serine

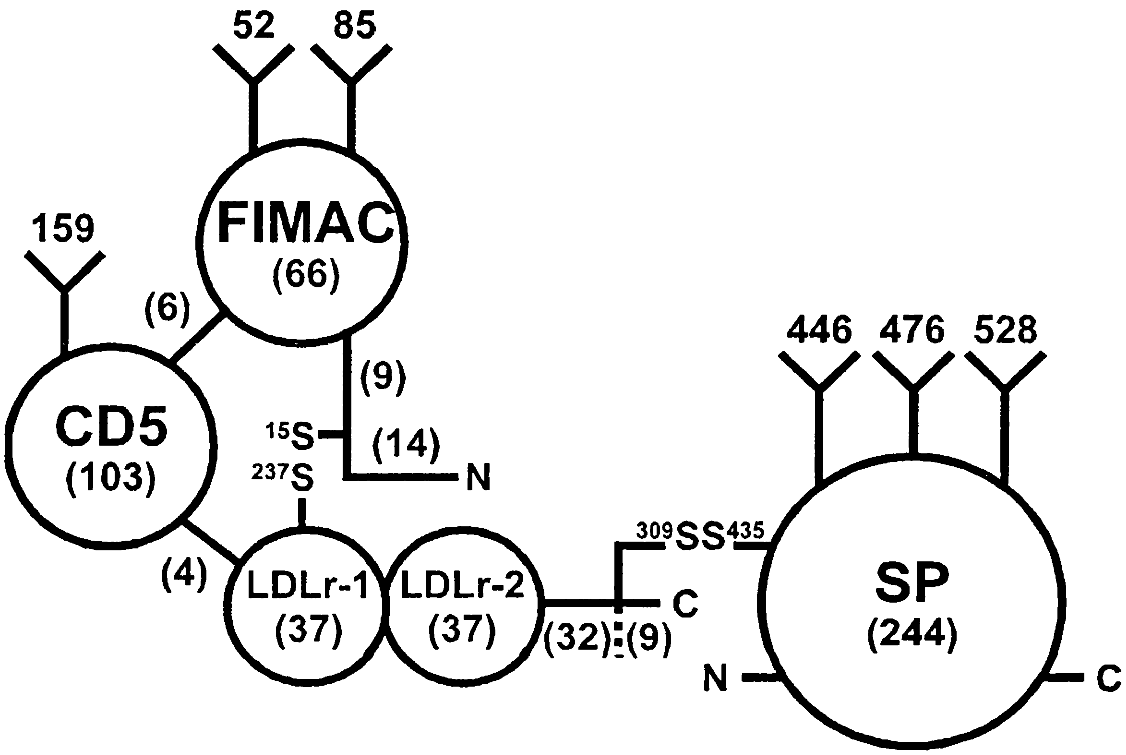


Figure 7.1: Domain structure of factor I. This is constructed from the FIMAC, CD5, LDLr-1, LDLr-2 and SP domains. Cys15-Cys247 are postulated to be bridged to link the FIMAC, CD5 and LDLr-1 domains in a triangular arrangement. The heavy and light chains are linked by Cys309-Cys435. Putative N-linked oligosaccharide chains are located at Asn52, Asn85 and Asn159 in the heavy chain and at Asn446, Asn476 and Asn528 in the light chain. There are 561 residues in processed factor I. The number of residues in each part of factor I is indicated in brackets.

protease (SP) domain which is disulphide-linked to the heavy chain. To appreciate the function of the five domains in factor I, we have expressed and characterised recombinant factor I (rFI) in a baculovirus system (Ullman *et al.*, 1998). rFI was determined to be folded correctly, but differed in its molecular weight from serum-derived factor I (sFI). This is attributable to the replacement of the complex-type oligosaccharide chains by high mannose-type chains in baculovirus (Jarvis and Finn, 1995).

The three-dimensional arrangement of the domains in factor I is poorly understood. X-ray and neutron scattering studies of factor I in solution had shown that its overall length is between 12.8-15 nm, and electron microscopy of factor I *in vacuo* stained with uranyl acetate had shown that factor I was 13 nm in length and bilobal, however no molecular explanation of these findings had been provided (Perkins *et al.*, 1993; DiScipio, 1992). Since that time, solution scattering has been improved by the establishment of a procedure to calculate scattering curves from known crystal structures (Smith *et al.*, 1990; Perkins *et al.*, 1993a; Ashton *et al.*, 1997). A new automated curve-fit procedure employs constraints based on homologous atomic structures for individual domains and their known covalent connectivity to yield molecular models for domain arrangements in the intact protein (Mayans *et al.*, 1995; Beavil *et al.*, 1995; Boehm *et al.*, 1996; reviewed in Perkins *et al.*, 1998). This constrained modelling method is now applicable to factor I as the result of (i) recently-determined crystal structures that are relevant for the FIMAC, LDLr and SP domains of factor I and (ii) the recently-determined disulphide bridges in the globular structure of the CD5 domain (Hoehnester *et al.*, 1997; Ullman and Perkins, 1997; Fass *et al.*, 1997; Spraggon *et al.*, 1995; Resnick *et al.*, 1996). Here, homology modelling suggested that the FIMAC, CD5 and LDLr-1 domains form a compact triangular arrangement stabilised by a disulphide bridge between Cys15-Cys237. Together with the SP domain, this defined two globular entities within a bilobal structure for factor I. This structure was tested using automated X-ray and neutron curve fit analyses of 9600 bilobal structures for sFI. Since solution scattering is sensitive to oligosaccharide conformations (Boehm *et al.*, 1996), independent X-ray and neutron curve fits with rFI were performed to take advantage of the different oligosaccharide structure in rFI. The functional significance of the resulting domain structure is discussed.

(7.2) Materials and Methods.

(7.2.1) Expression and Purification of Factor I for Scattering.

Four preparations of sFI were obtained from 0.6 l outdated human plasma for each one (Sim *et al.*, 1993; Ullman *et al.*, 1998). Using separate apparatus to avoid cross-contamination with sFI, four preparations of purified active rFI were obtained from a recombinant baculovirus expression system using 2-3 l of culture for each one. Full details will be presented elsewhere (Ullman *et al.*, 1998; Chapter 6). Samples were stored frozen at -20°C. When needed for scattering, samples were subjected to gel filtration to remove trace aggregates using a Superdex-200 Prep grade column (1.6 × 60 cm) (Pharmacia) and stored at 4°C. For X-ray scattering and neutron scattering in H_2O buffers, samples were dialysed into Dulbecco's phosphate buffered saline at pH 7.0 (137 mM NaCl, 2.7 mM KCl, 8.1 mM Na_2HPO_4, 1.5 mM KH_2PO_4) (Sigma) together with 0.1 mM Pefabloc-SC (Pentapharm) and 0.5 mM EDTA. For neutron scattering, the samples were dialysed as above but now using ^2H_2O buffers with four buffer changes over 36 h at 6°C. Before and after data collection, samples were checked by SDS-PAGE. Fourier transform infrared and circular dichroism spectroscopy was performed on rFI and sFI to verify the similarity of their folded protein structures (Ullman *et al.,* 1998; Chapter 6).

To assay for free Cys residues in factor I, the use of Ellman's reagent (5,5'-dithio-bis(2-nitrobenzoic acid)) with 1.5 nmol of sFI in 0.1 M sodium phosphate pH 7.3, using 0.5 - 10 nmol of reduced glutathione as a reference, detected less than 0.3 free thiol groups per sFI. In 5 M guanidine, similar results were obtained for unfolded sFI and rFI. In addition, the use of 2 μCi ^{14}C-labelled iodoacetamide with 20 μg sFI denatured with 8 M guanidine in 0.2 M Tris, pH 8.2 confirmed this. A second sFI sample was previously incubated with 0.5 mM iodoacetamide for 10 min at 37°C followed by dialysis (negative control). A third sFI sample was previously incubated with 40 mM dithiothreitol for 2 h in the dark (positive control). The three samples were dialysed extensively and counted for bound radioactivity. Significant radioactivity was present only in the dithiothreitol-treated sample.

(7.2.2) <u>X-ray and Neutron Scattering Data Collection.</u>

X-ray data were obtained in one beam session at Station 2.1 using a camera with a quadrant detector at the Synchrotron Radiation Source, Daresbury, U. K.. A sample-detector distance of 3.58 m was used, with beam currents of 154-184 mA and a storage ring energy of 2.0 GeV. This resulted in a usable Q range of 0.1 to 2.3 nm^{-1} ($Q = 4 \pi \sin \theta / \lambda$; scattering angle = 2θ; wavelength = λ). Samples were measured between 1.7-3.3 mg/ml at 15°C in cells of path thickness 1 mm with mica windows. Data acquisition times were 10 min, obtained as 10 time frames of 1 min each in order to confirm the absence of radiation damage. Neutron data were obtained in one session on Instrument D22 at the Institut Laue-Langevin, Grenoble, which is analogous to Instrument D11 (Lindner, *et al.*, 1992). Sample-detector distances of 1.4 m and 5.6 m were used. Using λ of 1.00 nm, a 64 × 64 cm detector, and a rectangular beam aperture of 7 × 10 mm, the two detector positions resulted in a usable Q range of 0.06 to 2 nm^{-1}. Samples were measured between 0.3-1.4 mg/ml at 15°C in rectangular quartz Hellma cuvettes of path length 2 mm for acquisition times between 10-30 min. Neutron data were also obtained in two sessions on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, U. K., using a proton beam current of 170-185 μA to generate neutrons. Based on a fixed sample-detector distance of 4.3 m, the usable Q range was 0.1 to 2.0 nm^{-1}. For concentrations between 1.7-4.3 mg/ml, the data acquisition time was 3-4 h at a sample temperature of 15°C. Other details, data reduction, and references are given elsewhere (Ashton *et al.*, 1997; Chapter 2).

(7.2.3) Guinier and Distance Distribution Function Analyses.

Guinier analyses at low Q gave the radius of gyration R_G and the forward scattering at zero angle I(0) (Glatter and 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 value is a measure of structural elongation if the internal inhomogeneity of scattering densities has no effect. The I(0)/c values (where c is the sample concentration) leads to molecular weights M_r. For elongated macromolecules, the mean radius of gyration of the cross-section R_{xs} and the mean cross-sectional intensity at zero angle $[I(Q).Q]_{Q\text{-}0}$ (Hjelm, 1985) were obtained from:

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

Combination of the R_G and R_{XS} analyses lead to triaxial dimensions (Perkins *et al.*, 1993). Indirect transformation of the scattering data in reciprocal space I(Q) into that in real space P(r) was performed using GNOM (Semenyuk and Svergun, 1991).

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

P(r) corresponds to the distribution of distances r between volume elements, from which the R_G and I(0) values can be determined as well as the maximum dimension L. A range of assumed maximum lengths for sFI and rFI were tested to optimise the calculation of the P(r) curve (Ashton *et al.*, 1997).

(7.2.4) Homology Modelling of the Domains of Factor I.

Homology models for the FIMAC, LDLr and SP domains were constructed using the sequence alignment of Figure 7.2 and INSIGHT II 95.0, BIOPOLYMER, HOMOLOGY and DISCOVERY software (Biosym/MSI, San Diego, U.S.A.) on Silicon Graphics INDY Workstations. Loops were built using the pdb_select.1995-jun-01 database derived from 349 crystal structures at 0.2 nm resolution or better (Hobohm and Sander, 1994; Hobohm *et al.*, 1992). Energy refinements were based on the consistent valence force field. Iterations were made using combinations of the steepest descent and conjugate algorithms to improve the connectivity of the model and minimize bad contacts or stereochemistry. Models were stereochemically verified using PROCHECK (Laskowski *et al.*, 1993). Solvent accessibilities were calculated using COMPARER (Lee and Richards, 1971; Šali and Blundell, 1990).

```
              1              ▼                                         50
Factor I      KVTYTSQEDL VEKKCLAKKY THLSCDKVFC QPWQRCIE.. ...GTCVCKLPYQCP
SPARC         .......... .......... ...PCQNHHC KHGKVCELDE NNTPMCVCQDPTSCP
                                                  <–B1>      <-B2>
             51   cho                                         cho           100
Factor I      K...NGTAVCATN RRSFPTYCQQ KSLECLHPGT ......KFLNNGTCTA EGKFSVSLKH
SPARC         APIGEFEKVCSND NKTFDSSCHF FATKCTLEGT KKGHKLHLDYIGPCK. ..........
                      <B3     B4 <--A1-->         <-A2>  <B5>
            101                                                  150
Factor I      GNTDSEGIVE VKLVDQDKTM FICKSSWSMR EANVACLDLG FQQGADTQRR

            151       ch.o                                       200
Factor I      FKLSDLSINS TECLHVHCRG LETSLAECTF TKRRTMGYQD FADVVCYTQK

            201              *         ●   ●   *  ●● ▼           250
Factor I      ADSPMDDFFQ CVNGKYISQM KACDGINDCG DQSDELCCKA CQGKGFHCKS
LDLr-5        ..PCSAFEFH CLSGECIHSS WRCDGGPDCK DKSDEENCAP CSAFEFHCLS

            251*         ○   ●  *   ●●                           300
Factor I      GVCIPSQYQC NGLVDCITGE DEVGCAGFAS VAQEETEILT ADMDAERRRI
LDLr-5        GECIHSSWRC DGGPDCKDKS DEENCA.... .......... ..........

            301              317     322                              350
Factor I      KSLLPKLSCG VKNRMHIrrk rIVGGKRAQL GDLPWQVAIK DA.....SGITCGGI
Plas Act              CG QKTLRP.     IIGGEFTTI ENQPWFAAIY RRHRGGSVTYVCGGS
                                                     <--B ->     <C>
            351           #                                      400
Factor I      YIGGCWILTA AHCLRASKTH RYQIWTTVVD WIHPDLKRIV IEYVDRIIFH
Plas Act      LMSPCWVISA THCFIDYPKK EDYIVYLGRS RLNSNTQGEM KFEVENLILH
                   <D->   #           <-E->             < ---F--->
            401            #                                        cho50
Factor I      ENYNAG..TYQN DIALIEMK.KD GNKKDCELP.R SIP.ACVPWSP YLFQPNDTCI
Plas Act      KDYSADTLAHHN DIALLKIRSKE GR...CAQPSR TIQTICLPSMY NDPQFGTSCE
                           # <--H-->                                  <--
            451                            c..ho                      500
Factor I      VSGWGREK..DN ERVFSLQWGE VKLISN..CSK.F YGNRFYEKEM ECAGTYDGSI
Plas Act      ITGFGKENSTDY LYPEQLKMTV VKLISHRECQQPH YYGSEVTTKM LCAADPQWKT
              J->               <--- K-->                   < -L->
            501     #           cho                              550
Factor I      DACKGDSGGP LVCMDANNVT YVWGVVSWGE NCGKPEFPGV YTKVANYFDW
Plas Act      DSCQGDSGGP LVCSLQGRMT LT.GIVSWGR GCALKDKPGV YTRVSHFLPW
                    #  < –M-->  <-- --N---->            < O>
            551            565
Factor I      ISYHVGRPFI SQYNV
Plas Act      IRSHTKEE
```

Figure 7.2: Sequence alignment used for homology modelling of factor I. The numbering is that of human factor I. The domain sequences are aligned with those corresponding to crystal structures for human SPARC (code 1bmo), the fifth domain of the human low density lipoprotein receptor (LDLr-5; 1ajj), and human plasminogen activator (Plas Act; 1lmw). Underlined sequences do not correspond to known homologous crystal structures. Symbols above the factor I sequence denote the locations of two predicted exposed unpaired Cys residues (▼: see text), the four acidic residues that constitute a potential Ca^{2+} binding site in the LDLr domain (●; ○ when not conserved; * other related residues) and six putative N-linked oligosaccharide sites (cho). The catalytic triad at His362, Asp411 and Ser507 is marked by #.

The FIMAC model was constructed using residues 54-159 in the crystal structure of SPARC (Brookhaven code 1bmo: Hoehnester *et al.*,1997; Ullman and Perkins, 1997). Using the rigid body fragment assembly method, nine structurally conserved regions based on α-helix and β-sheet residues and Cys residues (total of 27 residues) and eight designated loops (total of 27 residues) were defined. Three loops (11 factor I residues 37-40, 50-53 and 79-81) that correspond to deletions were constructed using database searches. A total of 24 conformationally unassigned N-terminal residues were added to the FIMAC model using the end repair command to represent the N-terminus of factor I, while 6 additional residues were added at the C-terminus as a designated loop based on the 1bmo structure to facilitate the connection of the FIMAC and CD5 domains (Figures 7.1 and 7.2). Energy refinements were performed at the six loop splice junctions, then the five disulphide bridges were created. The final energy refinements were performed on the sidechain atoms of mutated residues in the structurally conserved regions, the sidechain atoms of both types of loop residues, and the added N- and C-terminal residues. The secondary structure backbone was retained by fixing the mainchain atoms in the conserved regions, and tethering these in the loop regions. The length of this model is 4.14 nm (Ser24-Thr89).

The globular CD5 structure visualised by electron microscopy (Resnick *et al.*, 1996) was represented by an immunoglobulin fold of the same size. Residues 3-107 in the V_L domain of human IgG1 HIL (code 8fab; chain A) were used because their total was close to the 102 residues found in the CD5 superfamily (see below). Since the construction of a Debye sphere model required the correct amino acid volume (see below), the IgG1 HIL sequence was directly replaced by that of CD5 (Figure 7.2). The structure was treated as a designated loop with the exception of a reconstruction of a Trp residue as a searched loop for steric reasons. Energy refinements were performed on all atoms to minimise bad contacts.

The LDLr-1 and LDLr-2 domains were constructed as a double-domain structure LDLr-1/2 using residues 4-40 in the crystal structure of LDLr (code 1ajj; Fass *et al.*, 1997; Ullman *et al.*, 1995). Since no gaps or insertions occurred in the sequence alignment, and the linker region between the two domains is flanked by Cys247 and Cys250 with no additional residues between them (Figure 7.2), both were combined into a single structure in which Pro4 of the second domain was changed into Ala4 for steric reasons. An extended link between the two domains was created, since (as shown by Ramachandran plots) less extended links were disfavoured for reason of steric obstruction between the two domains. The sidechains were replaced with those in factor I. The connection with the CD5 domain was facilitated by adding 6 N-terminal linker residues to LDLr-1 by an end repair. Energy refinements were performed as for the FIMAC domain, all residues being considered as a structurally conserved region. The length of this double-domain model is 4.25 nm (Ser203-Ala276).

The SP domain was constructed by Dr C. G. Ullman from residues 1-245 in the two-chain crystal structure of human urokinase-type plasminogen activator (code 1lmw: Spraggon *et al.*, 1995). The light chain of factor I was based on 12 conserved regions corresponding to the 12 β-strands and single α-helix of a SP domain (Perkins and Smith, 1993) and 12 Cys residues (total of 201 residues) and 3 designated loops (total of 11 residues). Nine searched loops (factor I residues 419-425, 439, 443, 458-461, 475-476, 479-481 and 523; a total of 25 residues) to correspond to insertions and deletions were constructed from database searches. Five disulphide bridges were created. The C-terminus of the heavy chain (residues 309-317) was modelled independently, then combined with the light chain by creating the Cys309-Cys435 disulphide bridge. Energy refinements were performed at the 18 loop splice junctions, then on the sidechain and mainchain residues as for the FIMAC domain except that the catalytic triad (His362, Asp411 and Ser507) was fixed in position. Seven unassigned residues were added at the C-terminus by an end repair.

The oligosaccharide chains (Figure 7.3) were modelled on the nine-residue structure in the Fc fragment of human IgG1 KOL (code 1fc1; Deisenhofer, 1981), to which extra residues were added to generate a tetraantennary complex-type structure $NeuNAc_4Gal_4Man_3GlcNAc_6$. The high mannose-type structure was formed by an adaptation of this structure to $Man_7GlcNAc_2$. The chains were positioned at Asn residues in extended conformations from the protein surface (Figure 7.1).

(7.2.5) Construction of Extended and Bilobal Domain Structures in Factor I.

The linear extended domain model for factor I was formed by positioning the long axes of the FIMAC, CD5 and LDLr-1/2 models and the N-terminal Cys308-Met315 fragment of the SP model on a common axis in arbitrary rotational orientations about this common axis. Bent domain models without steric overlap between the models were created from this by manually rotating the FIMAC and CD5 models about the N-terminal α-carbon atom of the LDLr-1/2 model as origin.

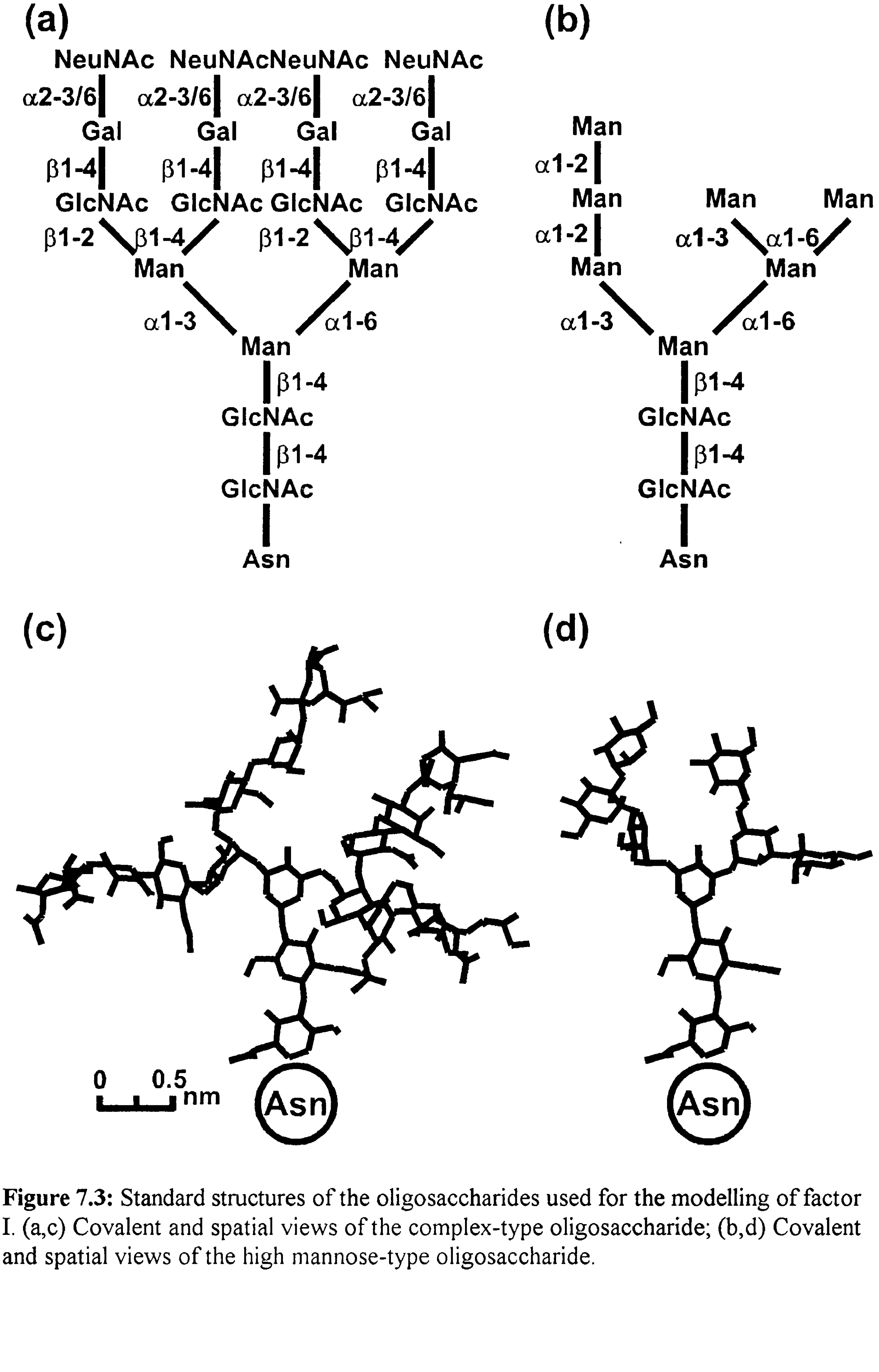


Figure 7.3: Standard structures of the oligosaccharides used for the modelling of factor I. (a,c) Covalent and spatial views of the complex-type oligosaccharide; (b,d) Covalent and spatial views of the high mannose-type oligosaccharide.

In the bilobal domain model for factor I, a triangular domain arrangement to represent one lobe was created by positioning the FIMAC and LDLr-1/2 domains in order to create a disulphide bridge between Cys15-Cys237, and inserting the CD5 model into the gap of 2.7 nm between the α-carbon atoms at Thr89 (C-terminus of the FIMAC domain) and Ser203 (N-terminus of the LDLr-1/2 domains). Cartesian axes were assigned with the origin set as the α-carbon atom of Lys249 (LDLr-2), the X-axis defined by the C2 atom of the GlcNAc residue on one of the four antennae of the oligosaccharide at Asn159, and the Y-plane defined by the α-carbon atom of Gln258 (LDLr-2). This set the longest axis of the triangular model as the X-axis, which is approximately equivalent to the long axes of the FIMAC and LDLr-1/2 domains. For the SP model used to represent the other lobe, the origin was set as the C-terminal α-carbon atom of Val565 (Figure 7.2), and the X-axis and Y-plane were defined by the α-carbon atoms of Lys368 and Ser367 respectively. This defined the X-axis of the SP domain so that the three oligosaccharide chains were located in a single Z-Y quadrant that was easily moved using 90° rotations about the X-axis.

Automated conformational searches optimised the best relative position of the two lobes in bilobal models in factor I (Perkins *et al.*, 1998). Full models were created using translations of the SP model relative to the triangular model using an INSIGHT MSI/Biosym Command Language (BCL) macro in conjunction with Unix shell scripts. The origins of the two lobe models were positioned 9.66 nm apart, and the X-axes and the X-Z planes of both lobe models were set parallel to each other. The SP model was translated in 20 × 0.5 nm steps along the Z-axis and in 30 × 0.5 nm steps along the X-axis for each of four 90° X-axis orientations of the triangular model to create 4 × 600 factor I models. The translations positioned the SP model on all sides of the triangular model as well as passing through it, while the four X-axis rotations of the triangular model explored the consequence of the asymmetric positioning of oligosaccharides on it. Other searches were performed for three further 90° Y-axis rotations of the triangular model which were equivalent to parallel X-axis rotations of the SP model. This gave a final total of 4 × 4 × 600 = 9,600 bilobal models to provide a comprehensive test of possible structures.

(7.2.6) Scattering Curve Modelling of Factor I.

The Debye sphere models for calculating scattering curves were derived using standard procedures. Each coordinate model was placed in a three dimensional grid of cubes of side length 0.3775 nm. By varying a cutoff scheme, a sphere of the same volume as the cube was created at the centre of each cube if a specified number of atoms were present in the cube. The cutoff was based on the constraint that the total volume of the spheres equalled that of the 561 amino acid and 102 complex-type carbohydrate residues in sFI (Perkins, 1986), and this enabled the unknown coordinates for the C-terminal 32 residues of the heavy chain (Figure 7.1) to be disregarded. The sFI models contained about 1904 spheres (102.5 nm^3). Models that incorporated the 54 high mannose-type carbohydrate residues in rFI were obtained using cubes of side length 0.371 nm and contained about 1751 spheres (89.4 nm^3). The X-ray and neutron scattering curves I(Q) were calculated from the Debye sphere models using SCT (Perkins and Weiss, 1983). Dry models were used for neutron curve fits, while a hydration of 0.3 g H_2O/g glycoprotein was used for X-ray curve fits (Smith *et al.*, 1993; Perkins *et al.*, 1993a; Ashton *et al.*, 1997). As no significant scattering density differences between protein and carbohydrate were observed in the curve fits, single-density spheres were used in modelling (Perkins *et al.*, 1993; Ashton *et al.*, 1997; Boehm, *et al.*, 1996). For X-ray fits, no corrections were applied for wavelength spread or beam divergence as these are considered negligible. For neutron fits for D22 and LOQ data, beam corrections were performed as described in Mayans *et al.*, (1995). The R_G and R_{XS} values were calculated from Guinier fits of the modelled curves in the same Q ranges used for experimental data. The sFI models were filtered to retain those for which 3.9 nm ≤ R_G ≤ 4.3 nm, 1.55 nm ≤ R_{XS} (X-rays) ≤1.85 nm, and 1.4 nm ≤ R_{XS} (neutrons) ≤1.6 nm (Table 7.1), and at least 95% of the expected total of 1810 spheres were present. The R-factor goodness-of-fit parameter for curve fits was defined by analogy with crystallography, for which I(0) was set as 1000 (Smith *et al.*, 1990).

(7.3) <u>Results and Discussion.</u>

(7.3.1) <u>X-ray and Neutron Scattering Data on sFI and rFI.</u>

X-ray and neutron scattering were used to compare the domain structures of sFI and rFI. The X-ray data visualise a hydrated structure in a high positive solute-solvent contrast, while the neutron data visualise a dry structure in a high negative solute-solvent contrast. These opposite contrasts act as a control for large internal density effects that may be caused by the 26% and 15% carbohydrate contents of sFI and rFI respectively (Boehm *et al.*, 1996; Perkins, 1986). X-ray and neutron studies of sFI and rFI also compared the effect of the replacement of the complex-type oligosaccharides in sFI with high mannose-type ones in rFI that is established to occur in baculovirus expression systems (Ullman *et al.*, 1998; Jarvis and Finn 1995; Chapter 6).

During X-ray data acquisition on Station 2.1, analyses of the 10 time-frames of 1 min each revealed small time-dependent effects due to irradiation. Only the first time-frame was used for data analysis to avoid this effect. X-ray data at low Q values on sFI measured between 1.7-3.3 mg/ml concentrations yielded a mean R_G value of 4.04 ± 0.27 nm from linear Guinier plots in an acceptable $Q.R_G$ range of 0.9-1.6 (Figure 7.4a; Table 7.1). Those for rFI measured at 2.0 mg/ml also yielded a mean X-ray R_G value of 4.06 ± 0.12 nm, in good agreement with the sFI data. The X-ray data showed that the overall domain structures of sFI and rFI were similar and were unaffected by the change in oligosaccharide contents.

The neutron scattering data from Instruments D22 and LOQ also resulted in linear Guinier R_G plots for sFI and rFI in 100% ^2H_2O (Figure 7.4b), from which similar mean R_G values of 4.00 ± 0.14 nm and 4.18 ± 0.07 nm respectively were obtained (Table 7.1). In molecular weight calculations based on I(0)/c values from the D22 Guinier fits, the I(0)/c values were determined to be 0.28 ± 0.02 for sFI and 0.23 ± 0.03 for rFI on the basis of absorbance coefficients (280 nm, 1%, 1 cm) of 14 and 12 respectively (Ullman *et al.*, 1998; Chapter 6). The 18% reduction in the D22 I(0)/c value for rFI confirmed the expected 13% reduction in molecular weight of rFI when the complex-type oligosaccharides were replaced by high mannose-type ones. The mean neutron I(0)/c value from the LOQ Guinier fits was 0.076 ± 0.008 for sFI and rFI relative to a polymer standard. This value corresponds to the molecular weight range of 74,500 for rFI and 85,300 for sFI by comparison with I(0)/c values measured on LOQ for nine other proteins of molecular weights 27,000-254,000. It is concluded that the neutron R_G values validate the X-ray R_G values, and that both sFI and rFI are monomeric in solution with

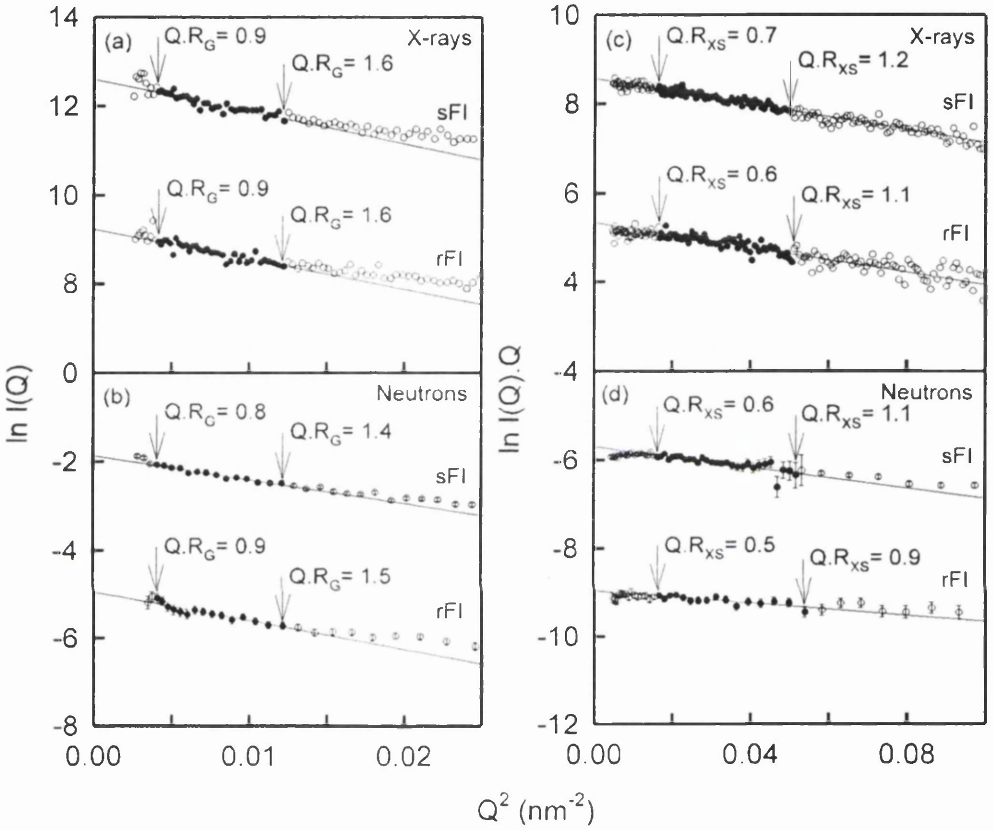


Figure 7.4: X-ray and neutron Guinier analyses of sFI and rFI. Filled circles between the indicated Q.R$_G$ and Q.R$_{XS}$ ranges show the data points used to determine the R$_G$ and R$_{XS}$ values (Table 7.1). Statistical error bars are shown when large enough to be visible. (a,c) X-ray Guinier R$_G$ and R$_{XS}$ plots are shown for sFI and rFI at concentrations of 3.3 and 2.0 mg/ml. (b,d) Neutron Guinier R$_G$ and R$_{XS}$ plots are shown for sFI and rFI in 100% ^2H_2O buffer at concentrations of 0.5 and 1.7 mg/ml respectively, measured using Instruments D22 and LOQ respectively.

Experimental	Guinier analyses[1]		GNOM analyses[2]
	R_G (nm)	R_{XS} (nm)	R_G (nm)
sFI (X-rays)	4.04 ± 0.27 (4)	1.70 ± 0.15	4.31 ± 0.31
sFI (neutrons)	4.00 ± 0.14 (6)	1.51 ± 0.08	4.08 ± 0.21
rFI (X-rays)	4.06 ± 0.12 (6)	1.57 ± 0.10	4.35 ± 0.29
rFI (neutrons)	4.18 ± 0.07 (4)	1.22 ± 0.06	4.29 ± 0.12

Modelling	Guinier analyses[1]		R-factor[2]
	R_G (nm)	R_{XS} (nm)	(%)
Linear sFI model	6.14	0.88	16.1
Part-bent sFI model	5.32	0.93	13.2
Half-bent sFI model	4.75	1.73	12.2
Fully-bent sFI model	3.71	2.25	14.2
Best-fit sFI model (X-rays)	4.10	1.85	10.2
(neutrons)	4.17	1.44	10.2
Best-fit rFI model (X-rays)	4.01	1.61	11.4
(Neutrons)	4.20	1.25	10.0

[1] The number of scattering curves measured for each sample is shown in brackets below. The Q range used for the R_G determinations was 0.20-0.35 nm^{-1} (Figures 7.5a and 7.5b), while that used for the R_{XS} determinations was 0.45-0.81 nm^{-1} (Figures 7.5c and 7.5d).
[2] The R-factor goodness-of-fit parameter was defined (Smith *et al.*, 1990) for X-ray data in the Q range between 0.20 and 2.0 nm^{-1} and neutron data in the Q range between 0.29 and 1.8 nm^{-1}. These corresponded to the Q ranges used to calculate the P(r) curves (Figure 7.4).

Table 7.1: Experimental and modelled scattering analyses for sFI and rFI.

molecular weights as expected from their carbohydrate contents.

As factor I has an elongated structure (Perkins *et al.*, 1993; DiScipio, 1992), cross-sectional analyses were performed to provide information on the mean dimensions of the two shorter axes of factor I (Hjelm, 1985). Linear cross-sectional X-ray and neutron Guinier R_{XS} plots were obtained for both sFI and rFI in an acceptable $Q.R_{XS}$ range of 0.5 to 1.2 (Figures 7.4c and 7.4d). The decrease in R_{XS} on going from X-rays to neutrons is consistent with both the observation of a dry structure by neutron scattering as well as a small contrast effect. Both the X-ray and neutron R_{XS} values (Table 7.1) were 0.23-0.29 nm larger for sFI compared to rFI. This difference is attributable to the altered oligosaccharide structures of sFI and rFI on the basis that these are predominantly located on the two shorter axes of factor I. The antennae in the branches of a typical complex-type structure ($NeuNAc_4Gal_4Man_3GlcNAc)_6$ are two residues longer than a high mannose-type one ($Man_7GlcNAc_2$) and contain extended β(1,4) linkages instead of sterically bent α(1,2) linkages. Molecular graphics showed that the maximal dimensions of this complex-type structure are 2.8 × 4.2 nm, while those for the high mannose-type structure are 2.4 × 1.7 nm (Figure 7.3).

The distance distribution functions P(r) calculated from the entire scattering curves I(Q) up to Q of 2 nm^{-1} confirmed the Guinier analyses at low Q values (Figure 7.5). The P(r) analyses resulted in similar R_G values of 4.08-4.35 nm. The most frequently occurring interatomic vector in sFI and rFI corresponds to the peak maximum M of the P(r) curve. The mean values of M for the X-ray P(r) curves of sFI and rFI were 3.9 ± 0.2 nm and 3.7 ± 0.2 nm respectively. Those for the neutron P(r) curves were smaller at 3.5

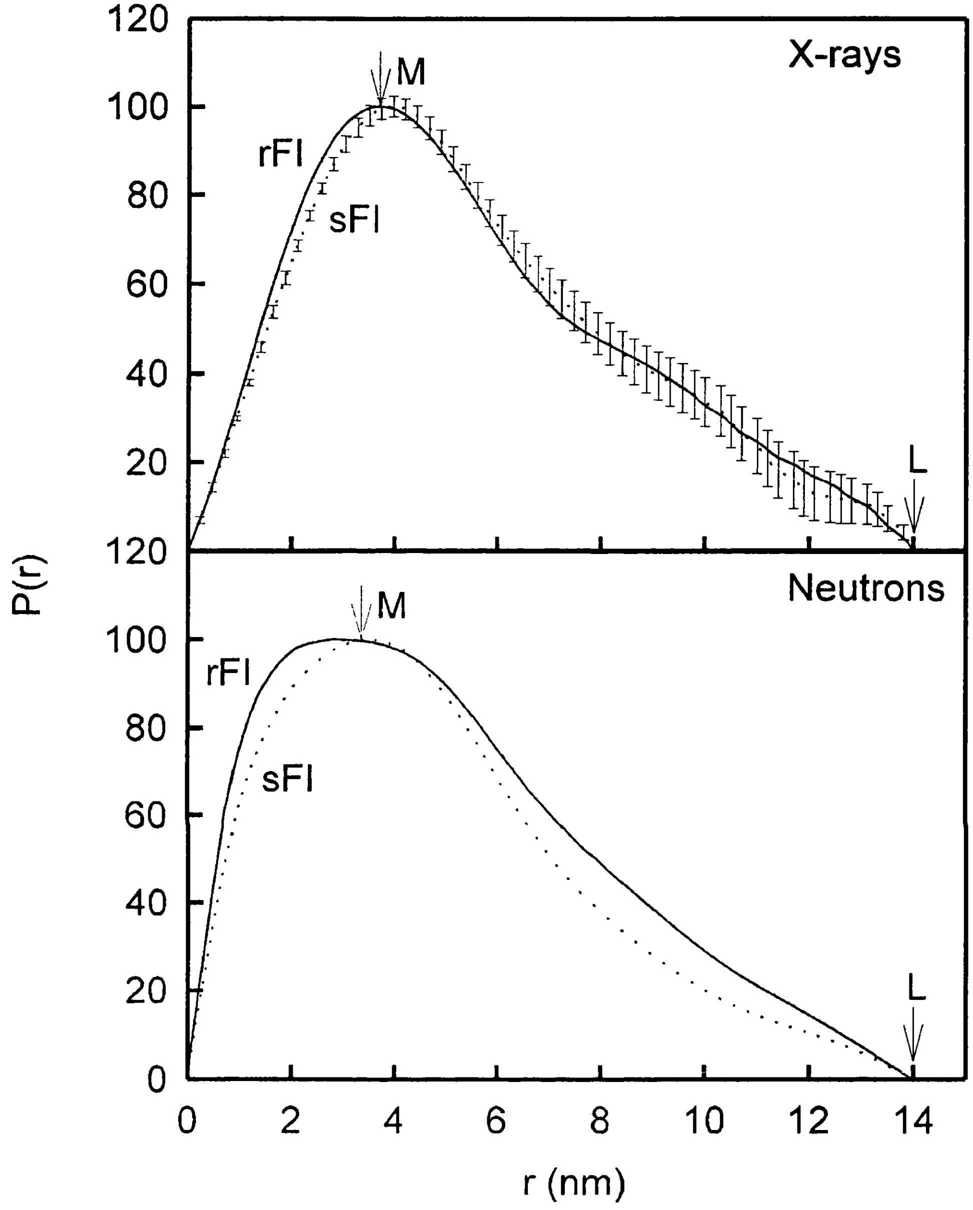


Figure 7.5: X-ray and neutron distance distribution functions P(r) for sFI and rFI. The four P(r) curves were calculated from the I(Q) curves used in Figure 2. The dotted X-ray and neutron P(r) curves correspond to sFI data. The maximum of the P(r) curve is denoted by M, the most frequently occurring distance within sFI and rFI, and the maximum dimension is denoted by L. Representative error bars are shown for the X-ray P(r) curve for sFI.

± 0.2 nm and 2.8 ± 0.2 nm respectively. In parallel with the slightly smaller M values for rFI, both P(r) curves in Figure 7.5 showed greater intensities at low r values between 0 and 3 nm for rFI compared to sFI. The greater proportion of short interatomic vectors in rFI compared to sFI is consistent with the shorter oligosaccharide structures present in rFI. The slight decrease in M on going from X-rays to neutrons is consistent with both a change to a dry structure by neutron scattering as well as a small contrast effect.

The distance distribution functions P(r) also provide the length L of sFI and rFI. The best P(r) analyses suggested that the maximum dimension L for both sFI and rFI was 14 nm. Further sets of L values were calculated from the Guinier analyses (Perkins *et al.*, 1993). From the X-ray R_G and R_{XS} values, L was 14.2 ± 0.6 nm, and from the neutron R_G and R_{XS} values, L was 13.5 ± 0.7 nm, both of which were in good agreement with the P(r) analyses. The present study is consistent with the previous neutron contrast variation study of sFI (Perkins *et al.*, 1993). The present R_G values of 4.05 nm from Guinier analyses are slightly higher than the previous values, but they are now consistent with the present and previous P(r) analyses.

(7.3.2) <u>Homology Modelling for the FIMAC Domain.</u>

The FIMAC domain was modelled (Figure 7.2) on the basis of a follistatin domain in the crystal structure of SPARC (Hoehnester *et al.*, 1997), where a distant sequence relationship exists between the follistatin and FIMAC sequences (Ullman and Perkins, 1997). Evidence to support a structural relationship was obtained as follows:

(i) The follistatin structure is a hybrid of an N-terminal epidermal growth factor (EGF) domain with a C-terminal ovomucoid domain. The disulphide bridge pattern in the

EGF domain is 1-2, 3-4 and 5-6, the first two of which occur in the follistatin structure. Sequence comparisons had already indicated a high similarity of this region to the EGF domain (Catterall *et al.*, 1987).

(ii) The consensus secondary structure prediction from 52 follistatin and FIMAC sequences (Ullman and Perkins, 1997; Figure 7.6) gave a ββββαβ pattern in full agreement with the follistatin crystal structure, and is 81% accurate on a residue-by-residue basis. The consensus predictions from each of the follistatin and FIMAC sequences were similar, in particular for the C-terminal ββαβ motif corresponding to the ovomucoid domain.

(iii) Construction of the FIMAC model involved only deletions at three surface loops in the follistatin structure without disruption of the structural core (Figure 7.2). The β-hairpin between B1 and B2 is shortened, the sequence PIG is removed to give a loop the same length as that in ovomucoid, and the extra α-helix A2 containing the Cu^{2+} binding site in SPARC could be excised (Ullman and Perkins, 1997; Figure 7.6). The FIMAC model contained two semi-conserved N-linked oligosaccharide sites at Asn52 and Asn84 in human, mouse and xenopus factor I, both of which occurred at solvent-exposed sites with accessibilities of 80% and 50% as required to support this modelling.

(7.3.3) Modelling of the CD5 Domain.

Since no atomic structure is known for the CD5 domain, a multiple sequence alignment of 52 CD5 sequences was constructed to define its disulphide bridges and its secondary structure (Figure 7.7). The CD5 consensus length is 102 residues, and an alignment was readily obtained (Resnick *et al.*, 1994). Ten residues demonstrated greater than 90% conservation, which included four Cys and five other hydrophobic residues,

```
                                  5           10   15   20         25     30       35   40   45      50   55   60             65   70
ACCESSION   PROTEIN NAME      ----+--        --+----+----+---      -+ - ---+---    -+----+----+   ----+----+----+----          +----+----
CFAI_HUMAN  FACTOR I(Human)   LSCDKVF........CQPWQRCIEGT...........CV.C.KLPYQCPK....NGTAVCATNR....RSFPTYCQQKSLECLHPGT...........KFLNNGTCT
XLC3BC4B    FACTOR I(Xenopus) LSCHKVF........CAPWQRCVAGV...........CR.C.KLPYQCPK...NATTEVCTDGK....RKLQSYCQLKSVECSNPLNSK........YRFSSEAPCT
MMU47810    FACTOR I(Mouse)   RSCNKVF........CQPWQRCIEGT...........CI.C.KLPYQCPR....AGTPVCAMNG....RSYPTYCHQKSFECLHPE...........IKFSHNGTCA
CO6_HUMAN   C6 repeat 1       LTKLKGH........CQLGQKQSGSE...........CI.CMSPEEDCSH....HSEDLCVFDTDSNDYFTSPACKFLAEKCLNNQQ..........LHFLHIGSCQ
CO6_HUMAN   C6 repeat 2       ESCGYDT........CYDWEKCSASTSK.........CV.C.LLPPQCFK...GGNQLYCVKMGSSTSEKTLNICEVGTIRCANRK...........MEILHPGKCL
CO7_HUMAN   C7 repeat 1       LTQAVPK........CQRWEKLQNSR...........CV.C.KMPYECGP.....SLDVCAQDERSKRILPLTVCKMHVLHCQGRN...........YTLTGRDSCT
CO7_HUMAN   C7 repeat 2       KACGA..........CPLWGKCDAESSK.........CV.C.REASECEE....EGFSICVEVNG..KEQTMSECEAGALRCRGQS............SVTSIRPCA
SPRC_BOVIN  SPARC             NPCQNHH........CKHGKVCELDENNTPM......CV.C.QDPTSCPA.PIGEFEKVCSNDN....KTFDSSCHFFATKCTLEGTKKGHK.....LHLDYIGPCK
SPRC_HUMAN  SPARC             NPCQNHH........CKHGKVCELDENNTPM......CV.C.QDPTSCPA.PIGEFEKVCSNDN....KTFDSSCHFFATKCTLEGTKKGHK.....LHLDYIGPCK
SPRC_MOUSE  SPARC             NPCQNHH........CKHGKVCELDESNTPM......CV.C.QDPTSCPA.PIGEFEKVCSNDN....KTFDSSCHFFATKCTLEGTKKGHK.....LHLDYIGPCK
SPRC_RAT    SPARC             NPCQNHH........CKHGKVCDGESNTPM.......CV.C.QDPTSCPA.PIGEFEKVCSNDN....KTFDSSCHFFATKCTLEGTKKGHK.....LHLDYIGPCK
SPRC_CHICK  SPARC             NPCQNHH........CKHGKVCEVDDNNSPM......CV.C.QDPSSCPA.HSGVFEKVCGTDN....KTYDSSCHFFATKCTLEGTKKGHK.....LHLDYIGPCK
SPRC_XENLA  SPARC             NPCLNHH........CKHGKVCEVDESNTPL......CV.C.QDPSTCPT.SVGEFEKICGTDN....KTYDSSCHFFATKCTLEGTKKGHK.....LHLDYIGPCK
SPRC_CAEEL  SPARC             NPCEDHQ........CGWGKECVGKNGEPT.......CE.C...ISKCPELDGDPMDKVCANNN....QTFTSLCDLYREEGLCKRKSKECSKAFNAKVHLEYLGECK
QR1_COTJA   QR1               DLCRNFH........CKRGKVCQADKQGKPS......CI.C.QDPAACPS..TKDYKRVCGTDN....KTYDGTCQLFGTHCQLEGTKMGRQ.....LHLDYMGACK
SC1_RAT     SC1               DSCMNFQ........CKRGHTCNTDQHGKPH......CV.C.QDPETCPP..AKILDQACGTDN....QTYASSCHLFATKCMLEGTKKGHQ.....LQLDYFGACK
HSHEVIN     HEVIN(Human)      DSCMSFQ........CKRGHICKADQQGKPH......CV.C.QDPVTCPP..TKPLDQVCGTDN....QTYASSCHLFATKCRLEGTKKGHQ.....LQLDYFGACK
DYGAGR      AGRIN(Marine ray) RTCSDLH........CQYGATCVQSIGRAY.......CE.C..PPSICPK...NKQFKVCGSDG....VTYANECQLKTIACRQGSV..........INILHQGPCQ
AGRI_CHICK  AGRIN repeat 1    DACRGML........CGFGAVCEKSPTDPSQAS....CV.C..EKTACPR....TVAPVCGSDT....STYSNECELEKAQCNQQRR..........IKVISKGPCG
AGRI_RAT    AGRIN repeat 1    DVCRGML........CGFGAVCEPSVEDPGRAS....CV.C..KKNACPA....TYAPVCGSDA....STYSNECELQRAQCNQQRR..........IRLLRQGPCG
AGRI_CHICK  AGRIN repeat 2    DPCAEVT........CSFGSTCVRSADGQTAG.....CV.C...PASCPR...VAESIVCGSDG....KDYRSECDLNKHACDKQEN..........VFKKFDGACD
AGRI_RAT    AGRIN repeat 2    DPCANVT........CSFGSTCVPSADGQTAS.....CL.C...PTTCPG...APDGTVCGSDG....VDYPSECQLLSHACASQEH..........IFKKFNGPCD
AGRI_CHICK  AGRIN repeat 3    .PCKGIL........NDMNKVCKVNPRTRRV......EL.L.SRPENCPR....KREPVCGDDG....VTYASECVMGRTGAIRGLE..........IQKVRSGQCQ
AGRI_RAT    AGRIN repeat 3    .PCQGSM........SDLNHICKVNPRTRHPE.....ML.L..RPENCPA....QHTPICGDDG....VTYENDCVMGRIGATRGLL..........LQKVRSGQCQ
AGRI_CHICK  AGRIN repeat 4    DKCKDE.........CKFNAVCLHRWHAR........CS.C..DRITCPG....TYRPVCARDS....RTYSNDCERQKAECHQKAA..........IPVKHSGPCD
AGRI_RAT    AGRIN repeat 4    DQCPET.........CQFNSVCLSRRGRPH.......CS.C..DRVTCPG....SYRPVCAQDG....HTYNNDCWRQQAECRQQRA..........IPPKHQGPCD
AGRI_CHICK  AGRIN repeat 5    SPCLSVE........CTFGATCVVFNREPV.......CE.C...QQVCQG....RYDPVCGSDN....RTYGNPCELNAMACVLKRE..........IRVKHKGPCD
AGRI_RAT    AGRIN repeat 5    SPCHGVQ........CAFGAVCTVKNGKAE.......CE.C...QRVCPG....IYDPVCGSDG....VTYGSVCELESMACTLGRE..........IQVARRGPCD
AGRI_CHICK  AGRIN repeat 6    .RCGK..........CQFGAICEAETGR.........CV.C...PTECVP....SSQPVCGTDG....NTYGSECELHVRACTQQKN..........ILVAAQGDCK
AGRI_RAT    AGRIN repeat 6    .PCGQ..........CRFGSLCEVETGR.........CV.C...PSECVE....SAQPVCGSDG....HTYASECELHVHACTHQIS..........LYVASAGHCQ
AGRI_CHICK  AGRIN repeat 7    .SCGTTV........CSFGSTCVGGQ...........CV.C....PRCEQ...QPLAQVCGTDG....LTYDNRCELRAASCQQQKS..........IEVAKMGPCE
AGRI_RAT    AGRIN repeat 7    .TCGEKV........CTFGAVCSAGQ...........CV.C....PRCEH...PPPGPVCGSDG....VTYLSACELREAACQQQVQ..........IEEAHAGPCE
AGRI_CHICK  AGRIN repeat 8    DECGSGGSGSGDGSECEQDR.CRHYGGWWDEDAEDDRCV.C...DFTCLA...VPRSPVCGSDD....VTYANECELKKTRCEKRQN..........LYVTSQGACR
AGRI_RAT    AGRIN repeat 8    AECGSGGSGSGEDDECEQEL.CRQRGGIWDEDSEDGPCV.C...DFSCQS...VPRSPVCGSDG....VTYGTECDLKKARCESQQE..........LYVAAQGACR
AGRI_CHICK  AGRIN repeat 9    KSCEEMS........CEFGATCVEVNGFAH.......CE.C..PSPLCSE...ANMTKVCGSDG....VTYGDQCQLKTIACRQGQL..........ITVKHVGQCH
AGRI_RAT    AGRIN repeat 9    VTCVEIH........CEFGASCVEKAGFAQ.......CI.C..PTLTCPE...ANSTKVCGSDG....VTYGNECQLKAIACRQRLD..........ISTQSLGPCQ
FSA_HUMAN   FOLLISTATIN rpt 1 ETCENVD........CGPGKKCRMNKKNKPR......CV.C...APDCSN..ITWKGPVCGLDG....KTYRNECALLKARCKEQPE..........LEVQYQGRCK
FSA_PIG     FOLLISTATIN rpt 1 ETCENVD........CGPGKKCRMNKKNKPR......CV.C...APDCSN..ITWKGPVCGLDG....KTYRNECALLKARCKEQPE..........LEVQYQGKCK
FSA_RAT     FOLLISTATIN rpt 1 ETCENVD........CGPGKKCRMNKKNKPR......CV.C...APDCSN..ITWKGPVCGLDG....KTYRNECALLKARCKEQPE..........LEVQYQGKCK
FSA_SHEEP   FOLLISTATIN rpt 1 ETCENVD........CGPGKKCRMNKKNKPR......CV.C...APDCSN..ITWKGPVCGLDG....KTYRNECALLKARCKEQPE..........LEVQYQGKCK
FSA_XENLA   FOLLISTATIN rpt 1 ........................NKKNKPR......CV.C...APDCSN..ITWKGSVCGIDG....KTYKDECALLKAKCKGVPE..........LDVQYQGKCK
FSA_HUMAN   FOLLISTATIN rpt 2 KTCRDVF........CPGSSTCVVDQTNNAY......CVTC...NRICPE.PASSEQYLCGNDG....VTYSSACHLRKATCLLGRS..........IGLAYEGKCI
FSA_PIG     FOLLISTATIN rpt 2 KTCRDVF........CPGSSTCVVDQTNNAY......CVTC...NRICPE.PTSSEQYLCGNDG....VTYSSACHLRKATCLLGRS..........IGLAYEGKCI
FSA_RAT     FOLLISTATIN rpt 2 KTCRDVF........CPGSSTCVVDQTNNAY......CVTC...NRICPE.PSSSEQSLCGNDG....VTYSSACHLRKATCLLGRS..........IGLAYEGKCI
FSA_SHEEP   FOLLISTATIN rpt 2 KTCRDVF........CPGSSTCVVDQTNNAY......CVTC...NRICPE.PTSSEQYLCGNDG....VTYPSACHLRKATCLLGRS..........IGLAYEGKCI
```

```
FSA_XENLA   FOLLISTATIN rpt 2 KTCRDVL........CPGSSSCVVDQTNNAY......CVTC...NRICPE.PTSPDQYLCGNDG....ITYGSACHLRKATCLLGRS...........IGLAYEGKCI
FSA_HUMAN   FOLLISTATIN rpt 3 KSCEDIQ........CTGGKKCLWDFKVGRGR.....CSLC...DELCPD..SKSDEPVCASDN....ATYASECAMKEAACSSGVL...........LEVKHSGSCN
FSA_PIG     FOLLISTATIN rpt 3 KSCEDIQ........CTGGKKCLWDFKVGRGR.....CSLC...DELCPE..SKSEEPVCASDN....ATYASECAMKEAACSSGVL...........LEVKHSGSCN
FSA_RAT     FOLLISTATIN rpt 3 KSCEDIQ........CGGGKKCLWDFKVGRGR.....CSLC...DELCPD..SKSDEPVCASDN....ATYASECAMKEAACSSGVL...........LEVKHSGSCN
FSA_SHEEP   FOLLISTATIN rpt 3 KSCEDIQ........CTGGKKCLWDFKVGRGR.....CSLC...GELCPE..SKSEEPVCASDN....ATYASECAMKEAACSSGVL...........LEVKHSGSCN
FSA_XENLA   FOLLISTATIN rpt 3 KSCEDIQ........CSAGKKCLWDSRVGRGR.....CGLS...DDLCGE..SKSDDTVCASDN....TTYPSECAMKQAACSTGIL...........LEVKHSGSCN
S51362      FRP (Mouse)       KICANVF........CGAGRECAVTEKGEPT......CL.C...IEQCKP....HKRPVCGSNG....KTYLNHCELHRDACLTGSK...........IQVDYDGHCK
                                  5           10   15   20         25      30     35   40   45       50   55   60              65   70
RESIDUE CONSERVATION          ----+--        --+----+----+---      -+ -  ---+---  -+----+----+    ----+----+----+----           +----+----
>90% conservation               C            C     C               C  C      C          ICA         Y   C      C                L     G C
>70% conservation               C            C G   C               CV C      C          ICA D      TY   C L    C                L L Y G C
PREDICTED STRUCTURE
GOR I                         ttttttt        ttttttEEtttttttt      tE E  ttttttt  ttttEEEEtttt    tttttHHHHHHHHHHHHHH           EEEttttttt
GOR III                       ctttttt        cctttcEEttttccct      EE c  ccttccc  ccctcEEEEctt    ccccttHHHHHHHHHHt*H           EEEEEtcccc
Chou-Fasman                   ttttEEE        ttttHEEEEtttttEE      EE E  ttttttt  *tttEEEEtttt    *EEttHHHHHHHHHHHHHH           E*H*tttttt
SAPIENS                       ooiooio        ioiooHHHHooooooo      ii i  ooooioo  ioooooiiiooo    EEEooHHHHHHHHHHHHHH           EEEEEoioio
PHD                           lllEEll        lllllEEEElllllll      EE l  lllllll  llllllEEElll    llllllHHHHHHlllllll           EEllElllll
Averaged Structure                                EEEE             EE E               EEEEE        EE  HHHHHHHHHHHHHH           EEEEE
PREDICTED ACCESSIBILITY
Hydropathy                    ooiooio        ioiiooioiooooooii     ii i  oiooiio  iooooiiiiooi    ooiiooioiooioioooio           ioioioioio
PHD Solvent Accessibility     ooiooi.        i.iooi.ioooooooo      ii i  o.oiioo  oo.oooiiiooo    .iioioioiooioiooooo           iiioiooo.o
SAPIENS Solvent Accessibility ooiooio        ioioooiooooooooo      ii i  ooooioo  ioooooiiiooo    ooioooioiooooiooooo           ioioooioio
Averaged Accessibility        ooiooio        ioioooioiooooooo      ii i  ooooioo  ioooooiiiooo    ooioooioiooioiooooo           ioioioioio
```

Figure 7.6: Multiple sequence alignment of 52 FS/FIMAC sequences. The two copper binding regions of the SPARC sequences are in bold. The alignment shows the consensus sequence length of 74 residues which is conserved in over 50% of sequences, residues that show 90% and 70% conservation, and the predicted secondary structure and accessibility profiles of the domains. The averaged predicted structures are based on the presence of at least 2/5 secondary structures or 2/3 accessibility states at each residue position. Abbreviations are as follows; α-helix, H; β-sheet, E; loop/coil/turn, l, c, t; solvent accessible, o; solvent inaccessible, i. Based on Ullman and Perkins, 1997.

```
                                                             5    10   15     20   25   30           35   40   45   50     55   60            65   70     75   80   85    90    95  100
                                                          ----+ -- --+----+---  -+----+----+----      +----+----+----+-  ---+----+-         ---+----+---- +----+----+---- +-  ---+----+--
ACCESSION    PROTEIN NAME
CFAI_HUMAN   FACTOR I (Human)                             VSLKH.GN.TDSEGIVEVKLVDQDKTMFICKSSWSMRE.......ANVACLDLGFQQGADTQR.RFKLSDLSIN..........STECLHVHCRGLE..TSLAECTFTKRRTMG.....YQDFADVVCYT
XLC3BC4B     FACTOR I (Xenopus)                           FTLTQ.NG.EPGKGIIKVKLPTFEQELFLCGKQWSNRE.......ANVVCRQLGSTKGADASA.SDKVFSLVTEK........PPEHCIQATCRGLE..NSLAECALRKLPMQD......NQVAKVTCYT
MMU47810     FACTOR I (Mouse)                             VSLIY.GR.TKTEGLVQVKLVDQDERMFICKNSWSMAE.......ANVACVDLGFPLGVRDIQGSFNISGNLHIN.........DTECLHVHCRGVE..TSLAECAFTKRRTEL.....SNGLAGVVCYK
MSRE_BOVIN   MACROPHAGE SCAVENGER RECEPTOR (Bovine)       VRLVG.GS.GPHEGRVEIFH..EGQWGTVCDDRWELRG.......GLVVCRSLGYKGVQSVH..KRAYFGKGTG..........PIWLNEVFCFGKE..SSIEECRIRQWGVRA.CS..HDEDAGVTCTT
MSRE_HUMAN   MACROPHAGE SCAVENGER RECEPTOR (Human)        VRLVG.GS.GPHEGRVEILH..SGQWGTICDDRWEVRV.......GQVVCRSLGYPGVQAVH..KAAHFGQGTG..........PIWLNEVFCFGRE..SSIEECKIRQWGTRA.CS..HSEDAGVTCTL
RABMSRIA     MACROPHAGE SCAVENGER RECEPTOR (Rabbit)       VRLVG.GR.GPHEGRVEILH..NGQWGTVCDDHWELRA.......GQVVCRSLGYRGVKSVH..KKAYFGQGTG..........PIWLNEVPCLGME..SSIEECKIRQWGVRV.CS..HGEDAGVTCTL
MSRE_MOUSE   MACROPHAGE SCAVENGER RECEPTOR (Mouse)        VRLVG.GS.GAHEGRVEIFH..QGQWGTICDDRWDIRA.......GQVVCRSLGYQEVLAVH..KRAHFGQGTG..........PIWLNEVMCFGRE..SSIENCKINQWGVLS.CS..HSEDAGVTCTS
MMMAMA       MAMA (Mouse)                                 MRLVN.GA.SANEGRVEIFY..RGRWGTVCDNLWNLLD.......AHVVCRALGYENATQAL..GRAAFGPGKG..........PIMLDEVECTGTE..SSLASCRSLGWMVSR.CG..HEKDAGVVCSN
HUMMAC2A     MAC-2 BINDING PROTEIN (Human)                MRLAD.GG.ATNQGRVEIFY..RGQWGTVCDNLWDLTD.......ASVVCRALGFENATQAL..GRAAFGQGSG..........PIMLDEVQCTGTE..ASLADCKSLGWLKSN.CR..HERDAGVVCTN
MUSCYCLOC    CYCLOPHILIN C-ASSOCIATED PROTEIN (Mouse)     MGLVN.GA.SANEGRVEIFY..RGRWGTVCDNLWNLLD.......AHVVCRALGYENATQAL..GRAAFGPGKG..........PIMLDEVECTGTE..SSLASCRSLGWMVSR.CG..HEKDAGVVCSN
SPER_STRPU   SPERACT RECEPTOR (Sea urchin) Repeat-1       IRLIH.GR.TENEGSVEIYH..ATRWGGVCDWWWHMEN.......ANVTCKQLGFPGARQFY..RRAYFGAHVT..........TFWVYKMNCLGNE..TRLEDCYHRPYGRPWLCN..AQWAAGVECLP
SPER_STRPU   SPERACT RECEPTOR (Sea urchin) Repeat-2       LRMIL.GD.VPNEGTLETFW..DGAWGSVCHTDFGTPD.......GNVACRQMGYSRGVKSIK.TDGHFGFSTG..........PIILDAVDCEGTE..AHITECNMPVTPYQHACPYTHNWDVGVVCKP
SPER_STRPU   SPERACT RECEPTOR (Sea urchin) Repeat-3       IRLMD.GS.GPHEGRVEIWH..DDAWGTICDDGWDWAD.......ANVVCRQAGYRGAVKASGFKGEDFGFTWA..........PIHTSFVMCTGVE..DRLIDCILRDGWTHS.CY..HVEDASVVCAT
SPER_STRPU   SPERACT RECEPTOR (Sea urchin) Repeat-4       VRIV..GM.GQGQGRVEVSL..GNGWGRVCDPDWSDHE.......AKTVCYHAGYKWGASRAAGSAEVSAPFDLEA........PFIIDGITCSGVEN.ETLSQCQMKVSADMT.C...ATGDVGVVCEG
WC11_BOVIN   ANTIGEN WC1.1 Repeat-1                       LRLKD.GV.HRCEGRVEVKH..QGEWGTVDGYRWTLKD.......ASVVCRQLGCGAAIGFP..GGAYFGPGLG..........PIWLLYTSCEGTE..STVSDCEHSNIKDYRNDGYNHGRDAGVVCSG
WC11_BOVIN   ANTIGEN WC1.1 Repeat-2                       VRLAG.GD.GPCSGRVEVHS..GEAWIPVSDGNFTLAT.......AQIICAELGCGKAVSVL..GHELFRESSA..........QVWAEEFRCEGEE..PELWVCPRVPCPGGT.CH..HSGSAQVVCSA
WC11_BOVIN   ANTIGEN WC1.1 Repeat-3                       VRLMTNGS.SQCEGQVEMNI..SGQWRALCASHWSLAN.......ANVICRQLGCGVAISTP..GGPHLVEEGD..........QILTARFHCSGAE..SFLWSCPVTALGGPD.CS..HGNTASVICSG
WC11_BOVIN   ANTIGEN WC1.1 Repeat-4                       LRLVD.GG.GPCAGRVEILD..QGSWGTICDDGWDLDD.......ARVVCRQLGCGEALNAT..GSAHFGAGSG..........PIWLDNLNCTGKE..SHVWRCPSRGWGQHN.CR..HKQDAGVICSE
WC11_BOVIN   ANTIGEN WC1.1 Repeat-5                       LRMVS.ED.QQCAGWLEVFY..NGTWGSVCRNPMEDIT.......VSTICRQLGCGDSGTLN..SSVALREGFR..........PQWVDRIQCRKTD..TSLWQCPSDPWNYNS.CS..PKEEAYIWCAD
WC11_BOVIN   ANTIGEN WC1.1 Repeat-6                       IRLVD.GG.GRCSGRVEILD..QGSWGTICDDRWDLDD.......ARVVCKQLGCGEALDAT..VSSFFGTGSG..........PIWLDEVNCRGEE..SQVWRCPSWGWRQHN.CN..HQEDAGVICSG
WC11_BOVIN   ANTIGEN WC1.1 Repeat-7                       VRLAG.GD.GPCSGRVEVHS..GEAWTPVSDGNFTLPT.......AQVICAELGCGKAVSVL..GHMPFRESDG..........QVWAEEFRCDGGE..PELWSCPRVPCPGGT.CL..HSGAAQVVCSV
WC11_BOVIN   ANTIGEN WC1.1 Repeat-8                       VQLMKNGT.SQCEGQVEMKI..SGRWRALCASHWSLAN.......ANVVCRQLGCGVAISTP..RGPHLVEGGD..........QISTAQFHCSGAE..SFLWSCPVTALGGPD.CS..HGNTASVICSG
WC11_BOVIN   ANTIGEN WC1.1 Repeat-9                       LRLVD.GG.GPCGGRVEILD..QGSWGTICDDDWDLDD.......ARVVCRQLGCGEALNAT..GSAHFGAGSG..........PIWLDDLNCTGKE..SHVWRCPSRGWGRHD.CR..HKEDAGVICSE
WC11_BOVIN   ANTIGEN WC1.1 Repeat-10                      LRMVS.ED.QQCAGWLEVFY..NGTWGSVCRSPMEDIT.......VSVICRQLGCGDSGSLN..TSVGLREGSR..........PRWVDLIQCRKMD..TSLWQCPSGPWKYSS.CS..PKEEAYISCEG
WC11_BOVIN   ANTIGEN WC1.1 Repeat-11                      LRLRG.GD.SECSGRVEVWH..NGSWGTVCDDSWSLAE.......AEVVCQQLGCGQALEAV..RSAAFGPGNG..........SIWLDEVQCGGRE..SSLWDCVAEPWGQSD.CK..HEEDAGVRCSG
HSM130AC2    M130 ANTIGEN (Human) Repeat-1                 LRLVD.GE.NKCSGRVEVKV..QEEWGTVCNNGWSMEA.......VSVICNQLGCPTAIKAP..GWANSSAGSG..........RIWMDHVSCRGNE..SALWDCKHDGWGKHSNCT..HQQDAGVTCSD
HSM130AC2    M130 ANTIGEN (Human) Repeat-2                 MRLTR.GG.NMCSGRIEIKF..QGRWGTVCDDNFNIDH.......ASVICRQLECGSAVSFS..GSSNFGEGSG..........PIWFDDLICNGNE..SALWNCKHQGWGKHN.CD..HAEDAGVICSK
HSM130AC2    M130 ANTIGEN (Human) Repeat-3                 LRLVD.GV.TECSGRLEVRF..QGEWGTICDDGWDSYD.......AAVACKQLGCPTAVTAI..GRVNASKGFG..........HIWLDSVSCQGHE..PAVWQCKHHEWGKHY.CN..HNEDAGVTCSD
HSM130AC2    M130 ANTIGEN (Human) Repeat-4                 LRLRG.GG.SRCAGTVEVEI..QRLLGKVCDRGWGLKE.......ADVVCRQLGCGSALKTS..YQVYSKIQAT..........NTWLFLSSCNGNE..TSLWDCKNWQWGGLT.CD..HYEEAKITCSA
HSM130AC2    M130 ANTIGEN (Human) Repeat-5                 PRLVG.GD.IPCSGRVEVKH..GDTWGSICDSDFSLEA.......ASVLCRELQCGTVVSIL..GGAHFGEGNG..........QIWAEEFQCEGHE..SHLSLCPVAPRPEGT.CS..HSRDVGVVCSR
HSM130AC2    M130 ANTIGEN (Human) Repeat-6                 IRLVN.GK.TPCEGRVELKT..LGAWGSLCNSHWDIED.......AHVLCQQLKCGVALSTP..GGARFGKGNG..........QIWRHMFHCTGTE..QHMGDCPVTALGASL.CP..SEQVASVICSG
HSM130AC2    M130 ANTIGEN (Human) Repeat-7                 LRLVN.GG.GRCAGRVEIYH..EGSWGTICDDSWDLSD.......AHVVCRQLGCGEAINAT..GSAHFGEGTG..........PIWLDEMKCNGKE..SRIWQCHSHGWGQQN.CR..HKEDAGVICSE
HSM130AC2    M130 ANTIGEN (Human) Repeat-8                 LRLTSEASREACAGRLEVFY..NGAWGTVGKSSMSETT.......VGVVCRQLGCADKGKIN..PASLDKAMSI..........PMWVDNVQCPKGP..DTLWQCPSSPWEKRL.AS..PSEETWITCDN
HSM130AC2    M130 ANTIGEN (Human) Repeat-9                 IRLQE.GP.TSCSGRVEIWH..GGSWGTVCDDSWDLDD.......AQVVCQQLGCGPALKAF..KEAEFGQGTG..........PIWLNEVKCKGNE..SSLWDCPARRWGHSE.CG..HKEDAAVNCTD
CD6_HUMAN    CD6 (Human)    Repeat-1                      VRLTN.GS.SSCSGTVEVRL..EASWEPACGALWDSRA.......AEAVCRALGCGGAEAASQLAPPTPELPPP.........PAAGNTSVAANATLAGAPALLCSGAEWRL...C.....EVVEHACRS
CD6_HUMAN    CD6 (Human)    Repeat-2                      LRLVD.GG.GACAGRVEMLE..HGEWGSVCDDTWDLED.......AHVVCRQLGCGWAVQAL..PGLHFTPGRG..........PIHRDQVNCSGAE..AYLWDCPGLPGQHY..CG..HKEDAGVVCSE
CD6_HUMAN    CD6 (Human)    Repeat-3                      WRLTG.GA.DRCEGQVEVHF..RGVWNTVCDSEWYPSE.......AKVLCQSLGCGTAVERP...KGLPHSLSG.............RMYYSCNGEE..LTLSNCSWRFNNSNL.CS..QSLAARVLCSA
CD5_HUMAN    CD5 (Human)    Repeat-1                      ARLTR.SN.SKCQGQLEVYL..KDGWHMVCSQSWGRSSKQWEDPSQASKVCQRLNCGVPLSLG..PFLVTYTP...............QSSIICYGQL..GSFSNCSHSRNDM...C.....HSLGLTCLE
CD5_MOUSE    CD5 (Mouse)    Repeat-1                      VMLSG.SN.SKCQGQVEIQM..ENKWKTVCSSSWRLSQDHSKNAQQASAVCKQLRCGDPLALG..PFPSLNRP...............QNQVFCQGSP..WSISNCNNTSSQDQ..C.....LPLSLICLE
RNCD5RN      CD5 (Rat)      Repeat-1                      VMLSG.SN.SKCQGLVEVQM...NGMKTVCSSSWRLSQDLWKNANEASTVCQQLGCGNPLALG..HLTLWNRP...............KNQILCQGPP..WSFSNCSTSSLGQ...C.....LPLSLVCLE
SHPDCC5      CD5 (Sheep)    Repeat-1                      ........................EWYAVHGQSWGQSSLYQVMPKQFFKLCQKLQCRDPLLLS..SLHHFKDRPF.............QKLMICHGQL..GSFSNCSLNRGHQV........GPLALICSE
CD5_BOVIN    CD5 (Bovine)   Repeat-1                      MRLSG.SG.SRCQGRLEVSN..GTEWYAVHSQSWGQLSLYQVAPRQFLKLCQELQCRDPLLLS..SSRYFKEVQF.............QKLIICHGQL..GSFSNCSLNRGRQV........DSLALICLE
CD5_HUMAN    CD5 (Human)    Repeat-2                      LVAQS.GG.QHCAGVVEFYS..GSLGGTISYEAQDKTQDL.....ENFLCNNLQCGSFLKHLPETEAGRAQDPGEPREHQPLPIQWKIQNSSCTSLE.....HCFRKIKPQKS......GRVLALLCSG
CD5_MOUSE    CD5 (Mouse)    Repeat-2                      LVPGH.EG.LRCTGVVEFYN..GSWGGTILYKAKDRPLGL.....GNLICKSLQCGSFLTHLSGTEAAGTPAPAELRDPRPLPIRWEAPNGSCVSLQ.....QCFQKTTAQEG.....GQALTVICSD
RNCD5RN      CD5 (Rat)      Repeat-2                      LVPGH.EG.LRCTGVVEFYN..GSRGGTILYKAKARPVDL.....GNLICKSLQCGSFLTHLSRIETAELQLCSELRDPRPLPIRWEAQNGSCTSLQ.....QCFQKTTVQEG.....SQALAVVCSD
SHPDCC5      CD5 (Sheep)    Repeat-2                      LVAEP.GG.LRCAGLVEFYS..GGVGGTIGIEPQDEIKDL.....GQLICAALQCGSFLKPLPETEEAQTPKP...GGQRPLPIRWEIQNPRCNSLE.....QCFRKVQPRAG.....GQALGLICSD
CD5_BOVIN    CD5 (Bovine)   Repeat-2                      LVAEP.GG.LRCAGVVEFYS..GGLGGTIGIEPQNDIKDL.....GQLICAALQCGSFLKPLPETEEAQTQKP...EGQRPLPIRWEIQNPKCTSLE.....QCFRKVQPWVG.....GQALGLICSD
CD5_HUMAN    CD5 (Human)    Repeat-3                      SRLVG.GS.SICEGTVEVRQ..GAQWAALCDSSSARSSLR.....WEEVCREQQCGSVNSYR......VLDAG..........DPTSRGLFCPHQK...LSQCHELWERNSY.C.....KKVFVTCQD
CD5_MOUSE    CD5 (Mouse)    Repeat-3                      SRLVG.GS.SVCEGIAEVRQ..RSQWEALCDSSAARGRGR.....WEELCREQQCGDLISFH......TVDA...........DKTSPGFLCAQEK...LSQCYHLQKK.KH.CN....KRVFVTCQD
RNCD5RN      CD5 (Rat)      Repeat-3                      SRLVG.GS.SVCEGIAEVRQ..RSQWAALCDSSAARGPGR.....WEELCQEQQCGNLISFH......VMDA...........DRTSPGVLCTQEK...LSQCYQLQKK.TH.C.....KRVFITCKD
SHPDCC5      CD5 (Sheep)    Repeat-3                      SRLVG.GS.DMCEGSVEVRSGKGQQWDTLCDSSWAKGTAR.....WEEVCREQQCGNVSFYQ......GLDPS..........EKTLGGLYCPSGI...LSQCHKLEEKKSY.C.....KRVFVTCQN
CD5_BOVIN    CD5 (Bovine)   Repeat-3                      SRLVG.GS.DVCEGSVEVRSGKGQKWDTLCDDSWAKGTAR.....RVEVCREQQCGNVSSYR......GLDPS..........EKTLGGFYCPPGI...LSQCHKLEEKKSH.C.....KRVFVTCQN
```

```
                              5     10   15     20   25   30           35   40   45   50     55   60             65   70     75   80   85    90     95  100
RESIDUE CONSERVATION      ----+ -- --+----+---  -+----+----+----        +----+----+----+-  ---+----+-          ---+----+----  +----+----+---- +-  ---+----+--
90% conserved                          A ID           I                    IC                                           C           C                      I C
70% conserved             ***   *    * * ***       *  **   *            * ***  * *                                     * * *    *  *          *         * *
PREDICTED STRUCTURE
GOR-I                     EEEtt tc cctccEEEEEt  ttttctEtttttcctt        HHEEEEttttttEEEEc  ccccccc*cc          ctEEttttttttt  ttttttttttttttt tt  tttttEEEEtt
GOR-III                   EEEEt tc ccttt*EEEEE  ttttEEEEcctHccHt        HHHHEEtttccccEEEc  ccEEEccccc          ccEEtEEHHttct  ttEEHHttttttttt tt  cctttEEEEEc
Chou Fasman               E*EHt tt ttttHHHHHHt  tttEEEEEtttHHHHH        *EEEEEEEttttHHHHH  HHHHHtHttH          HHHHHHEEttttH  HHHEHE*HEHEEttt tt  HHHHEEEEEtt
SAPIENS                   ioEio oo ooioioEEEEE  oooiooiioooioooo        ooiiiooioiooiioHi  HHHHHooooo          oiiiooioioooo  ooiioiooooioooo io  ooooioiiioo
PHD                       llEll ll lllllEEEEEl  llllEEEEElllllll        HHHHHHHllllllEEll  llllEEllll          lEEEEEEEEElll  lllllllllllllll ll  llllEEEEEEl
Averaged Structure          E            EEEE       EEEE                    EE      EE                           EE   E                                 EEEE
```

Figure 7.7: Multiple sequence alignment and secondary structure analysis of 52 CD5 sequences. The sequences are identified by their accession codes to the left. The alignment shows the consensus sequence length of 102 residues which is conserved in over 50% of sequences, in which 10 residue types that show 90% conservation are identified and 27 that show 70% conservation are asterisked. The outcome of five averaged secondary structure predictions (see Ullman *et al.*, 1995 and Ullman and Perkins 1997 for details) is summarised below the alignment. The consensus secondary structure prediction is based on the presence of α-helix (H) or β-strand (E) in at least 3 of the 5 predictions at each residue position. Other abbreviations are as follows: loop/coil/turn, l, c, t.

and 27 residues were conserved to better than 70%. In a typical CD5 sequence, between 6 to 8 Cys residues occur at ten positions (10, 26, 39, 44, 65, 70, 80, 85, 90 and 100 in Figure 7.7). The conservation pattern of Cys residues in different sequences showed that Cys10-Cys44, Cys26-Cys65, Cys26-Cys90 and Cys85-Cys90 were bridged. An experimental determination for the CD5 domain in the macrophage scavenger receptor showed that Cys26-Cys90, Cys39-Cys100 and Cys70-Cys80 were bridged (Resnick *et al.*, 1996). In application to factor I, the Cys26-Cys90 bridge is replaced by that between Cys26-Cys65 for reason of a deletion at residues 90-91. The presence of this bridge in factor I is supported by the predominantly hydrophobic nature of residue 65 (Figure 7.7). In the numbering of Figure 7.2, this corresponds to Cys136-Cys196, Cys123-Cys163 and Cys168-Cys178. The CD5 consensus secondary structure was predicted from the sequence alignment by five different averaging methods (Ullman and Perkins, 1997; Ullman *et al.*, 1995) to give 19% β-sheet with six β-strands and no α-helix (Figure 7.7). The protein fold recognition program THREADER was used to score the 52 CD5 sequences for compatibility with 254 known protein folds. While no strong matches were identified, five folds of 99-115 residues in size scored the best, all of which had a preponderance of several β-strands. As both the secondary structure predictions and THREADER analyses showed that the CD5 structure resembled β-sheet proteins, an immunoglobulin fold structure of the same dimensions (4.4 nm × 3.4 nm) as those for the CD5 domain seen by electron microscopy (5.4 ± 1.0 nm × 3.5 ± 0.9 nm) and the same volume calculated from its sequence was used to satisfy the requirements for scattering curve modelling (Resnick *et al.*, 1996; Perkins, 1986).

(7.3.4) Homology Modelling of the LDLr Domains.

The LDLr-1 and LDLr-2 domains were modelled on the basis of a multiple sequence alignment and one crystal and two NMR structures (Fass *et al.*, 1997; Ullman *et al.*, 1995; Daly *et al.*, 1995a; Daly *et al.*, 1995b). The three LDLr disulphide bridges create a compact fold which is stabilised by Ca^{2+} in the crystal structure but not in the NMR structures. In factor I, amino acid changes in LDLr-1/2 make it unlikely that Ca^{2+} binds, in agreement with functional studies (Crossley, 1980). In LDLr-1, even though four acidic Ca^{2+}-binding residues and Asp231 are conserved, Glu215 is missing and two conserved Cys residues are missing from positions 204 and 216 (Figure 7.2). In LDLr-2, a Ca^{2+}-binding residue is replaced by Asn261, while Glu252 and Asp268 are missing (Figure 7.2). Nonetheless, as the sequence length of the LDLr crystal structure coincided exactly with each of LDLr-1 and LDLr-2 in factor I, homology models could be created (Methods). Support for this modelling is provided by the 100% solvent exposure of the N-linked oligosaccharide site as required at Asn244 in mouse factor I (Lys244 in Figure 7.2). A significant feature of the LDLr-1 model is that the solvent accessibilities of Cys237 and Cys238 were 100% and 10% respectively. Visual inspection showed that Cys237 clearly protruded from the surface of the LDLr-1 domain, while Cys238 formed an internal link with Cys223. In addition, the absence of linker residues between the LDLr-1 and LDLr-2 domains meant that the construction of a double domain structure required an extended domain arrangement in order to avoid steric conflicts between the two domains.

(7.3.5) Homology Modelling of the SP Domain.

The SP fold in factor I contains a pair of four-stranded β-sheet Greek key motifs, each adjacent to a β-hairpin with two β-strands (Perkins and Smith, 1993). A homology model was based on the crystal structure of the urokinase-type plasminogen activator which has the same six disulphide bridges as factor I (Spraggon *et al.,* 1995). The modelling was supported by the observed secondary structure in the plasminogen activator which showed that all 12 β-strands were close to their predicted positions in factor I (Kabsch and Sander, 1983a). The β-strands B, C, D, F form the first Greek key with the β-hairpin G, H, followed by the β-strands J, K, L, M and the β-hairpin N, O respectively (Figure 7.2). All the main insertions or deletions in the model occurred in loop regions outside this core of 12 β-strands and 12 Cys residues (Figure 7.2). The validity of the SP model was supported by the accessibilities of 60%, 50% and 70% at three putative N-linked oligosaccharide sites at Asn446, Asn476 and Asn528. It is of interest that all three Asn residues were located on the face of the SP model opposite to that containing the catalytic triad of His362, Asp411 and Ser507, all three Asn residues being in proximity to the attachment point of the heavy chain to the SP domain via Cys309-Cys435.

(7.3.6) <u>Constrained Scattering Curve Modelling of the Domain Arrangement in Factor I.</u>

The domain arrangement of sFI and rFI in solution was assessed by the combination of their X-ray and neutron data in curve fits that were constrained by the above models for the five domains. This was firstly investigated by models based on extended five-domain arrangements where the long axes of each domain were aligned on a common axis. This resulted in an overall length of 18 nm which is larger than the observed value of 14 nm. The curve calculated from this gave a poor fit to the experimental X-ray data (Figures 7.5 and 7.8). This model had an R_G value of 6.14 nm and an R_{XS} value of 0.88 nm which deviated widely from the observed values, and the R-factor of agreement of the curve fit was high at 16.1% (Table 7.1). Efforts to improve this by the use of part-bent, half-bent and fully-bent structures likewise gave poor X-ray curve fits, with different R_G and R_{XS} values from those observed (Table 7.1; Figure 7.8). It was concluded that a structural family of extended domain arrangements similar to the linear or bent four-domain structures found in the multidomain serine proteases factor VIIa and IXa of blood coagulation was not appropriate for factor I (Brandstetter, *et al.*, 1995; Banner *et al.*, 1996).

An improved curve fit approach was based on the observation of a bilobal structure in factor I by electron microscopy, although it had not been explained previously how such a structure could be formed (DiScipio, 1992). Here, the present modelling analyses showed that 19 disulphide bridges could be identified from multiple sequence alignments and known crystal structures, and accounted for 38 of the 40 Cys residues in factor I (Figure 7.1). The remaining two Cys residues were Cys15 and Cys237, where Cys15 is N-terminal to the FIMAC domain and Cys237 is exposed at the C-terminal end of the LDLr-1 domain. The strong inhibition of factor I by dithiothreitol or 2-merceptoethanol showed that the integrity of all 20 Cys bridges was required for activity (Crossley, 1980). The absence of free Cys residues in sFI was confirmed by two experiments based on Ellman's reagent and radiolabelled iodoacetamide (Methods). These considerations suggested that Cys15-Cys237 were disulphide linked to create a triangular globular domain structure (Figure 7.1) that, together with the large globular SP domain, would provide an explanation of the bilobal structure seen by electron

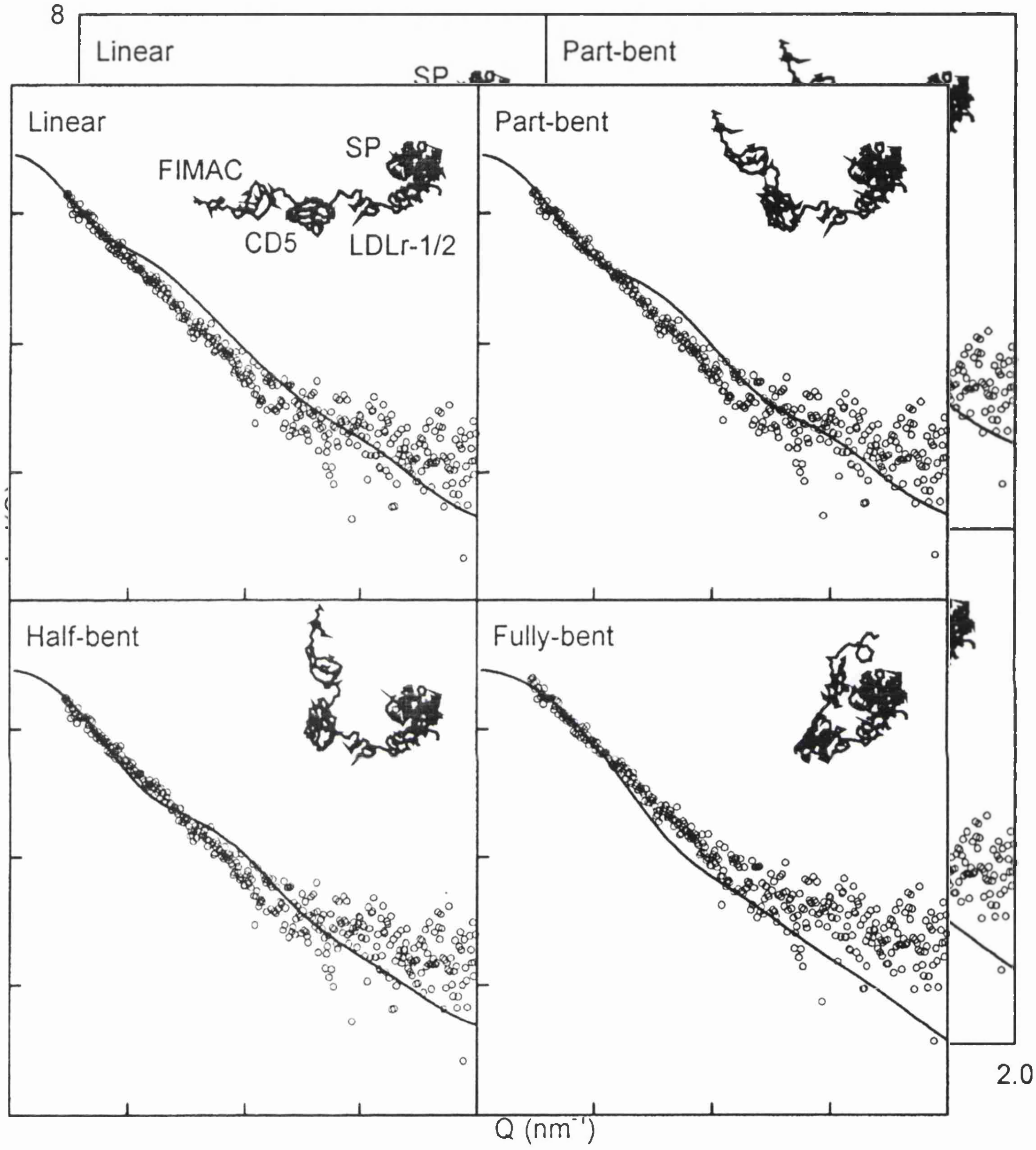


Figure 7.8: Comparison of the calculated and experimental wide-angle scattering curves I(Q) for linear, part-bent, half-bent and fully-dent extended domain models for sFI. The continuous lines represent the curves calculated from the best-fit modelled structure, and the points correspond to experimental data. Inside each panel, a schematic α-carbon view of each domain structure is shown, in all four of which the SP domain is fixed in position.

microscopy for sFI. Such a triangular model could be constructed from the above models for the FIMAC, CD5 and LDLr-1/2 domains. This assembly was stereochemically constrained by the predicted Cys15-Cys237 bridge, the polypeptide links between the four domains, the retention of solvent exposed glycosylation sites at Asn52, Asn85 and Asn244, and the surface attachment of the LDLr-2 domain. Only a limited range of structural variants were possible in the construction of this triangular model, and this meant that the relative positions of these four heavy chain domains was fixed within the structural resolution of solution scattering.

This definition of two globular entities permitted the molecular modelling of the scattering curves of sFI and rFI by an analysis of 9600 different bilobal arrangements of the triangular and SP models. Automated structural searches were performed to test whether any arrangement of these two entities within a bilobal structure for factor I would lead to better curve fits compared to those based on linear extended domain arrangements (Figure 7.8). A full search would involve three rotational parameters to define the orientation of each of the triangular and SP models and three translational parameters to define their separation. As this would have resulted in a computationally prohibitive total of models, the search was simplified in order to test the major features of a bilobal model. For this, it was noted that the triangular model has a long axis which contained oligosaccharide chains on one side, while the SP model also contained three oligosaccharide chains on one side. This asymmetry meant that a comprehensive test of bilobal models could be performed on the basis of 16 different relative 90° rotations of the two models about parallel X-axes in a common X-Z plane. In each of the 16 searches, the SP model was translated in 20 × 30 × 0.5 nm steps in the X-Z plane relative to the triangular model (which was held fixed), and this generated 600 bilobal models.

In order to summarise the outcome of the searches, 16 contour maps were generated to show the resulting number of spheres, R_G and R_{XS} values and R-factors of the curve fits when the calculated curves were compared with the experimental X-ray scattering curve for sFI (Figure 7.9). The application of cut-off filters for the number of spheres and the R_G and R_{XS} values of the sFI models showed that 12 searches each yielded between 10-22 similar good-fit structures (2-4% of the total), while 4 searches gave no solutions. The good-fit structures were all located at the lowest R-factor values in the fourth panel of Figure 7.9c after the application of these three filters in the other panels of Figure 7.9c. This showed that the searches had produced sensible outcomes. Even though the 16 searches had performed independent translations of the two models relative to each other, all the filtered good-fit structures from these searches corresponded to a single minimum in which the SP domain was positioned close to the C-terminus of the LDLr-2 domain. This is consistent with the covalent connection between the LDLr-2 and SP domains (Figure 7.1). There was no evidence for any other minima in these searches. From this single family of related best-fit structures, a final best-fit bilobal model is presented in Figure 7.10 using the smallest R-factor of 10.2% from the search of Figure 7.9c. The curve-fit from this is much improved over those for the extended domain models in Figure 7.8 (Table 7.1). The separation between the centres of mass of the heavy chain and SP domain models was 5.9 nm, and its longest dimension was 13 nm.

Large differences are visible between the X-ray and neutron curves in Figure 7.10

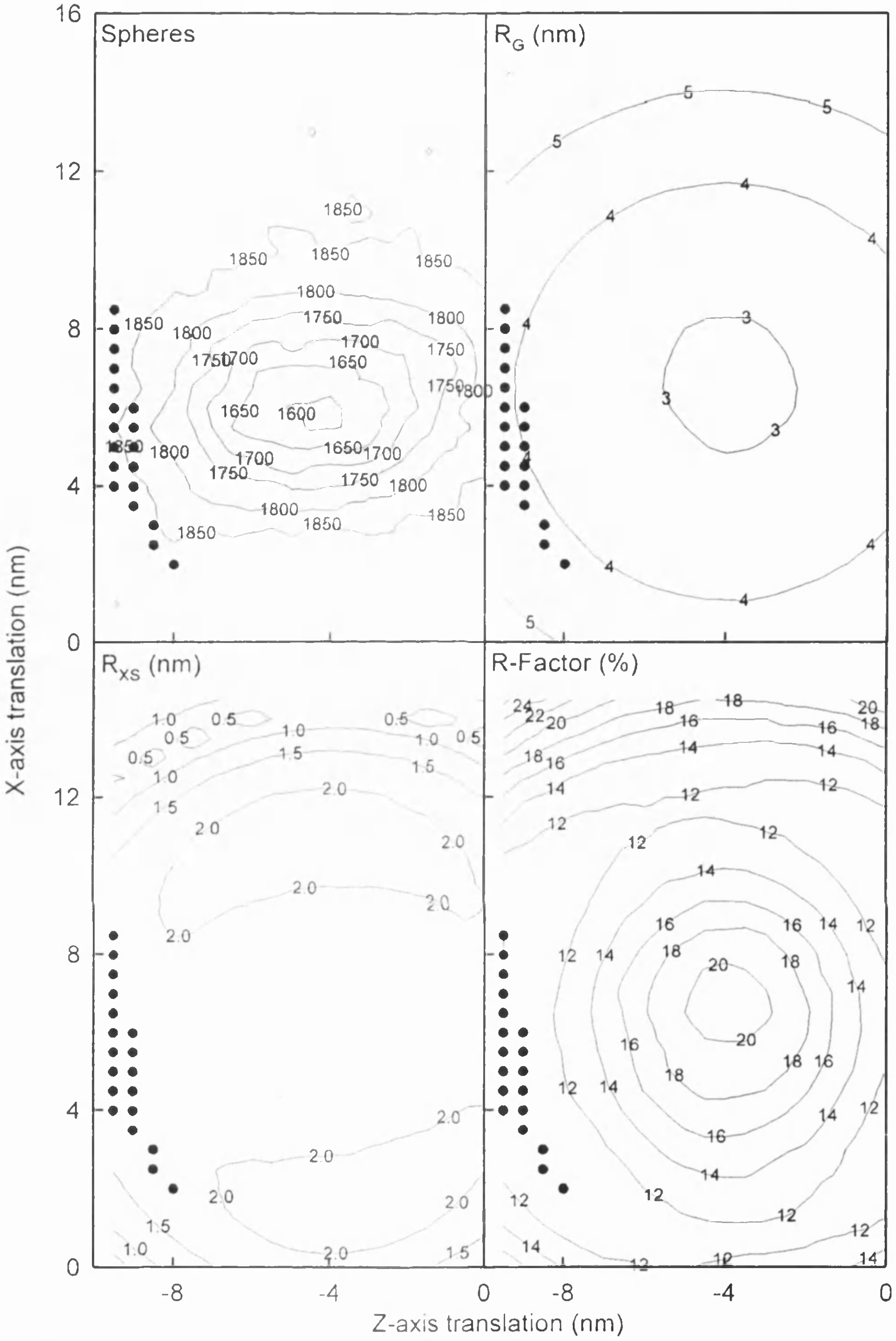


Figure 7.9a: Contour maps from a curve-fit search to determine the domain structure in sFI. Heavy chain rotated Y=90°; SP domain rotated X=0°. The contours correspond to the number of spheres and the R_G, R_{XS} and R-factor values calculated for the 600 models in an automated search in the X-Z plane. In the spheres map, the pronounced minimum corresponds to the complete overlap of the heavy chain and SP models. The 19 good-fit models are denoted by ● to show the positions of the SP domain relative to the heavy chain domains after the use of filters.

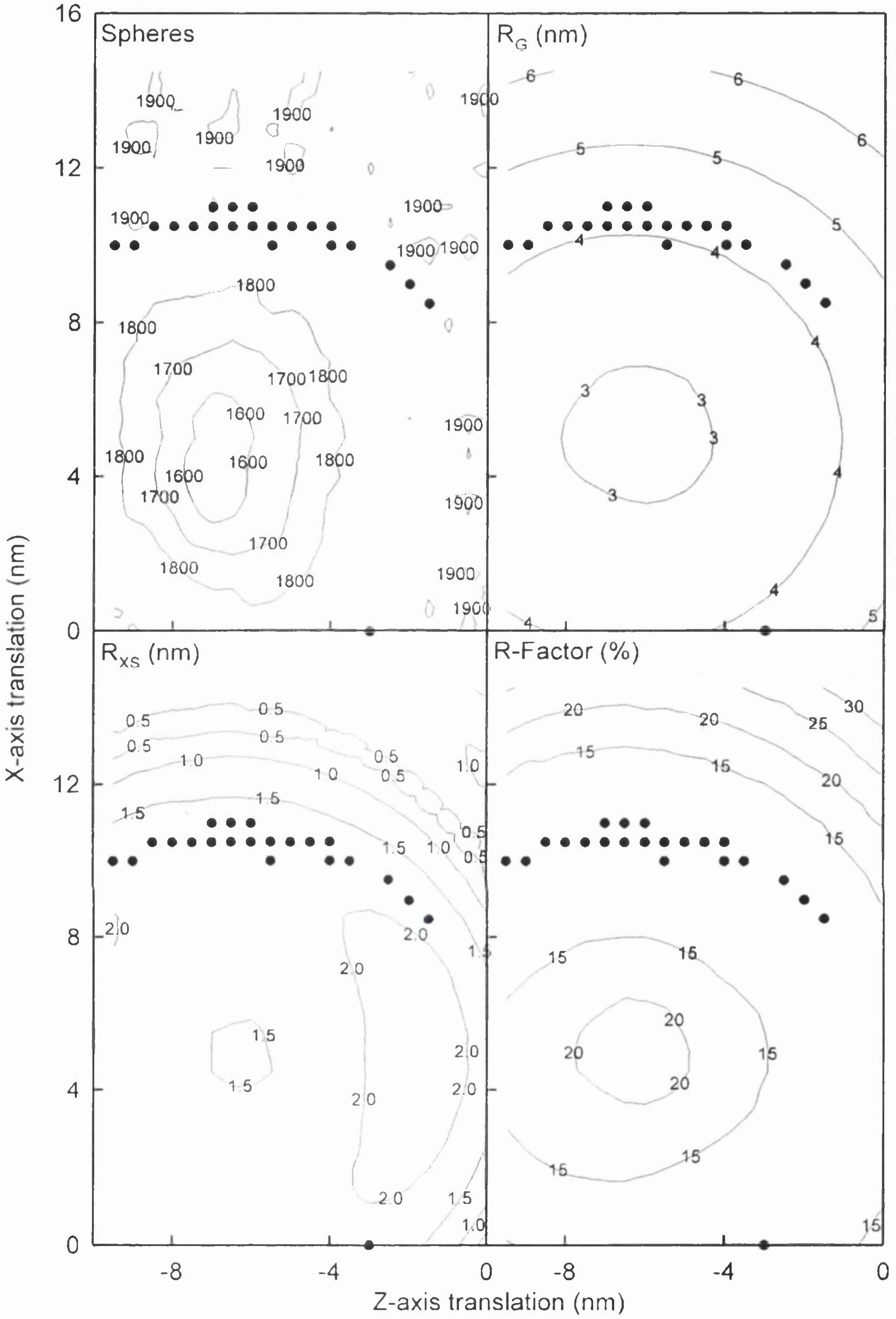


Figure 7.9b: Contour maps from a curve-fit search to determine the domain structure in sFI. Heavy chain rotated Y=90°; SP domain rotated X=90°. The contours correspond to the number of spheres and the R_G, R_{XS} and R-factor values calculated for the 600 models in an automated search in the X-Z plane. In the spheres map, the pronounced minimum corresponds to the complete overlap of the heavy chain and SP models. The 22 good-fit models are denoted by ● to show the positions of the SP domain relative to the heavy chain domains after the use of filters.

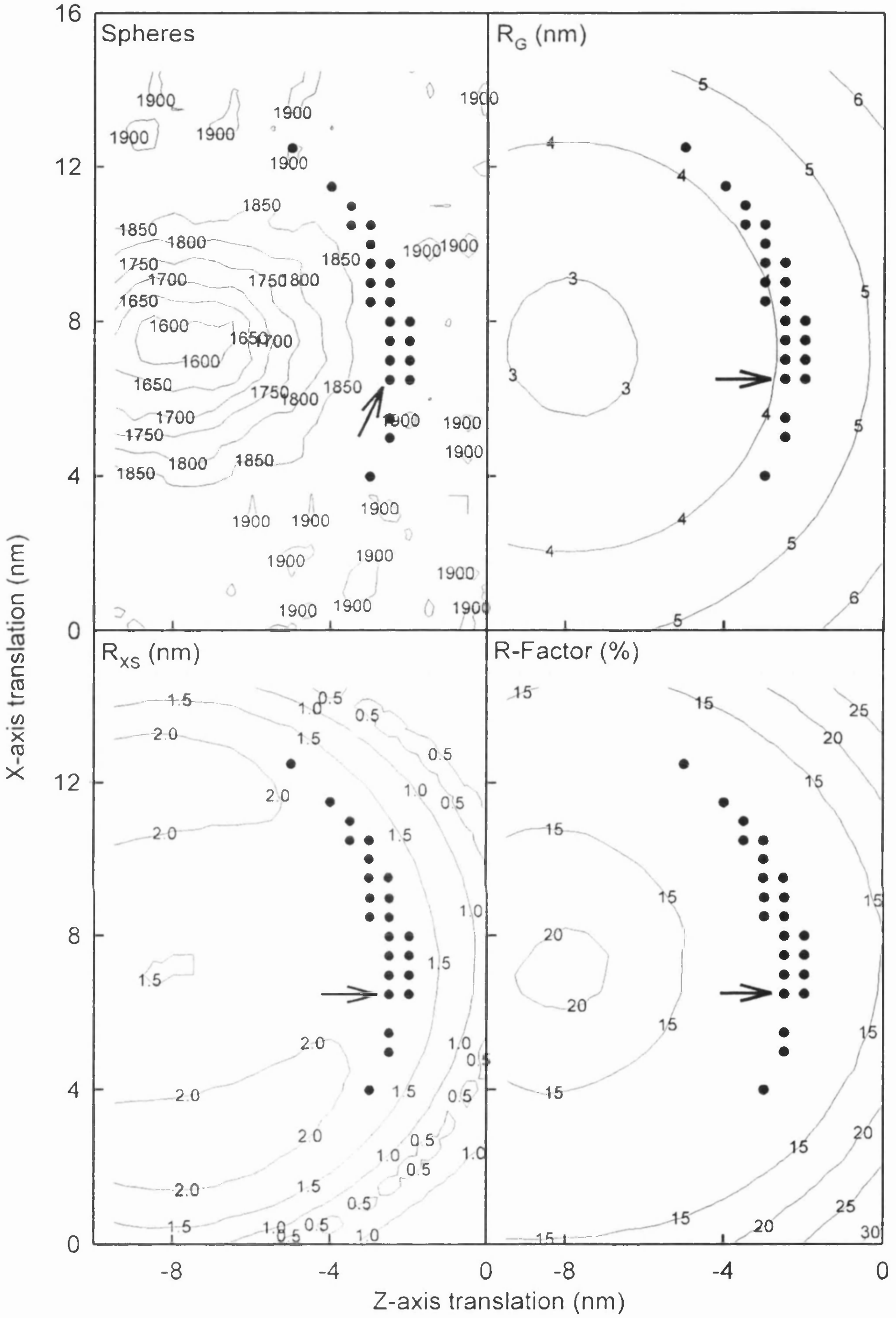


Figure 7.9c: Contour maps from a curve-fit search to determine the domain structure in sFI. Heavy chain rotated Y=90°; SP domain rotated X=180°. The contours correspond to the number of spheres and the R_G, R_{XS} and R-factor values calculated for the 600 models in an automated search in the X-Z plane. In the spheres map, the pronounced minimum corresponds to the complete overlap of the heavy chain and SP models. The 22 good-fit models are denoted by ● to show the positions of the SP domain relative to the heavy chain domains after the use of filters. The best-fit model shown in Figures 7.9 and 7.10 is arrowed.

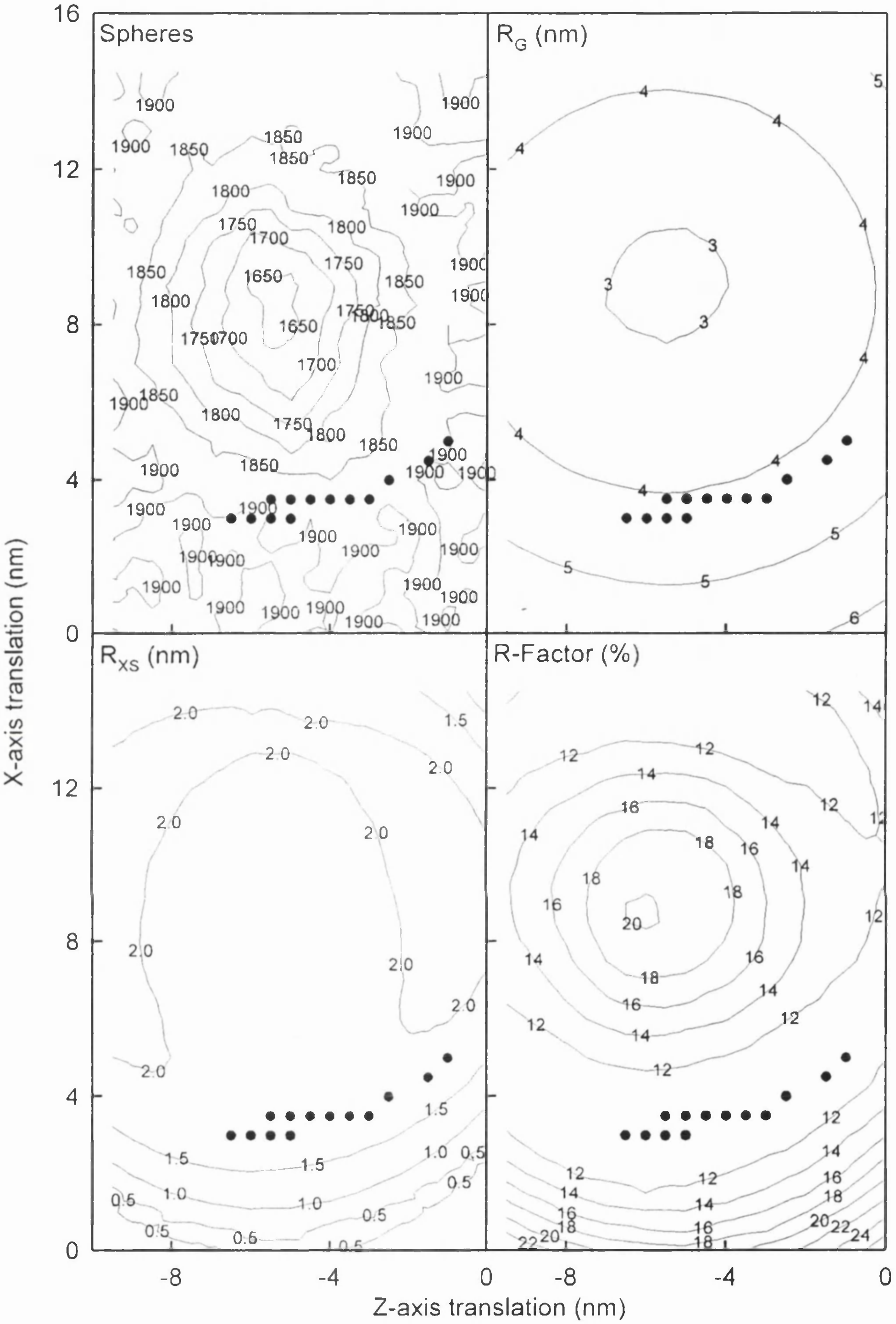


Figure 7.9d: Contour maps from a curve-fit search to determine the domain structure in sFI. Heavy chain rotated Y=90°; SP domain rotated X=270°. The contours correspond to the number of spheres and the R$_G$, R$_{XS}$ and R-factor values calculated for the 600 models in an automated search in the X-Z plane. In the spheres map, the pronounced minimum corresponds to the complete overlap of the heavy chain and SP models. The 13 good-fit models are denoted by ● to show the positions of the SP domain relative to the heavy chain domains after the use of filters.

for reason of the different instrumental geometries in use and the observation of a hydrated and unhydrated structure by X-ray and neutron scattering respectively. In addition, the scattering curve of rFI is distinct from that for sFI in that an inflexion is observed at 0.9 nm^{-1} in the rFI curve (arrowed in Figure 7.10) that is less pronounced in the sFI curve, and this is attributable to the altered oligosaccharide structures present in rFI. These differences provided the basis of three further searches using the neutron curve of sFI and the X-ray and neutron curves of rFI in combination with the above 9,600 models to see whether these would also give the same best-fit bilobal structure for factor I. In the rFI searches, the oligosaccharide structures were converted to high mannose type (Methods). All three further searches resulted in good curve fits with similar R-factors of 10.2%, 11.4% and 10.0% (Table 7.1). The best-fit models from all three searches again placed the LDLr-2 and SP domains in proximity to each other, which were reproduced within a positional range of 2.7 nm in the X-Z plane. The best-fit rFI model gave an unchanged X-ray R_G value when compared with the best-fit sFI model, while its X-ray R_{XS} value was reduced by 0.24 nm (Table 7.1). This effect is explained by the change from complex-type to high mannose-type carbohydrate structures. The best-fit sFI neutron model had an unchanged R_G value when compared with the sFI X-ray model, while its R_{XS} value was reduced by 0.41 nm, and the best-fit unhydrated rFI neutron model performed likewise. This effect is explained by the observation of hydrated structures in X-ray scattering and unhydrated structures in neutron scattering. That all four different modelling analyses consistently resulted in the same bilobal structure for factor I supports the conclusion of a single structural outcome that corresponds to the domain arrangement in factor I shown in Figure 7.10. The R-factors of 10.0-11.4% are slightly higher than the R-factors of 1.2%-8.7% obtained in other scattering analyses of

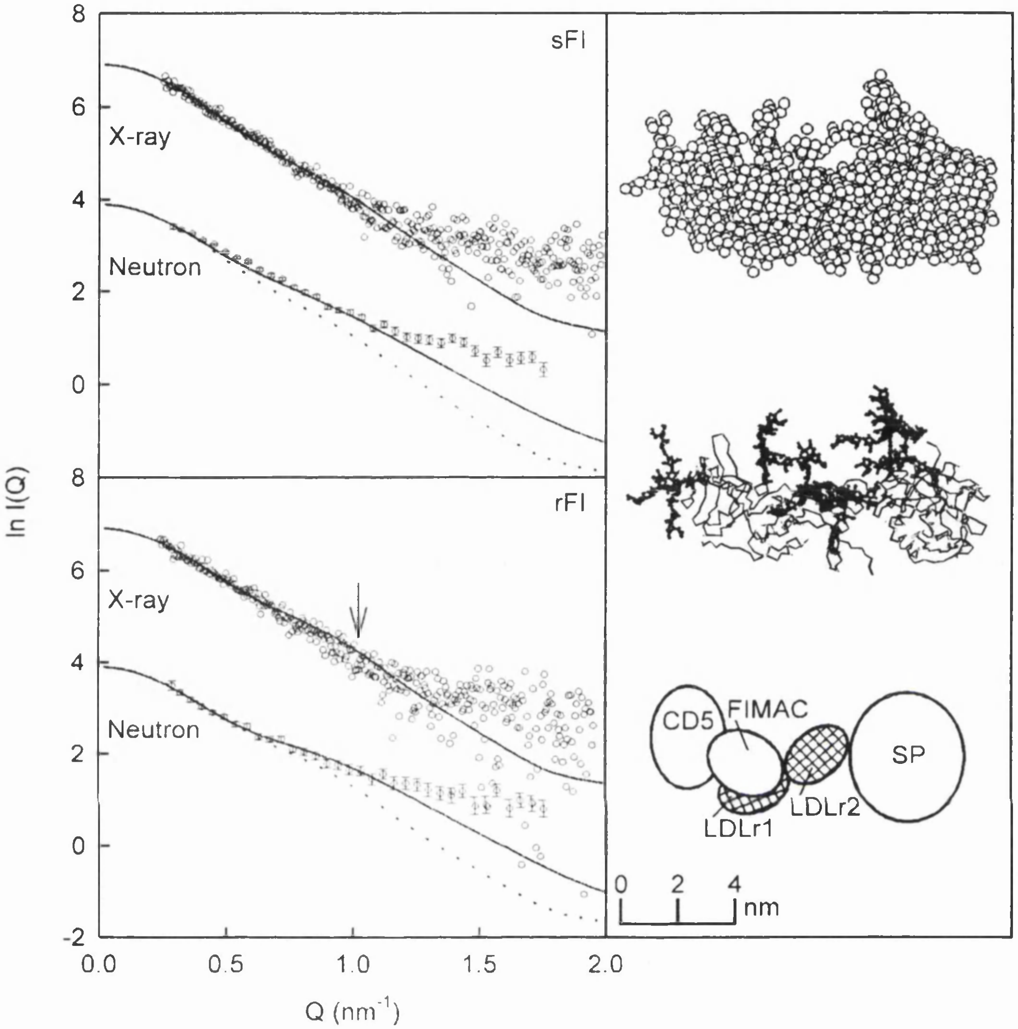


Figure 7.10: Comparison of the calculated and experimental wide-angle scattering curves I(Q) for sFI and rFI. For each of sFI and rFI, the continuous lines represent the curves calculated from the best-fit modelled structure, while the points correspond to the X-ray and neutron experimental data as indicated. The dashed line shown with each neutron curve corresponds to the modelled X-ray curve to show the joint effect of the corrections for hydration and instrumental geometry. The pronounced inflexion in the rFI curves is arrowed. For the best-fit sFI structure, the sphere model, an α-carbon view and a schematic domain outline are shown. The full oligosaccharide structures are shown in bold in the α-carbon view. Figure 7.11 shows enlarged versions of the sphere and coordinate models

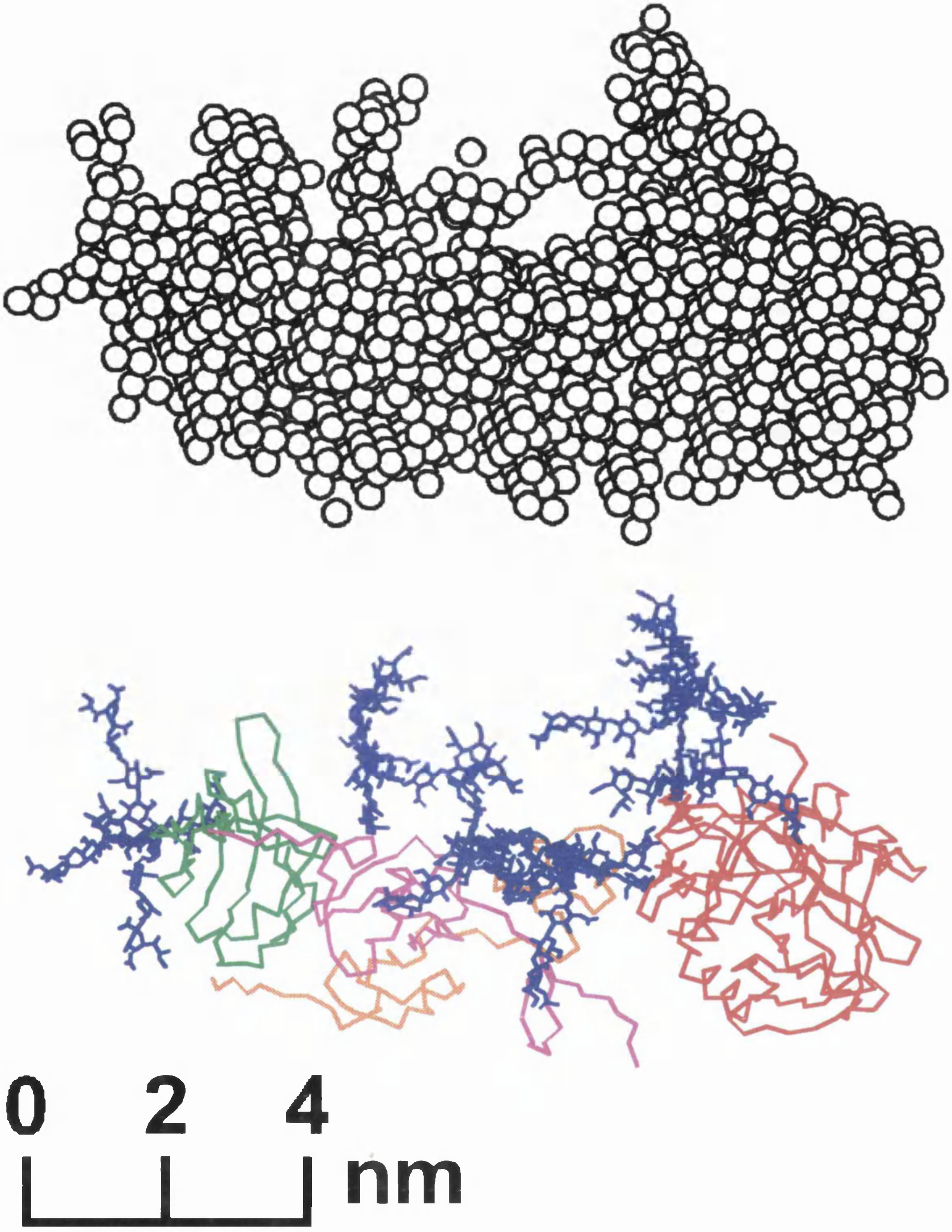


Figure 7.11: Enlarged sphere and α-carbon views of the best fit sFI model from Figure 7.10. The complex-type carbohydrate is shown in blue, FIMAC in purple, CD5 in green, LDLr1/2 in orange, and the SP domain in red.

multidomain proteins (Perkins *et al.*,1998), but is nonetheless consistent with the expected outcome of these modelling analyses.

(7.4) Conclusions.

The combination of neutron and X-ray scattering data, correctly-sized domain structures from homology modelling, and constrained scattering curve fit analyses has revealed new insights into the domain arrangement of the multidomain protein factor I. Unlike the majority of the multidomain plasma serine proteases of the complement, coagulation and fibrinolysis cascades, the factor I domains do not form an extended linear arrangement, as this was ruled out by scattering curve fits (Figure 7.8). Multiple sequence alignments for the five domains in factor I and the use of homology models were able to identify 19 disulphide bridges that were formed from the 40 Cys residues present in factor I (Ullman and Perkins, 1997; Figure 7.6; Ullman *et al.*, 1995; Perkins and Smith, 1993; Figure 7.7). This suggested that Cys15 and Cys237 were unpaired, yet no free Cys residues could be detected in assays of factor I. The combination of this result with the earlier observation of a bilobal structure in factor I by electron microscopy (DiScipio, 1992) suggested that a predicted Cys15-Cys237 disulphide bridge might cause the formation of a compact triangular assembly of the FIMAC, CD5 and LDLr-1 domains (Figure 7.1). The mean diameter of the SP model is 5.3 nm while that of the triangular model is 4.6 nm, and both these values agree well with the diameters of 5.4 nm and 4.9 nm for the two globular structures observed by electron microscopy (DiScipio, 1992). Experimental tests of this bilobal structural model resulted in four satisfactory X-ray and neutron scattering curve fits for both rFI and sFI which provide further support for its existence. The present scattering analyses have resulted in a model for the domain arrangement in factor I that is now most useful for the rational planning of future work to confirm it, in particular confirmatory disulphide bridge mapping to verify the predicted Cys15-Cys237 bridge and the synthesis of smaller fragments of factor I for structural studies.

The multiple sequence alignments for the four domain types in factor I (Ullman and Perkins, 1997; Figure 7.6; Ullman *et al.*, 1995, Perkins and Smith 1993; Figure 7.7) provide complementary information that is consistent with the scattering analyses. Summation of the Asp, Glu, Lys and Arg residues in the three FIMAC domains of human, mouse and xenopus factor I shows that these are basic (net charge of +4 to +7). The corresponding summation for the LDLr-1 and LDLr-2 domains in factor I shows that these are acidic (net charge of -2 to -5 each). The CD5 and SP domains are variable in net charge but close to neutral. The opposite charges on the FIMAC and LDLr domains would enable them to attract each other, which is consistent with the non-extended domain arrangement in factor I by X-ray and neutron scattering (Figure 7.10).

The proposed domain model of factor I provides insights on the roles of (i) the CD5, LDLr-1 and SP domains and (ii) the oligosaccharide chains in the function of factor I. Factor I cleaves C3b or C4b at two or three sites in the presence of a cofactor, which is either the soluble proteins factor H or C4b binding protein, or the membrane-bound proteins complement receptor type 1 or membrane cofactor protein. These cleavages control the activities of the convertase enzymes C3bBb and C4b2a within which C3b and C4b are incorporated (Sim *et al.*, 1993; Law and Reid, 1995). By analogy with the tissue factor-factor VIIa complex of blood coagulation (Banner *et al.*, 1996), the cofactors may induce a conformational change in C3b, C4b or factor I, or provide a binding site to orient the SP domain towards its substrate. It is known that factor I interacts with low affinity to either of factor H or C3(NH_3) in the absence of the third member of the complex at nonoverlapping sites, but the affinity is increased when all three components are present, and there is evidence for the binding of factor H to the heavy chain of factor I (Soames and Sim, 1997). The present molecular identification of two globular entities in a bilobal structure is consistent with these biochemical data in that the triangular domains are able to interact with the cofactor while the SP domain is able to interact with the substrate. The R_{XS} and P(r) analyses showed differences between sFI and rFI which are attributable to variable carbohydrate structures at the centre of factor I (Jarvis and Finn, 1995), and this was supported by the scattering modelling of the oligosaccharide chains at the centre of factor I (Figure 7.10). The oligosaccharides may contribute to the structure of the interlobal region in factor I (Varki, 1993), and evidence to support this is suggested by the 55% activity of rFI when compared with sFI (Ullman *et al.*, 1998). Such a central location for the oligosaccharides implies that protein surfaces will be left exposed for functional interactions. Figure 7.10 suggests that these will primarily involve regions surrounding the catalytic triad on the SP domain and most of the CD5 and LDLr-1 domains in the triangular region. As the pH optimum for factor I function is maximal at pH 4-6 and decreased sharply at pH 6-7, His residues have been implicated in the protein-protein interactions involved in factor I-mediated cleavage (Soames and Sim, 1997). Interestingly, summations of the His residues in the domains of factor I showed that these are most abundant in the SP domain. Given the ionic strength dependence of the factor I-mediated cleavage of C3b (Soames and Sim, 1997), it is of interest that there are as many as nine conserved charged groups in well-aligned regions of the three CD5 sequences from factor I, several of which are not conserved in the remainder of the CD5 superfamily (Figure 7.7).

The domain modelling also provides functional insight on the FIMAC domain in factor I. Follistatin domains in connective tissues can act as spacers (Hohenester *et al.*, 1997; Lane and Sage, 1994). The molecular domain arrangement in factor I suggests that this may constitute the most likely function of the FIMAC domain to link the CD5 and LDLr-1 domains. An alternative viewpoint is a proposal that the FIMAC domain may act as a protease inhibitor in view of its structural similarity with ovomucoid. Despite the intriguing presence of the SP domain in factor I (Hoehnester *et al.*, 1997; Ullman and Perkins, 1997; Lane and Sage, 1994), this is unlikely from the scattering curve fits. The molecular arrangement of the domains in the best curve-fit factor I model places the FIMAC and SP domains far apart (Figure 7.10). If the FIMAC and SP domains were placed in proximity to each other, poor curve fits were obtained (Figure 7.8). Inhibition has not been detected in functional assays using osteonectin or factor I (Hoehnester *et al.*, 1997; Lane and Sage, 1994; C. G. Ullman and S. J. Perkins, unpublished results). In addition, the FIMAC domain is masked by two oligosaccharide sites and the LDLr-1 and LDLr-2 domains in the best-fit model, and this would hinder possible interaction with the SP domain. In fact, one of these two glycosylation sites is located in the ovomucoid-like scissile loop of the homology model for the FIMAC domain that would block its interaction with a potential target if it were to be a protease inhibitor.

CHAPTER 8

MOLECULAR MODELLING OF THE OCTAMERIC SUBUNIT ARRANGEMENT OF RUVA FROM *MYCOBACTERIUM LEPRAE* IN THE PRESENCE AND ABSENCE OF A SYNTHETIC HOLLIDAY JUNCTION DNA AS VISUALISED BY NEUTRON CONTRAST VARIATION

(8.1) Introduction.

Homologous recombination is a universal biological process that plays a fundamental role in creating genetic diversity while also providing an important pathway for DNA repair. Central to this process is the pairing and exchange of strands between two homologous DNA molecules and the formation of crossover intermediates known as Holliday junctions (Figure 8.1). Holliday junctions are key structures in the recombination process as they provide the physical basis for the exchange of information between two DNA molecules (Figure 8.2). Specialised enzymes ensure the completion of the recombination or DNA repair process. The main pathway for the processing of Holliday junctions in *E. coli* is the RuvABC system which comprises two types of activities. One is the Holliday junction-specific endonuclease RuvC that nicks two DNA chains symmetrically across the junction to produce two duplex molecules (Dunderdale *et al.*, 1991). Similar enzymes have been identified in yeast and mammalian cells (reviewed in White *et al.*, 1997; Oram *et al.*, 1998). The second type of activity involves the RuvA and RuvB proteins, which promote the ATP-dependent movement of the DNA crossover point, or branch migration (Tsaneva *et al.*, 1992a; Iwasaki *et al.*, 1992). This results in elongation of the heteroduplex DNA which is essential for DNA repair and recombination events that involve gene conversion. Both genetic (Mandal *et al.*, 1993; Mahdi *et al.*, 1996; Sharples *et al.*, 1994) and biochemical (Eggleston *et al.*, 1997; van Gool *et al.*, 1997) data provide strong evidence that branch migration and resolution occur by a concerted mechanism with RuvABC possibly acting as a complex.

The *E. coli* RuvA and RuvB proteins have been extensively characterised *in vitro* (reviewed in West, 1996, 1997; Shinagawa and Iwasaki, 1996). RuvA binds specifically

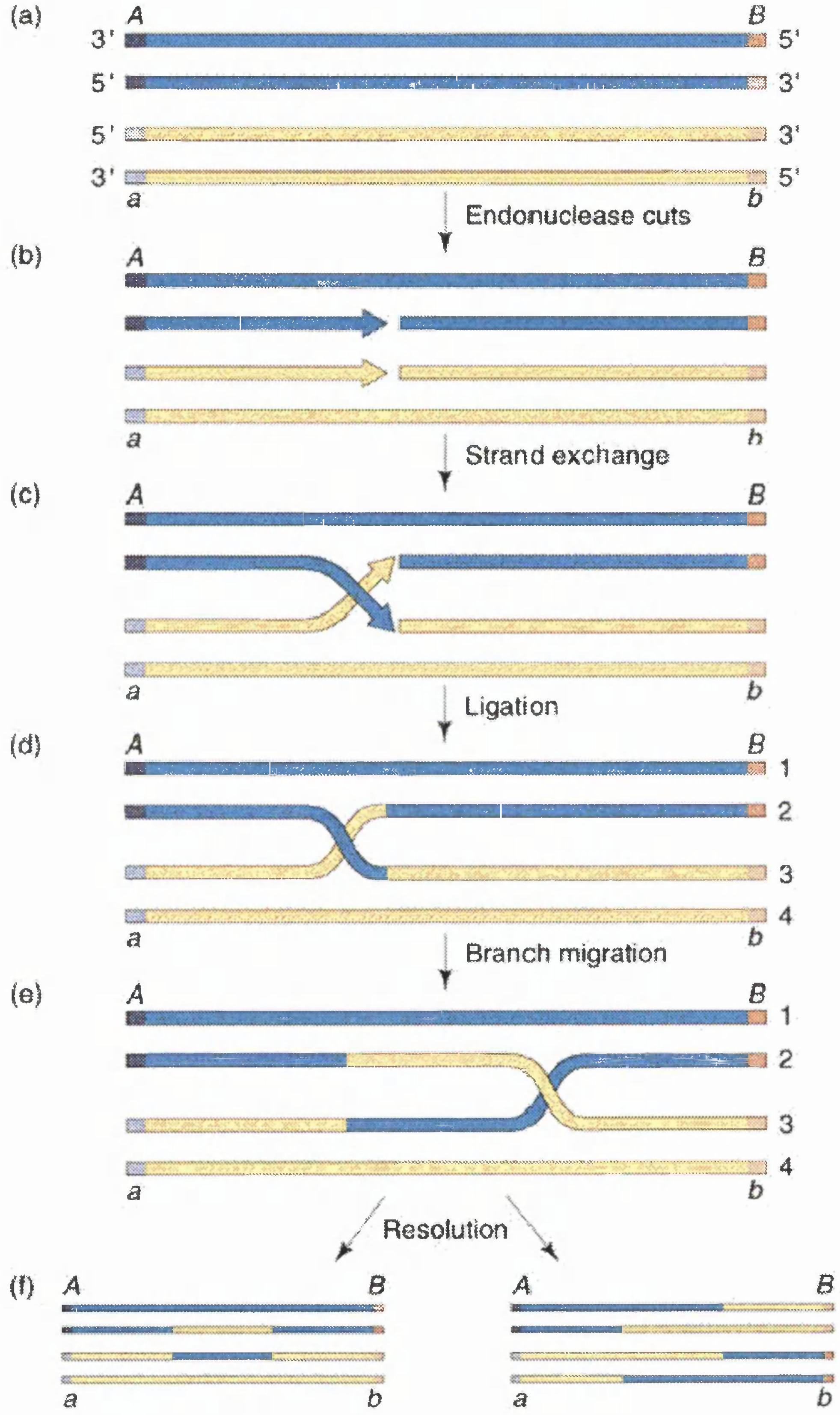


Figure 8.1: Schematic representation of the key stages of recombination. Taken from http://marley.biosci.arizona.edu/michod/.

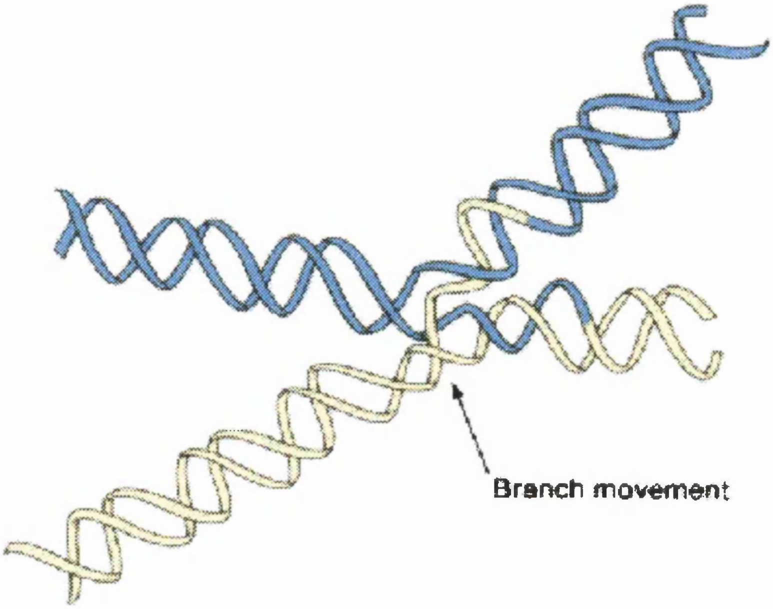


Figure 8.2: Schematic diagram demonstrating strand exchange in a Holliday junction after recombination. Taken from http://marley.biosci.arizona.edu/michod/.

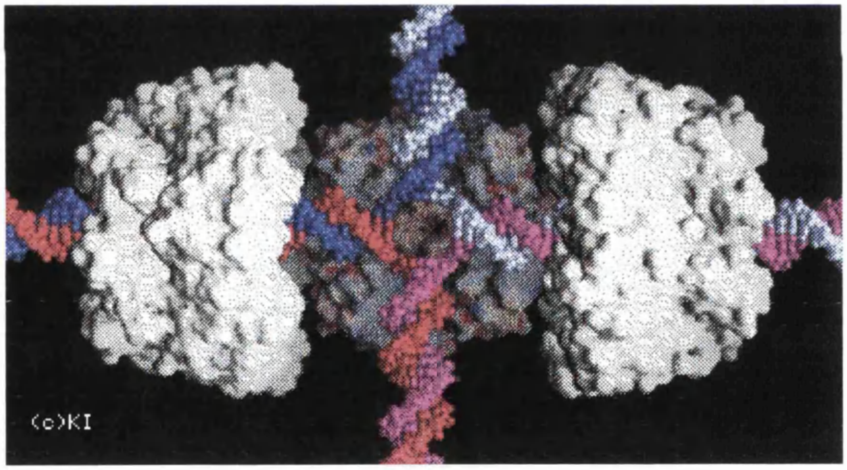


Figure 8.3: Simple model of a RuvA/RuvB/DNA complex (Rafferty *et al.*, 1996). A RuvA tetramer is shown at the back of the complex holding the DNA and targets two RuvB hexamers onto opposite arms of the DNA where they encircle the DNA duplexes and facilitate branch migration in concert with RuvA in an ATP dependent manner. Diagram taken from http://www.shef.ac.uk/~mbb/ruva.html.

to Holliday junctions to form stable protein-DNA complexes (Parsons *et al.*, 1992; Iwasaki *et al.*, 1992). In its complex with RuvA, the junction adopts an open square-planar configuration different from the folded stacked conformation of the free junction in the presence of Mg^{2+} (Parsons *et al.*, 1995). RuvA also interacts with RuvB and mediates the binding of RuvB to DNA (Muller *et al.*, 1993). RuvB is a DNA-dependent ATPase which is stimulated by RuvA (Shiba *et al.*, 1991) and the RuvAB complex exhibits an intrinsic DNA helicase activity (Tsaneva *et al.*, 1993; Figure 8.3). Electron microscopic studies have shown that RuvB forms regular hexameric ring structures on DNA (Stasiak *et al.*, 1994), and this appears to be the active form of RuvB in the branch migration complex (Mitchell and West, 1994; Parsons *et al.*, 1995). The formation of rings assembled around DNA may be a general feature of numerous hexameric helicases involved in DNA replication, transcription, recombination and repair (Egelman *et al.*, 1995; Mastrangelo *et al.*, 1989; Yu and Egelman, 1997). RuvA and RuvB homologues have been identified in the genomes of all bacterial species sequenced so far, and are likely to be ubiquitous in eubacteria, but only the *E. coli* RuvAB system has been studied in detail.

RuvA plays a pivotal role in the current model for the molecular mechanism of RuvAB action, which is based on biochemistry, electron microscopy and crystallography (Parsons *et al.*, 1995; Eggleston *et al.*, 1997; Rafferty *et al.*, 1996; West, 1997). A tetramer (or octamer) of RuvA binding to the DNA crossover is proposed to play a dual role (Figure 8.3). It recruits and promotes the formation of hexameric rings of RuvB on two arms of the RuvA-junction complex. At the same time, RuvA also locks the junction into an open square-planar configuration which favours branch migration. The pair of

RuvB rings are proposed to drive branch migration by the simultaneous and opposite rotation of two DNA arms through the RuvA-constrained junction complex. Such an elaborate molecular machine involves multiple protein-protein and protein-DNA interactions, and structural information is required to unravel its molecular mechanism.

The high resolution crystal structure of *E. coli* RuvA shows a four-fold symmetric tetramer where each monomer contains three well-ordered domains I, II and III, together with a partially disordered link between domains II and III (Rafferty *et al.*, 1996). A model for the four DNA arms of a Holliday junction could be docked into positively-charged grooves on one flat surface of the tetramer. These grooves and the presence of four negatively charged "pins" at the centre suggested a mechanism for how the junction could be constrained into a square planar configuration with unstacked base pairs in the centre (Figure 8.4). If one RuvA tetramer binds to one side of the junction, this would allow RuvC to recognise, bind and cleave the junction on its opposite side. However the RuvA structure would also permit the binding of RuvA tetramers to both sides of the junction. Electron microscopy is consistent with this sandwich model for RuvA in the RuvAB complex (Yu *et al.*, 1997). At present the high resolution structure of the RuvA-DNA complex is not known, and the solution properties of RuvA and its complex are poorly characterised.

Neutron scattering contrast variation is a powerful solution technique that provides information on the dimensions and the arrangement of protein and DNA within a protein-DNA complex as well as its stoichiometry (Perkins, 1988a). In particular, DNA is "invisible" in a buffer containing 65% ^2H_2O, and the use of this buffer permits the

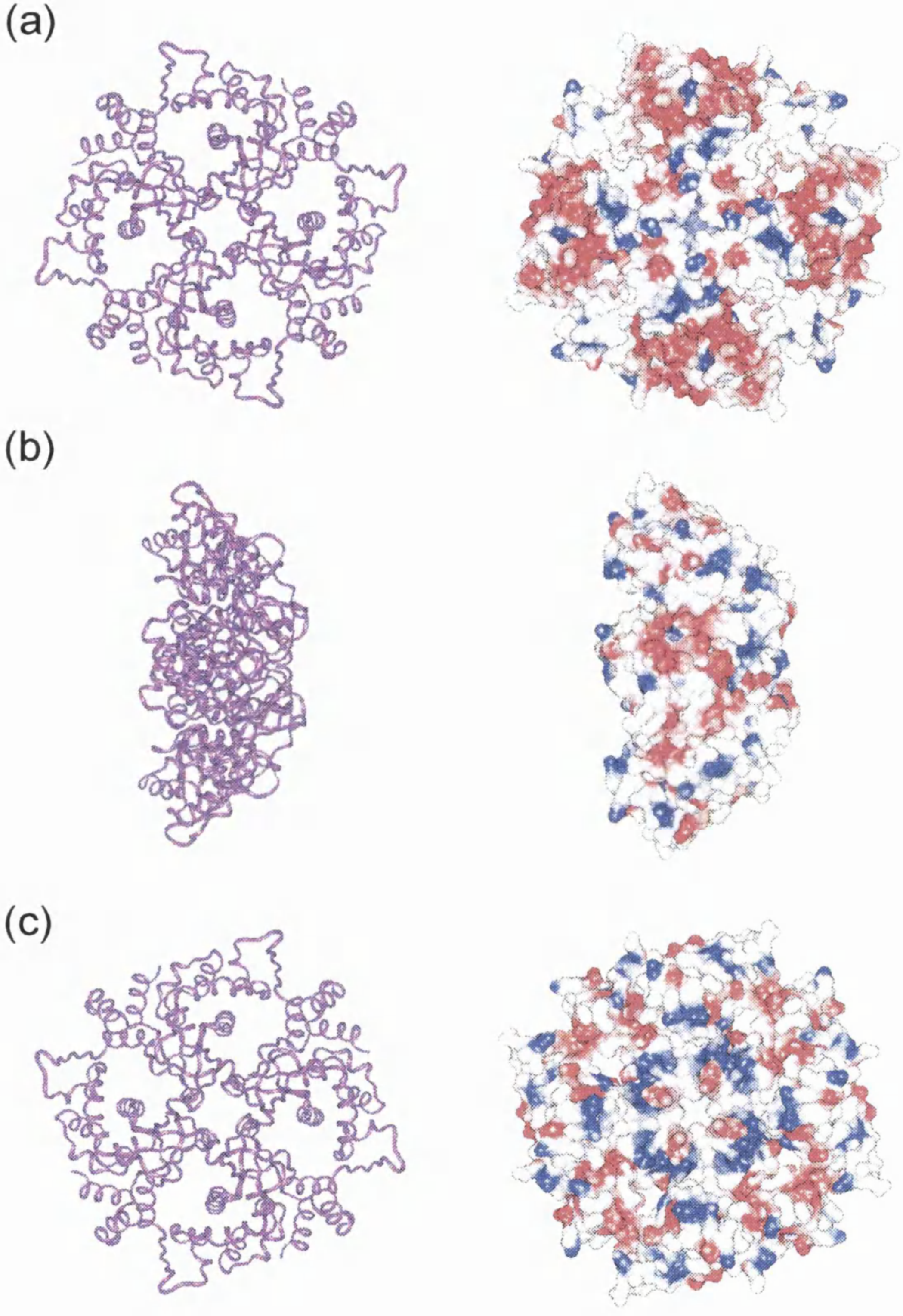


Figure 8.4: Ribbon and electrostatic plots of MleRuvA. Blue shading denotes areas of negative charge. Red shading denotes areas of positive charge. (a) Convex face of MleRuvA. (b) Side view of MleRuvA. (c) Concave DNA binding face of MleRuvA; Note the 4 negatively charged "pins" in the centre of the structure and the uncharged DNA binding grooves.

direct comparison of the unbound and complexed structures of RuvA. For data interpretation, a new method of constrained automated scattering curve fits based on known atomic structures has been developed (Perkins *et al.*, 1998). This approach was used to investigate the structure of the RuvA-Holliday junction complex in solution. Neutron scattering was applied to RuvA from *M. leprae* (MleRuvA), which is a functional homologue of *E. coli* RuvA (EcoRuvA) (Judit Arenas-Licea, Anthony Keeley and Irina R. Tsaneva, unpublished data), and its complex with a synthetic four-way junction containing 16 base pairs in each arm as an analogue of a Holliday junction. These new results show that MleRuvA is octameric in solution, both free and complexed with DNA. Modelling reveals that the solution structure of the complex contains the four-way junction sandwiched between two tetramers of MleRuvA. There is evidence for a conformational change in each RuvA tetramer in the complex. It is concluded that RuvA plays a more active role in recombination than previously expected.

(8.2) Materials and Methods.

(8.2.1) Cloning, Expression and Purification of *Mycobacterium leprae* RuvA (MleRuvA).

The open reading frame for MleRuvA was amplified by the polymerase chain reaction using cosmid L1177 from the *M. Leprae* genome sequencing project as a template and oligonucleotide primers that contained NdeI and HindIII linkers for cloning into the reading frame of the T7 expression vector pET21a(+) obtained from Novagen. The sequence of the forward primer was GAGACATATGATTTTCTCGGTACGC and that of the reverse primer was AGAGAAGCTTCATCGGGTCTTGCCCAGC. The restriction sites introduced by the linkers are underlined. Standard molecular biology protocols were followed for all DNA manipulations (Sambrook *et al.,* 1989). The recombinant plasmid pET21-*ruvA* was transformed in BL21(DE3) cells. For MleRuvA expression, cell cultures in LB media containing 100 μg/ml ampicillin were grown at 37°C to an optical density at 600 nm of 0.6. Induction was with 1 mM IPTG for 4 h. The induced cell pellets were frozen in a dry ice/ethanol bath and stored at -70°C.

For purification of MleRuvA, the induced cells were resuspended and lysed by treatment with lysozyme, Triton X-100 and 1 M NaCl (Tsaneva *et al.,* 1992b). The cleared crude lysate was dialysed against Buffer A (20 mM Tris-HCl, 1 mM EDTA, 0.5 mM DTT, 10% glycerol, pH 7.5). Protein precipitation during dialysis was removed by centrifugation, and the lysate was loaded onto a DEAE-BioGelA column (BioRad), and eluted with a 0 - 1 M KCl gradient in Buffer A. Fractions containing MleRuvA were dialysed against 10 mM K phosphate buffer, 0.5 mM dithiothreitol, 10% glycerol, pH 6.8, and applied to a Bio-Gel hydroxylapatite HTP (BioRad) column, which was eluted with a 10 to 700 mM gradient of K phosphate, 0.5 mM dithiothreitol, and 10% glycerol. MleRuvA fractions from the HTP column were dialysed against 10 mM K phosphate, 0.1 M KCl , 0.5 mM dithiothreitol, 10% glycerol, pH 6.8, and further fractionated on a HiTrap Heparin Sepharose (Pharmacia) column developed with a 0.1 - 1.0 M KCl gradient in 10 mM K phosphate buffer, pH 6.8, 0.5 mM dithiothreitol, 10% glycerol. The MleRuvA fractions from the Heparin column were finally subjected to gel filtration on a Superdex 200 pg column (Pharmacia) equilibrated with Buffer A containing 0.1 M NaCl. The pooled protein was concentrated using Centriplus30 concentrators (Amicon). Full details will be published elsewhere (Judit Arenas-Licea, Anthony Keeley and Irina R. Tsaneva, in preparation).

(8.2.2) Preparation of Four-Way Junction and the MleRuvA-Four-Way Junction Complex.

Four-way junction DNA was prepared using synthetic oligonucleotides designed to form four 16-base pair arms of different sequences as follows:

Oligonucleotide 1: TCACATACGCTTTGCTAGGACATCTTGATATC

Oligonucleotide 2: TGATATCAAGATGTCCATCTGTCCGTTCATC

Oligonucleotide 3: AGATGAACGGACAGATCATGGTGCTTTTAAAG

Oligonucleotide 4: TCTTTAAAAGCACCATGTAGCAAAGCGTATGTG

Oligonucleotides were synthesised on the 1.0 micromole scale on an ABI 394 DNA synthesiser using standard phosphoramidite monomers (Applied Biosystems). Purification was by reversed-phase HPLC, and purity was confirmed by capillary zone electrophoresis (Brown and Brown, 1992).

The four-way junction DNA was annealed from stoichiometric amounts of each oligonucleotide at a total DNA concentration of 1 mg/ml in a buffer containing 20 mM Tris-HCl, pH 7.5, 1 mM EDTA, 5% glycerol, with incubations for 3 min in a boiling water bath, 30 min at 75°C, 30 min at 65°C, 30 min at 37°C, and 30 min at room temperature. The reannealed DNA was stored at 4°C, and concentrated when required using Centricon10 concentrators (Amicon).

To prepare the MleRuvA-junction complex, MleRuvA was added to the junction at a mass ratio of about 4.7:1 protein:DNA in a buffer containing 20 mM Tris-HCl, pH 7.5, 1 mM EDTA, 1 M NaCl, 5% glycerol. The complex was assembled by salt step dialysis against 0.75 M, 0.5 M, and 0.1 M NaCl in the above buffer, concentrated using

Centriplus30 concentrators (Amicon) and purified by gel filtration on a Superdex 200 column (Pharmacia) in 20 mM Tris-HCl, 1 mM EDTA, 0.1 M NaCl, pH 7.5. The samples were finally concentrated using Centriplus30 concentrators (Amicon).

(8.2.3) Neutron Scattering Samples.

After extensive predialysis to ensure the removal of glycerol, the neutron samples were dialysed with four buffer changes during 36 h at 6°C in an EDTA buffer (20 mM Tris-HCl, 1 mM EDTA, 0.1 M NaCl, pH 7.5) or a Mg buffer (20 mM Tris-HCl, 1 mM $MgCl_2$, 0.1 M NaCl, pH 7.5) in 0%, 40%, 65% and 100% ^2H_2O. Samples were placed in quartz Hellma cells of thicknesses 1 mm (0% ^2H_2O) or 2 mm (40%, 65% and 100% ^2H_2O). The DNA concentration was determined from the absorbance at 260 nm using an absorption coefficient of 1 in a path length of 1 cm for 50 µg/ml of DNA as in fully paired duplex DNA (Sambrook *et al.*, 1989). In cases when the absorbance at 260 nm was too high to be reliable, the DNA concentration was determined from that at 280 nm using experimentally-determined A_{260}:A_{280} ratios of 1.95 for the free four-way junction and 1.86 for the MleRuvA-four-way junction complex. The MleRuvA concentration was determined both by Bradford assays using bovine serum albumin as a standard and from the absorbance at 280 nm. The theoretical extinction coefficient of MleRuvA at 280 nm (molecular weight 20,700) calculated from its amino acid composition was 2.76×10^5 M^{-1} cm^{-1}, corresponding to an absorption coefficient (1%, 1 cm) of 1.33 (Perkins, 1986). The ratio of the absorbances of native MleRuvA to that unfolded in 6 M guanidinium HCl, 20 mM potassium phosphate buffer (pH 6.8) was 1.5, and this leads to a corrected absorption coefficient of 2.01. An absorption coefficient of 1.88 gave the best agreement with the Bradford assays, and was used for all MleRuvA concentration measurements.

(8.2.4) Neutron Data Collection and Analysis.

Neutron scattering experiments were performed on Instrument D22 at the neutron reactor at the Institut Laue-Langevin, Grenoble, France, which is analogous to Instrument D11 (Lindner *et al.*, 1992). Data were obtained at 15 °C using sample-detector/collimation distances of 5.6 m/5.6 m and 1.4m/8.0 m, a wavelength λ of 1.00 nm and a rectangular beam aperture of 7 × 10 mm. These configurations avoided any need for detector deadtime corrections. The combined Q range was 0.07 to 2.5 nm^{-1} ($Q = 4\ \pi \sin \theta / \lambda$; scattering angle = 2θ; wavelength = λ). Data acquisition times ranged from typically 3-6 min in 100% ^2H_2O buffers for MleRuvA (0.4-3.8 mg/ml) and the complex (0.5-5.5 mg/ml) up to 30-60 min in 0% and 65% ^2H_2O buffers (0.8-7.0 mg/ml), and 90-120 min for four-way junction (0.4-2.4 mg/ml) in 0% and 100% ^2H_2O buffers. All concentrations for the complex refer to the combined total of protein and DNA. Neutron data using Instrument LOQ at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, U.K. (Heenan and King, 1993) were also obtained at 15°C using a proton beam current of 190 mA to generate neutrons. Acquisitions were for 70 min for MleRuvA concentrations of 2.4-4.5 mg/ml in 2 mm thick Hellma cells. Other details including data reduction protocols are given in Boehm *et al.* (1996) 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 and 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 and depends on the buffer contrast in use. The relative I(0)/c values (c =

sample concentration) for samples measured in the same buffer during a data session gives the molecular weights M_r of the proteins when referenced against a suitable standard (Jacrot and Zaccai, 1981; Wignall and Bates, 1987).

The dependence of the R_G on the reciprocal contrast difference between the macromolecule and the solvent $\Delta\rho^{-1}$ is analysed using Stuhrmann plots (Perkins, 1988a):

$$R_G^2 = R_C^2 + \alpha.\Delta\rho^{-1} - \beta.\Delta\rho^{-2}$$

where R_C is the R_G at infinite contrast (when $\Delta\rho^{-1}$ is zero), α is the radial distribution of scattering density fluctuations within the macromolecule. β is related to this radial distribution if its centre of gravity is at a different location to that of the centre of the macromolecular volume in a non-centrosymmetric macromolecule, and is negligible in the present application.

The matchpoint of a two-component system is determined by a plot of $\sqrt{I(0)/c.t.T_s}$ against volume percentage ^2H_2O (t = path thickness; T_s = sample transmission). Matchpoints were calculated from compositional data by summation of the nuclear scattering factors for all atoms present in the macromolecule, correction for the percentage ^2H_2O present in the buffer assuming full exchange at all NH and OH positions, and division by the dry volume derived from crystallography (Perkins, 1986, 1988).

Indirect transformation of the observed scattering curve in reciprocal space I(Q) into that in real space P(r) was performed using GNOM (Semenyuk and Svergun, 1991):

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

P(r) corresponds to the distribution of distances r between volume elements, from which the R_G and I(0) values can be determined as well as the maximum dimension L. A range of assumed maximum lengths for MleRuvA, the junction and the complex were tested to optimise the calculation of the P(r) curve (Ashton *et al.*, 1997).

(8.2.5) Neutron Scattering Curve Modelling.

In order to make the Debye sphere models used to calculate scattering curves, two sets of atomic coordinates were used. One was those from the crystal structure of EcoRuvA (Brookhaven code 1cuk; PIR accession code P08576). The second was a homology model for MleRuvA (PIR accession code P40832) based on the sequence alignment of Figure 8.5 and INSIGHT II 95.0, BIOPOLYMER, HOMOLOGY and DISCOVERY software (Biosym/MSI, San Diego, U.S.A.) on Silicon Graphics INDY Workstations. Using the rigid body fragment assembly method, a total of 15 structurally conserved regions (total of 109 MleRuvA residues) and 16 designated loops (total of 62 MleRuvA residues) were defined, based mostly on regions of α-helix and β-sheet in EcoRuvA that were identified using the DSSP program (Kabsch and Sander; 1983a). Nine loops (total of 31 MleRuvA residues) that correspond to insertions or deletions were constructed 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 and Sander 1994), and one C-terminal residue was added through an end repair. Energy refinements based on the consistent valence force field were performed at the loop splice junctions,

```
                  10        20        30        40        50
         ....|....|....|....|....|....|....|....|....|....|
MleRuvA: MIFSVRGEVLEVALDHAVIEAAGIGYRVNATPSALATLRQGSQARLV-TA
         ||  ||| |||     |||| ||||| |    |    |  |||  || |
EcoRuvA: MIGRLRGIIIEKQPPLVLIEVGGVGYEVHMPMTCFYELPEAGQEAIVFTH
            <--B1/B2->  <-B3->  <-B4-> <--A1->      <---B5-

                  60        70        80        90        100
         ....|....|....|....|....|....|....|....|....|....|
MleRuvA: MVVREDSMTLYGFSDAENRDLFLALLSVSGVGPRLAMATLAVHDAAALRQ
          |||||   ||||     | ||  ||   ||||||||| ||   |    |
EcoRuvA: FVVREDAQLLYGFNNKQERTLFKELIKTNGVGPKLALAILSGMSAQQFVN
         --->  <--B6->  <-----A2--->     <---A3--->  <---A4

                  110       120       130       140       150
         ....|....|....|..*.|.*..|....|....|....|....|....|
MleRuvA: ALADSDVASLTRVPGIGKRGAERIVLELRDKVGPVGASGLTVGTAADGNA
         ||   ||||| |||||||| |||||||||||   | |   | |      |
EcoRuvA: AVEREEVGALVKLPGIGKKTAERLIVEMKDRFKGLHGDLFTPAADLVLTS
         -->   <-A5>      <--------A6----->

                  160       170       180       190       200
         ....|....|....|....|....|....|....|....|....|....|....
MleRuvA: VRGSVVEALVGLGFAAKQAEEATDQVLDGELGKDGAVATSSALRAALSLLGKTR
           |   |     | |   |     |     |||      ||      || |
EcoRuvA: PASPATDDAEQEAVARLVALGYKPQEASRMVSKIARPDASSETLIREALRAAL-
                <-----A7---->   <---A8--->       <---A9--->
```

Figure 8.5: Sequence alignment of MleRuvA and EcoRuvA used for homology modelling of the MleRuvA structure. The underlined segment is not visible in the crystal structure of EcoRuvA (Brookhaven code 1cuk). The positions of six β-strands (B1-B6) and nine α-helices (A1-A9) are denoted underneath the alignment, and were determined here by DSSP analysis of the crystal structure (Methods). Residue identities or similarities (G=A=S; R=K=H; D=E; Q=N; S=T; I=L=V=M; W=F=Y=H) between the two sequences are indicated by vertical strokes. The conserved residues Lys117 and Gly121 implicated in RuvA octamer formation are asterisked.

then final energy refinements were performed on the sidechain atoms of mutated residues in the structurally conserved regions and the sidechain atoms of both types of loop residues. The secondary structure backbone was retained by fixing the mainchain atoms in the conserved regions. Iterations were made using combinations of the steepest descent and conjugate algorithms to improve the connectivity of the model and minimize bad contacts or stereochemistry. The model was stereochemically verified using PROCHECK (Laskowski *et al.*, 1993). The tetramer was generated from this structure from the crystallographic symmetry operations specified in the Brookhaven file 1cuk.

Neutron scattering curves were calculated using small single-density spheres to represent the MleRuvA structure, using Debye's Law adapted to spheres of a single density (Perkins and Weiss, 1983). To create the spheres, the tetrameric EcoRuvA or MleRuvA coordinates were placed in a three-dimensional grid of cubes of side 0.460 nm or 0.430 nm respectively. 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 MleRuvA protein volume of 107.7 nm^3 for the 812 residues of the EcoRuvA or MleRuvA tetramer (Perkins, 1986; Figure 8.5), and resulted in 1106 and 1347 spheres respectively. This requirement made allowance for the 8% volume increase on going from the MleRuvA sequence to that of EcoRuvA, despite the volume decrease of 5% for the 13 residues per monomer not visible in the EcoRuvA crystal structure (Figure 8.5). The hydration shell is invisible in neutron scattering and was not considered (Perkins, 1986; Perkins *et al.*, 1998). Modelled scattering curves for comparison with the experimental curves were calculated using SCT assuming a uniform sphere scattering density (Perkins and Weiss, 1983). 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). For D22 data, a full-width-half-height wavelength spread of 10% for λ of 1.00 nm and a beam divergence of 0.016 radians were used to correct the calculated curve (Ashton *et al.*, 1997). For LOQ data, a 16% spread in λ for a nominal λ of 0.6 nm and a beam divergence of 0.016 radians were used to correct the calculated neutron scattering curve (Mayans *et al.* 1995; 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 2.0 nm^{-1} (Smith *et al.*, 1990).

The scattering curve modelling of octameric MleRuvA in 100% and 65% ^2H_2O was based on two MleRuvA tetramers.from homology modelling. The two structures were superimposed on top of each other. The X- and Y-axes were defined to lie in the plane of each tetramer with the overall centre of mass as origin, and each passed through the centre of mass of domain III (Figure 8.13). The Z-axis was set perpendicular to the XY-plane to lie on the four-fold axis of symmetry. To initiate the generation of models, one tetramer was rotated 180° around its Y-axis, then translated by -6 nm from the origin along its Z-axis into its starting position. The second tetramer was then translated in 120 steps of -0.1 nm along its Z-axis from the origin to pass through the first tetramer and yielded 120 octamer models. Each model was converted into spheres for scattering curve calculation as described above. Identical calculations with two tetrameric EcoRuvA crystal structures were performed as a control.

The scattering curve modelling of the complex of octameric MleRuvA with the contrast-matched junction in 65% ^2H_2O was based on two tetrameric EcoRuvA crystal structures. The overall X-, Y- and Z-axes of each tetramer were defined as above (Figure 8.13). Individual X-, Y- and Z-axes were assigned to the C-terminal domain III fragment (EcoRuvA Thr156-Leu203 in Figure 8.5) that were approximately parallel to the overall X-, Y- and Z-axes of the tetramer, and their origin was set at the centre of mass of domain III. One tetramer was rotated 180° around its overall Y axis and translated by -10 nm on its overall Z-axis into its starting position. The second tetramer was then translated in 200 steps of -0.1 nm along its overall Z-axis from the origin to give 200 octamer models. In order to explore conformational changes in the complex, the individual centres of mass of the four domain III fragments were moved outwards from the central Z-axis of symmetry along the positive or negative overall X-axis or Y-axis in 20 steps of 0.1 nm for each of the 200 Z-axis steps to give a total of 4200 models. For reason of ambiguity caused by the explicit presence of the domain II-domain III link in the homology-modelled MleRuvA tetramers, the corresponding domain III search was not made using the MleRuvA tetramers.

(8.3) Results and Discussion.

(8.3.1) Neutron Contrast Variation of MleRuvA by Guinier Analyses.

The sequence alignment between EcoRuvA and MleRuvA (Figure 8.5) showed high similarity. Both sequences contained 203 residues and were readily aligned between residues 1-131 (corresponding to domains I and II in the crystal structure) with only a single gap at position 48. As shown, a total of 97 residues (48%) were identical or similar between the two sequences. Between residues 131-204, an improved alignment (not shown) involves the insertion of an 8-residue gap in the MleRuvA sequence just before the disordered linker region between domains II and III, and this increases the total of identical or similar residues to 104 (51%), most notably involving the identical C-terminal residues ALRAAL in both sequences. The calculated molecular weights of EcoRuvA and MleRuvA were 22,100 and 20,700 respectively. Purified EcoRuvA and MleRuvA gave a single band each on SDS-PAGE (Figure 8.6). Their apparent molecular weights were 27,000 and 24,000 respectively, both of which were abnormally high (Tsaneva *et al.,* 1992b).

For the molecular weight and matchpoint analyses, neutron Guinier plots were used to determine I(0) values for MleRuvA. The experimental data were obtained using Instrument D22 with 0%, 65% and 100% ^2H_2O buffers in a concentration range of 0.4-3.7 mg/ml (Figure 8.7). These plots were linear at the lowest Q values, and good Guinier fits in satisfactory Q.R_G ranges of 0.6-1.4 were obtained with a Q range of 0.16-0.49 nm^{-1}. No evidence of non-specific aggregation, association or dissociation phenomena was found from concentration series. No difference was observed between samples containing Mg^{2+} or EDTA. As MleRuvA has a low number of chromophores at 280 nm, its concentration could not be precisely determined by conventional absorbance measurements (Methods). Accordingly, relative neutron molecular weight calculations were based on the mean I(0)/c value in H_2O buffer of 0.061 ± 0.004 from c values based on absorbances and 0.065 ± 0.003 from c values based on Bradford assays, both normalised relative to the incoherent scattering of water as standard (Jacrot and Zaccai, 1981). This gave 120,000 ± 15,000 which was comparable with the value of 166,000 expected for octameric MleRuvA (Table 8.1). The analysis of relative √I(0) values as a

Purification of *M. Leprae* RuvA

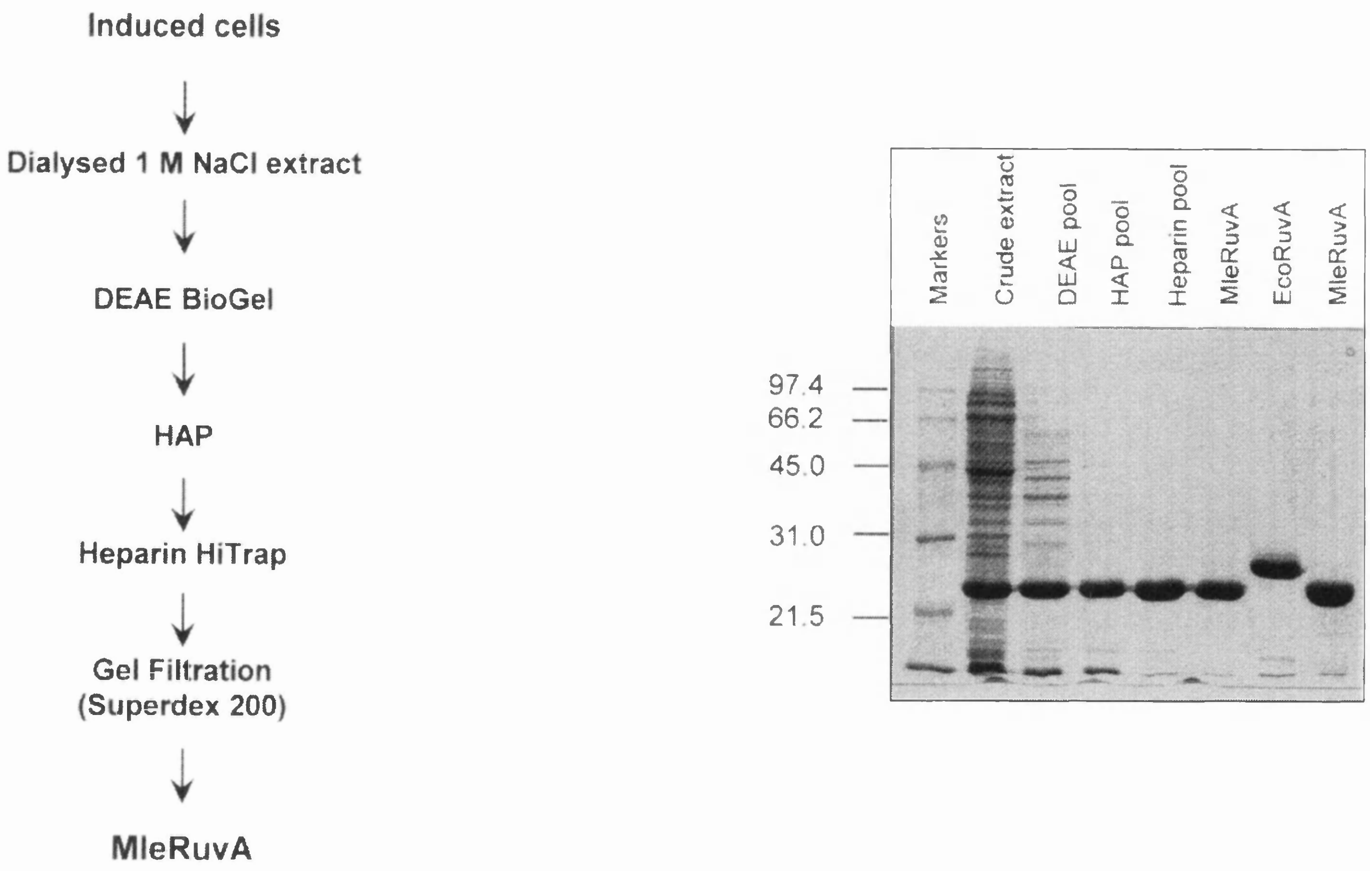


Figure 8.6: Flow chart and 15% SDS-PAGE gel detailing the purification scheme of recombinant *M. leprae* RuvA. This indicates the apparent molecular weights at 27,000 (EcoRuvA) and 24,000 (MleRuvA). Figure courtesy of Dr I.Tsaneva.

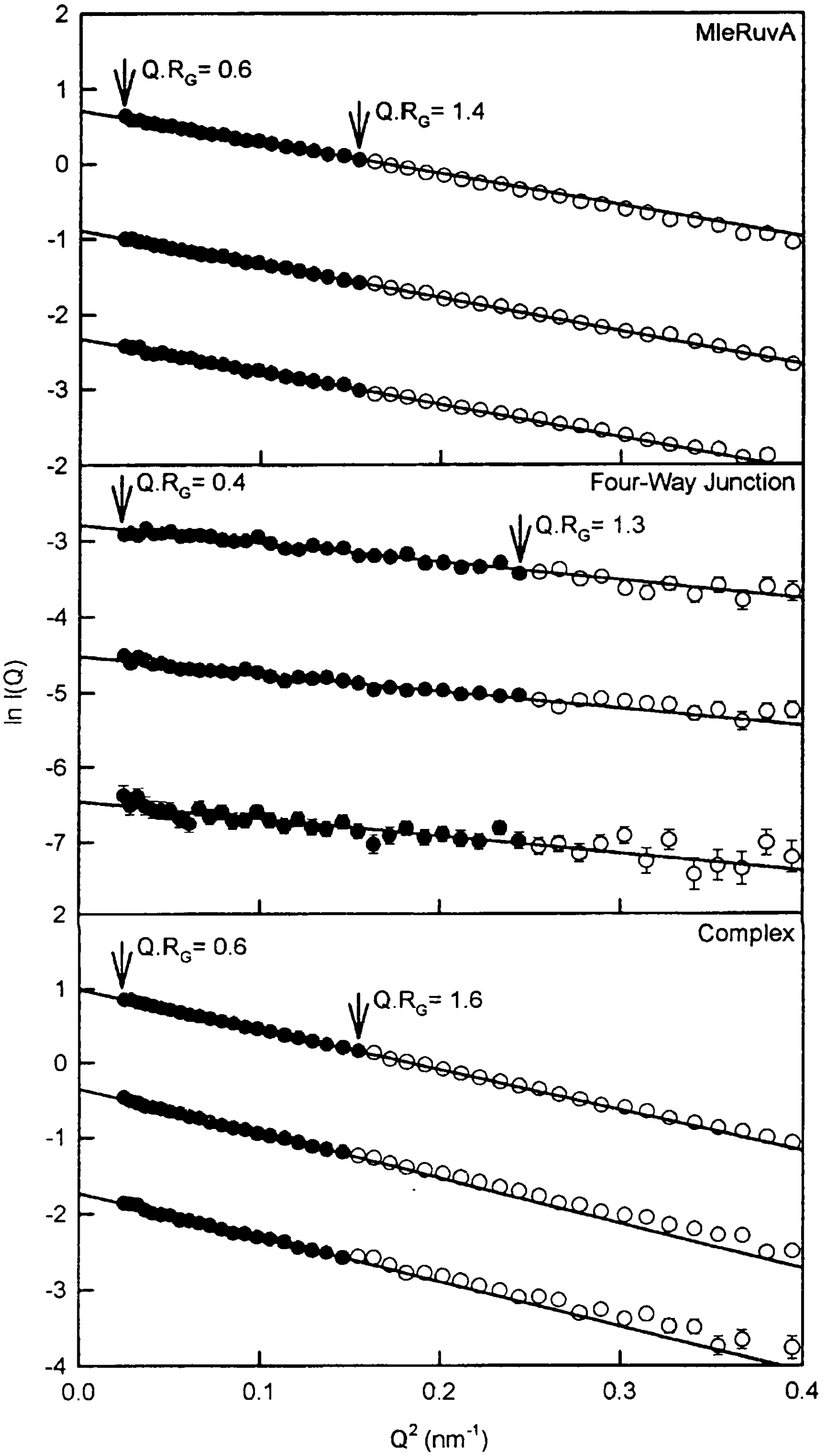


Figure 8.7: Neutron Guinier plots of ln I(Q) vs. Q^2 for MleRuvA, four-way junction and their complex in 100% ^2H_2O buffers. The concentrations of MleRuvA are 3.0, 1.0 and 0.3 mg/ml from top to bottom, while those for four-way junction are 2.4, 1.6 and 0.7 mg/ml, and those for the complex are 5.5, 2.2 and 0.5 mg/ml. Filled circles between the indicated Q.R_G ranges show the data points used to determine the R_G values (Table 8.1). Statistical error bars are shown when these are large enough to be visible.

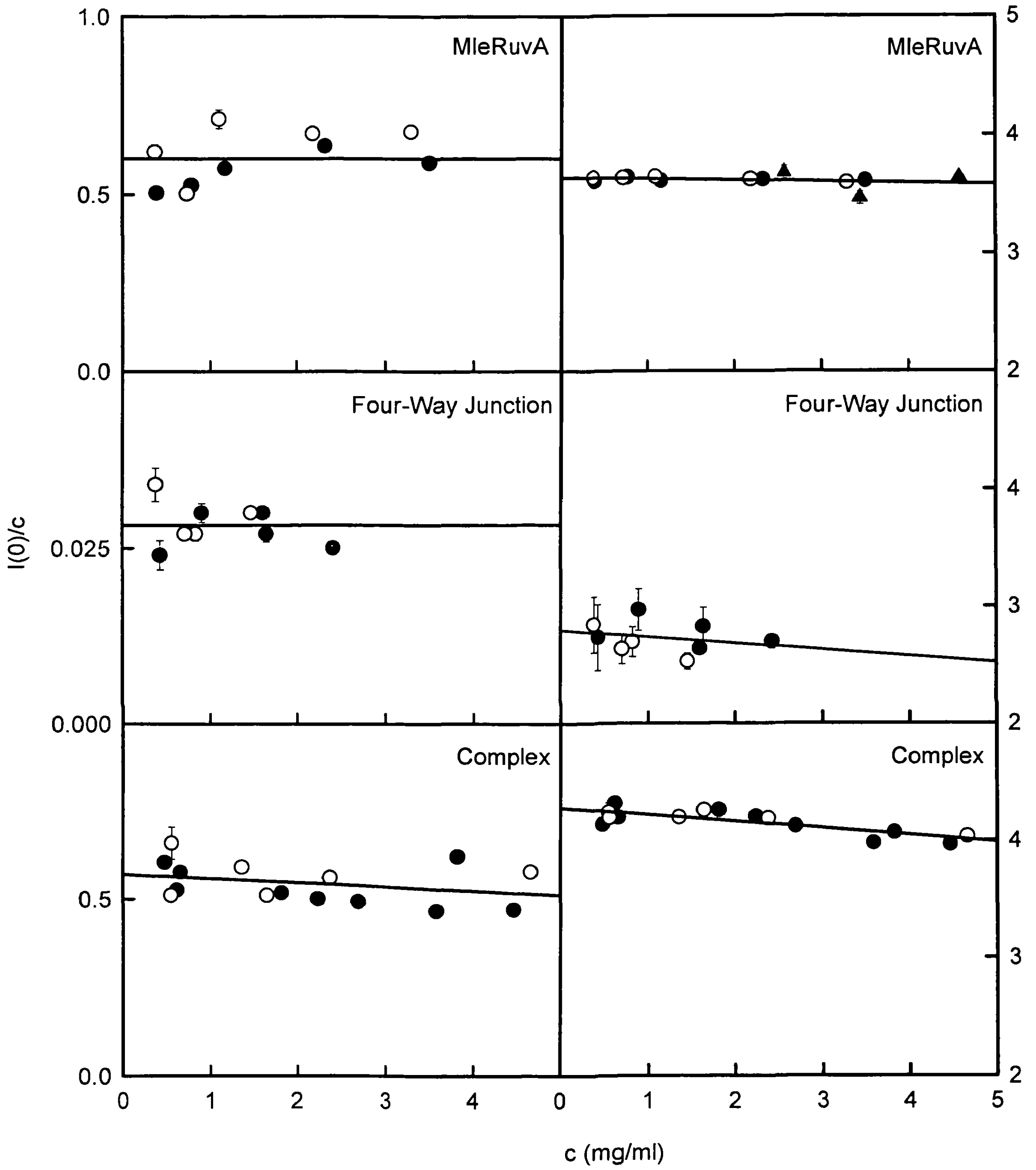


Figure 8.8: Concentration dependence of the Guinier R_G values and I0/c values as a function of the concentration c for MleRuvA, four-way junction and their complex in 100% ^2H_2O buffers. Statistical error bars are shown when these are large enough to be visible. ○, samples with Mg^{2+}; ●, samples with EDTA measured using D22; ▲, samples with EDTA measured using LOQ.

Figure 8.9 (continued): The Stuhrmann graph is the plot of the experimental R_G^2 against reciprocal contrast ($1/\Delta\rho$). The $R_{G\text{-}C}^2$ (y-axis intercept) is the radius of gyration at infinite contrast (i.e when $1/\Delta\rho = 0$) and corresponds to a particle of homogenous scattering density with the same shape and volume as the one in question. The term α is calculated from the slope of the line and reflects the radial distribution of internal scattering density fluctuations. The slope will be positive if there is an outer region of high scattering density surrounding a lower scattering density core, and negative if the reverse is true. Hence the R_G of glycoproteins is larger in positive contrasts (0% ^2H_2O) than in negative contrasts (80% ^2H_2O and 100% ^2H_2O) because the surface regions of carbohydrate and hydrophilic amino acids have a higher scattering density than that of the hydrophobic core (approximately equivalent to 50% ^2H_2O and 40% ^2H_2O, respectively). The value for α is therefore positive. β is a measure of the curvature of the parabola in the Stuhrmann plot. In practice, for globular proteins β is too small to be measured unless deuteration is used, and the Stuhrmann plot becomes simply a straight line of intercept $R_{G\text{-}C}^2$ and gradient α as seen in Figure 8.9.

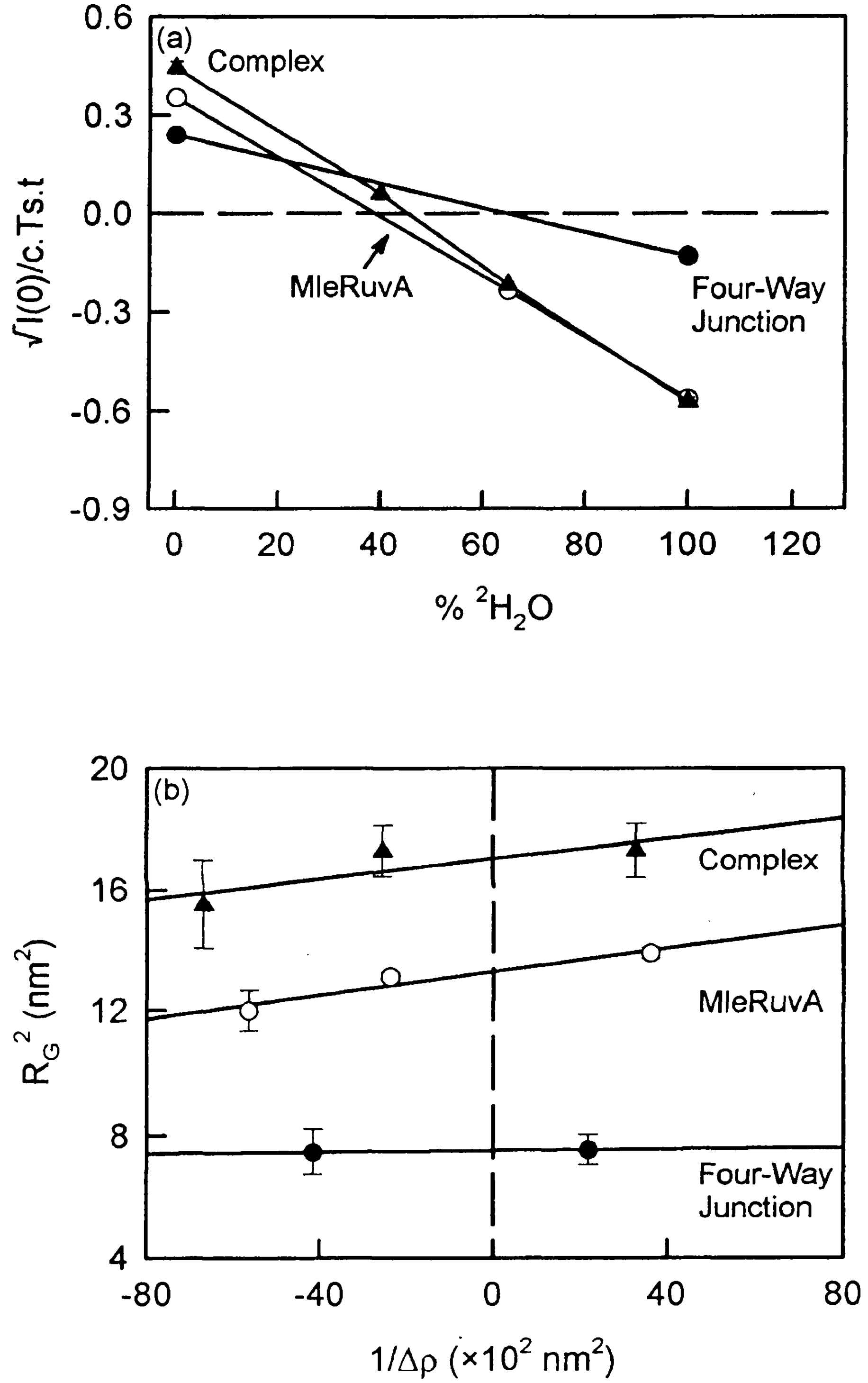


Figure 8.9: Contrast variation analysis of the Guinier I(0)/c and R_G values for MleRuvA, four-way junction and their complex. (a) The matchpoint analyses of √I(0)/c.t.T_s as a function of the ^2H_2O content in the solvent (v/v). (b) Stuhrmann analyses of $R_G{}^2$ as a function of the reciprocal solute-solvent contrast difference 1/Δρ. ●, four-way junction; ○, MleRuvA; ▲, complex. Further details are provided in Tables 8.1 and 8.2.

Table 8.1 Neutron scattering intensities for MleRuvA, four-way junction and their complex.

	From sequence	From experiment
Molecular weight		
MleRuvA	20,700 (monomer)	
	165,700 (octamer)	120,000 ± 15,000 (D22: in H_2O buffer)
		170,000 ± 5,000 (LOQ: in ^2H_2O buffer)
Four-way junction	39,400	42,000 ± 2,000 (D22: in H_2O buffer)
Complex	205,000	240,000 ± 20,000 (D22: in H_2O buffer)
Matchpoint[a]		
MleRuvA	39.6 % ^2H_2O	39.7 ± 0.7 % ^2H_2O (22)
Four-way junction	65.5 % ^2H_2O	65.4 ± 0.5 % ^2H_2O (12)
Complex	43.6 % ^2H_2O	43.9 ± 0.3 % ^2H_2O (29)

[a] Matchpoints are calculated from sequence data using standard neutron scattering lengths and unhydrated volumes and assuming full ^1H-^2H exchange in both protein and DNA components. The unhydrated protein $\bar{v}$ value is high at 0.783 ml/g for reason of the high content of Val and Leu residues in MleRuvA. The DNA $\bar{v}$ value of four-way junction is taken as 0.540 ml/g. The number of I(0)/c values used is bracketted.

Table 8.2 Experimental and modelled scattering analyses for MleRuvA, four-way junction and their complex.

Experimental (Figure 8.9b)	Guinier analyses[a] R_G (nm)	α (× 10^{-5})
MleRuvA (0% ${}^2H_2O$)	3.72 ± 0.02 (5)	
MleRuvA (65% ${}^2H_2O$)	3.46 ± 0.09 (6)	
MleRuvA (100% ${}^2H_2O$)	3.62 ± 0.02 (10)	
Infinite contrast	3.65 ± 0.02	19 ± 3
Four-way junction (0% ${}^2H_2O$)	2.75 ± 0.09 (2)	
Four-way junction (100% ${}^2H_2O$)	2.73 ± 0.13 (9)	
Infinite contrast	2.74 ± 0.06	1 ± 1
Complex (0% ${}^2H_2O$)	4.16 ± 0.11 (7)	
Complex (65% ${}^2H_2O$)	3.94 ± 0.18 (6)	
Complex (100% ${}^2H_2O$)	4.16 ± 0.10 (15)	
Infinite contrast	4.15 ± 0.03	15 ± 6

Final models (Figures 8.11 and 8.13)	Guinier analyses[a] R_G (nm)	R-factor[b] (%)	
Tetrameric MleRuvA	3.04	15.4	(D22)
		14.8	(LOQ)
Octameric MleRuvA (100% ${}^2H_2O$)	3.60	2.6	(D22)
	3.69	3.4	(LOQ)
Octameric MleRuvA (65% ${}^2H_2O$)	3.60	6.6	(D22)
Complex (65% ${}^2H_2O$)	4.25	6.9	(D22)

[a] The number of scattering curves measured for each sample is shown in brackets. The Q range used for the R_G determinations was 0.15-0.40 nm^{-1} for MleRuvA and its complex with four-way junction, and 0.15-0.50 nm^{-1} for four-way junction (Figure 8.7).

[b] The R-factor goodness-of-fit parameter corresponds to the neutron data in the Q range between 0.20 and 2.0 nm^{-1}, which is the Q range used to calculate the P(r) curves (Figure 8.10).

function of ^2H_2O gave a matchpoint of 38.9 ± 0.8% and 39.7 ± 0.7% ^2H_2O from the absorbance and Bradford determinations respectively (Figure 8.9). Both agreed with the matchpoint of 39.1% ^2H_2O calculated from the MleRuvA sequence (Perkins, 1986).

Neutron scattering data on MleRuvA at 2.6-4.6 mg/ml were also obtained on a different instrument LOQ. The mean I(0)/c value in 100% ^2H_2O was determined to be 0.188 ± 0.002 normalised relative to the scattering from a standard polymer. Comparison with a linear regression analysis of I(0)/c values for 11 other proteins measured on LOQ using 100% ^2H_2O buffers with molecular weights ranging between 27,000-254,000 (Ashton *et al.*, 1997) showed that MleRuvA had a molecular weight of 170,000 ± 5,000. Together with the linear Guinier plots and the matchpoint analyses, this proved that MleRuvA was octameric and monodisperse in solution.

For the structural analyses, the mean R_G values of MleRuvA in three contrasts were obtained from the Guinier analyses (Figure 8.7). No concentration dependence or difference between Mg^{2+} or EDTA was again observed (Figure 8.8). The Stuhrmann analysis of the R_G^2 values as a function of reciprocal contrast difference was linear with a positive slope as expected for globular proteins, and gave an R_C value at infinite contrast of 3.65 ± 0.02 nm and a slope α of 19 ± 3 ($\times$ 10^{-5}) (Table 8.2). The positive value of α corresponds to the location of hydrophilic residues with higher scattering densities in the surface regions of the protein. Since α is proportional to R_C^2, the use of α values for sixteen other compact globular proteins and glycoproteins with known R_C values showed that the expected value of α for RuvA would be 26 ± 22 ($\times$ 10^{-5}), in good agreement with that observed (Perkins, 1988a; Perkins *et al.*, 1993).

(8.3.2) Neutron Contrast Variation of Four-Way Junction by Guinier Analyses.

For the matchpoint analyses of the DNA junction, neutron Guinier plots for the I(0) values were obtained on D22 in 0% and 100% ^2H_2O only, for reason of signal-noise ratios, in a concentration range of 0.4-2.4 mg/ml. These plots were linear at the lowest Q values, and good Guinier fits in satisfactory $Q.R_G$ ranges of 0.3-1.3 were obtained with a Q range of 0.16-0.55 nm^{-1}. No concentration effects or dependence of I(0) or R_G on Mg^{2+} were detected (Figure 8.8). Neutron molecular weight calculations were based on the mean I(0)/c value in H_2O buffer of 0.026 ± 0.001. A value of 42,000 ± 2,000 was obtained which corresponded well with the calculated value of 39,300 from its sequence (Methods; Table 8.1). The analysis of relative $\sqrt{I(0)}$ values as a function of ^2H_2O gave a matchpoint of 65.4 ± 0.5% ^2H_2O (Figure 8.9). This agreed well with the matchpoint of 65.5% ^2H_2O calculated from the DNA sequence (Perkins, 1988a). These analyses showed that the junction was monodisperse, which was as expected from control experiments using neutral PAGE where the synthetic four-way junction migrated as a single species (not shown).

For the structural analyses of the DNA junction with 16 base pairs per arm, the mean R_G values in two contrasts were obtained from the Guinier analyses (Figure 8.7; Table 8.2). The R_G values are consistent with the open square-planar and compact stacked-X structures proposed for DNA junctions in the absence and presence of Mg^{2+} respectively (Lilley and Clegg, 1993), as the overall lengths of both structural types are similar and will lead to similar R_G values. The Stuhrmann analysis showed that the R_G^2 values were almost unchanged in 0% or 100% ^2H_2O, and this gave an R_C value at infinite contrast of 2.74 ± 0.06 nm and a slope α of 1 ± 1 ($\times 10^{-5}$).

(8.3.3) Neutron Contrast Variation of the MleRuvA-Junction Complex by Guinier Analyses.

The MleRuvA-junction complex was eluted from the gel-filtration column well separated from the free junction and overlapped slightly with free MleRuvA. The protein:DNA mass ratio of the complex was determined to be 3.4 ± 0.5 (mean ± standard deviation; 15 measurements) by a combination of protein Bradford assays and DNA absorbances. The expected ratio for an octamer of RuvA per junction was 4.21.

The matchpoint analysis of the I(0) values for the complex was performed using four contrasts in a concentration range of 0.5-7.0 mg/ml on D22 (Figure 8.9a). The Guinier plots were linear at the lowest Q values, and good fits in satisfactory Q.R_G ranges of 0.6-1.6 were obtained with a Q range of 0.16-0.49 nm^{-1}. No dependence of I(0) or R_G on Mg^{2+} was detected Fig 8.9a. This confirmed previous mobility shift experiments showing that in the presence of Mg^{2+} RuvA holds the junction in a square-planar conformation, similar to the conformation of the junction in the absence of bivalent metal ions (Parsons *et al.*, 1995). There was a small rise in the R_G values as the concentration decreased, which was consistent with a small interparticle interference effect caused by electrostatic interaction between different complex molecules (Figure 8.8). Neutron molecular weight calculations from the mean I(0)/c value in H_2O buffer of 0.082 ± 0.006 gave a value of 240,000 ± 20,000, which agreed well with the value of 205,000 calculated from sequences (Methods; Table 8.1). The analysis of relative √I(0) values as a function of ^2H_2O gave a matchpoint of 43.9 ± 0.3% ^2H_2O (Figure 8.9). This agreed well with a matchpoint of 43.6% ^2H_2O calculated from an octamer of RuvA per junction DNA (Perkins, 1986, 1988). In combination with the protein:DNA ratio, these analyses showed that free

MleRuvA was not detectable and that the complex was stable and monodisperse in solution.

For the structural analyses, the mean R_G values of the complex in three contrasts were obtained from the Guinier analyses (Figure 8.7). No concentration dependence or difference between Mg^{2+} or EDTA was again observed (Figure 8.8). The Stuhrmann analysis of the R_G^2 values as a function of reciprocal contrast difference was linear with a positive slope, and gave an R_C value at infinite contrast of 4.15 ± 0.03 nm and a slope α of 15 ± 6 (× 10^{-5}) (Table 8.2). The positive value of α corresponds to the location of residues with higher scattering densities in the surface regions of the complex. This value of α can be compared with those measured in seven neutron studies of nucleosomes from chromatin. These consist of a core of H2A, H2B, H3 and H4 histone proteins surrounded by almost two helical turns of DNA on the surface, and possess DNA contents ranging between 41-60% (w/w) and molecular weights of 150,000-235,000 (reviewed in Perkins, 1988a). Based on the R_C value of 4.15 nm for the MleRuvA complex, the α values for nucleosomes would result in a mean predicted value of α of 47 ± 5 (× 10^{-5}) for the MleRuvA complex if this had a similar protein-DNA arrangement. This high value compared to unbound MleRuvA is attributable to the scattering densities of DNA (65% ^2H_2O) relative to that of protein (40% ^2H_2O). The low value of α observed for the MleRuvA-junction DNA complex showed that the junction DNA was buried at the core of the complex, unlike the surface arrangement of DNA in nucleosomes.

(8.3.4) Contrast Variation Analysis of the Distance Distribution Functions.

The distance distribution function P(r) of MleRuvA and its complex with junction

DNA provides information on macromolecular dimensions (Figure 8.10). In 65% ^2H_2O where the DNA component was invisible, the P(r) curves for MleRuvA and the complex exhibited similar bell-shaped profiles. From the point at which P(r) becomes zero, the maximum length of unbound and complexed MleRuvA were found to be L1 = 9.5 nm and L2 = 12.5 nm (± 0.5 nm) respectively at large r. This showed that a large change had occurred in the protein structure. The most frequent distance in the protein corresponds to the maximum of the P(r) curve. This was also different at M1 = 4.3 nm for MleRuvA and M2 = 4.7 nm for the complex. It was noticeable that the outermost region of the P(r) curve for complexed MleRuvA was almost uniformly shifted to larger r values compared to unbound MleRuvA. Further examination of the P(r) curves in 100% ^2H_2O provided data on the full structures of MleRuvA, junction DNA and their complex, as the DNA component now contributed to the curve. Here, the P(r) curves for MleRuvA and the complex exhibited the same M1 and M2 values of 4.7 nm while the L1 and L2 values were different at 9.5 nm and 12.5 nm respectively. Importantly, the comparison with the L2 value for complexed MleRuvA in 65% ^2H_2O showed that the visibility of the DNA component in 100% ^2H_2O did not increase the value of L2. In combination with the result showing that MleRuvA is octameric, these data are most simply explained by postulating that the two tetramers associate directly with each other in unbound MleRuvA, while in the complex the two tetramers become separated by the insertion of junction DNA between them.

The P(r) curve of the unbound four-way junction showed a maximum M3 at 2.5 nm and an overall length L3 of 9.0 nm. The low value of M3 compared to that of L3 indicated that the junction DNA was not a compact globular structure in solution. If the

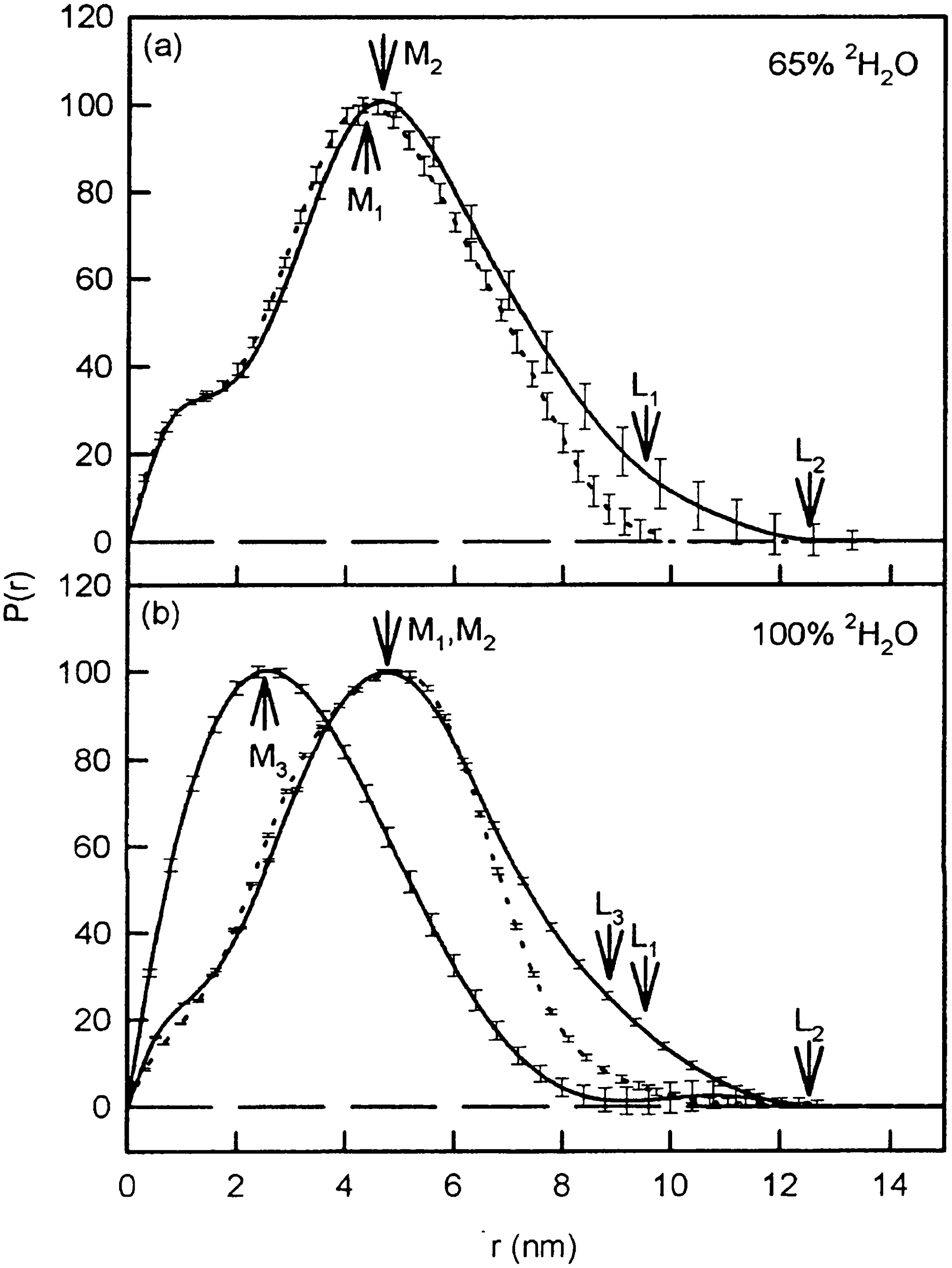


Figure 8.10: Distance distribution function analyses P(r) for MleRuvA and its complex with four-way junction. Peak heights were normalised to 100 for reason of clarity. Unbound MleRuvA, - - - -; complex and four-way junction, ———.

(a) The P(r) curves in 65% ^2H_2O correspond to the DNA matchpoint at which the four-way junction is invisible. M1 denotes the maximum in the P(r) curve of MleRuvA, and L1 denotes its maximum dimension. The corresponding values for the complex are denoted by M2 and L2.

(b) The P(r) curves in 100% ^2H_2O of MleRuvA, its complex and four-way junction, where the P(r) curve for junction DNA is denoted by M3 and L3.

spacing between adjacent base pairs is 0.34 nm in the commonly-occurring B-form of DNA, the equivalent length of two 16 base pair arms would be 10.9 nm, which is comparable with the value of L3. The slight decrease observed experimentally may reflect some flexibility in such a DNA structure. The P(r) curve also showed that the L3 value of 9.0 nm of junction DNA was comparable with that of 9.5 nm for unbound MleRuvA, and showed that the junction was sufficiently large to accommodate full interactions with MleRuvA in the complex.

(8.3.5) Molecular Modelling of the MleRuvA Structure.

In order to correlate the scattering curve of MleRuvA with the EcoRuvA crystal structure, the tetrameric EcoRuvA atomic structure was converted into a homology model of MleRuvA (Methods). The alignment in Figure 8.5 was used for simplicity as the incorporation of sequence gaps or insertions would not alter the outcome of the scattering curve fits. Single-density sphere models were used because of the relatively low dependence of the neutron curve on the contrast. A surface hydration monolayer of water molecules was not incorporated in the modelling since this is invisible in neutron scattering. Calculation of the scattering curve for a MleRuvA tetramer using previously calibrated procedures (Methods) gave a R_G value of 3.04 nm which was much less than the observed value of 3.65 nm. The shape of the calculated scattering curve for the tetramer was also quite different from the observed neutron curve in 100% ^2H_2O (Figure 8.11a). All combinations of tetramer and octamer models (below) showed no evidence for the presence of RuvA tetramers. Thus the existence of tetrameric MleRuvA in solution between 0.4-3.7 mg/ml was clearly ruled out.

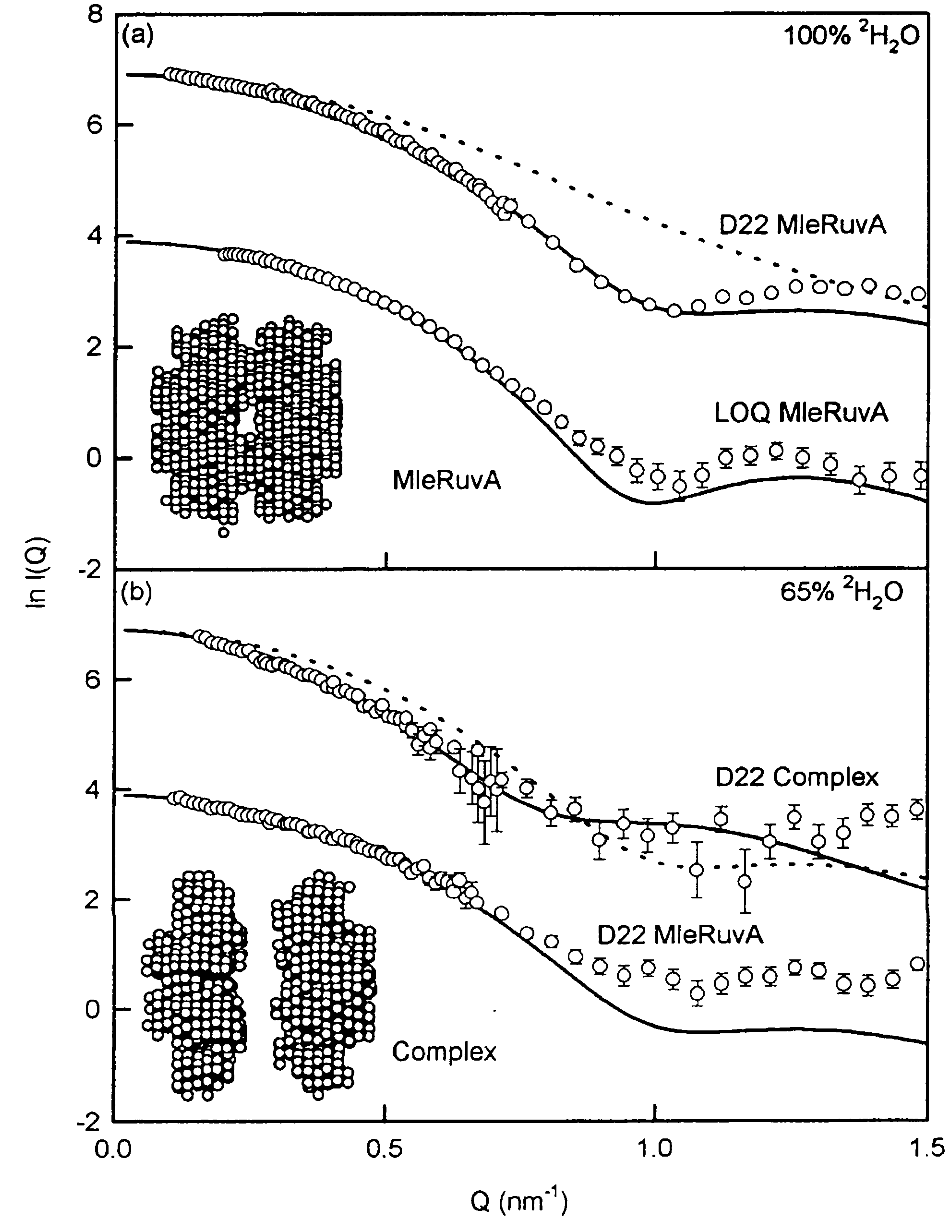


Figure 8.11: Scattering curve fits for the MleRuvA octamer and its complex with four-way junction. (a) The best-fit modelled curve for the MleRuvA octamer (continuous line) are compared with experimental data for MleRuvA obtained on Instruments D22 and LOQ (open circles). Experimental error bars are shown only when significant. The best-fit modelled curve was based on the EcoRuvA crystal structure (D22 fit) and the MleRuvA homology model (LOQ fit). The calculated curve for a tetramer is denoted by the dashed line. (b) Corresponding best curve fit for the MleRuvA complex with four-way junction (upper curve). For comparison, the curve fit of unbound MleRuvA is shown (lower curve). The dashed curve indicates the large difference between complexed and unbound MleRuvA. The inset shows a side view of the octameric sphere model used for the curve fits. The small deviations at large Q are attributable to small contributions from a flat incoherent scattering background.

The analyses of two MleRuvA tetramers as an octameric structure were constrained by the EcoRuvA crystal structure, the MleRuvA volume, the four-fold symmetry of the tetramer, and the presumed DNA-binding site in four grooves in the tetramer. Since each tetramer contained a four-fold symmetry axis at its centre, this defined the symmetry of the octamer. The tetramer has a flat DNA-binding face and a convex face. Only structures that interacted through either both DNA-binding or both convex faces were considered, as other arrangements would lead to the formation of indefinite assemblies of tetrameric RuvA. In order to assess such models, a translational search was made in which one tetramer was moved in 120 Z-axis translations of 0.1 nm relative to a second tetramer that was fixed in position at the half-way position of 6 nm (Figure 8.12a). The DNA-binding grooves were aligned parallel to each other in both tetramers. The 120 modelled curves were compared with the D22 scattering curve in 100% ^2H_2O, using the R-factor as a measure of the goodness of agreement. The R-factors were the lowest at 2.6% at two similar minima at Z-axis translations of 2.0 nm and 9.9 nm. The minima corresponded to the DNA-binding/DNA-binding and convex/convex face-to-face octamers respectively (Figure 8.12a). The R_G value at both minima was 3.60 nm. Between 6-8 models gave R_G values that were within ± 0.1 nm of the observed R_C value of 3.65 nm (Table 8.2), and this is equivalent to a positional precision of ± 0.3-0.4 nm. The total number of spheres in the two best-fit models was within 2% of the expected value at both R-factor minima, and Figure 8.12(a) showed that the two opposing tetrameric surfaces made contact with each other without steric overlap. Other curve fit searches tested the D22 scattering curve in 65% ^2H_2O and the LOQ scattering curves in 100% ^2H_2O with both the EcoRuvA crystal structure and the MleRuvA homology model. Similar minima resulted with the same translational precision

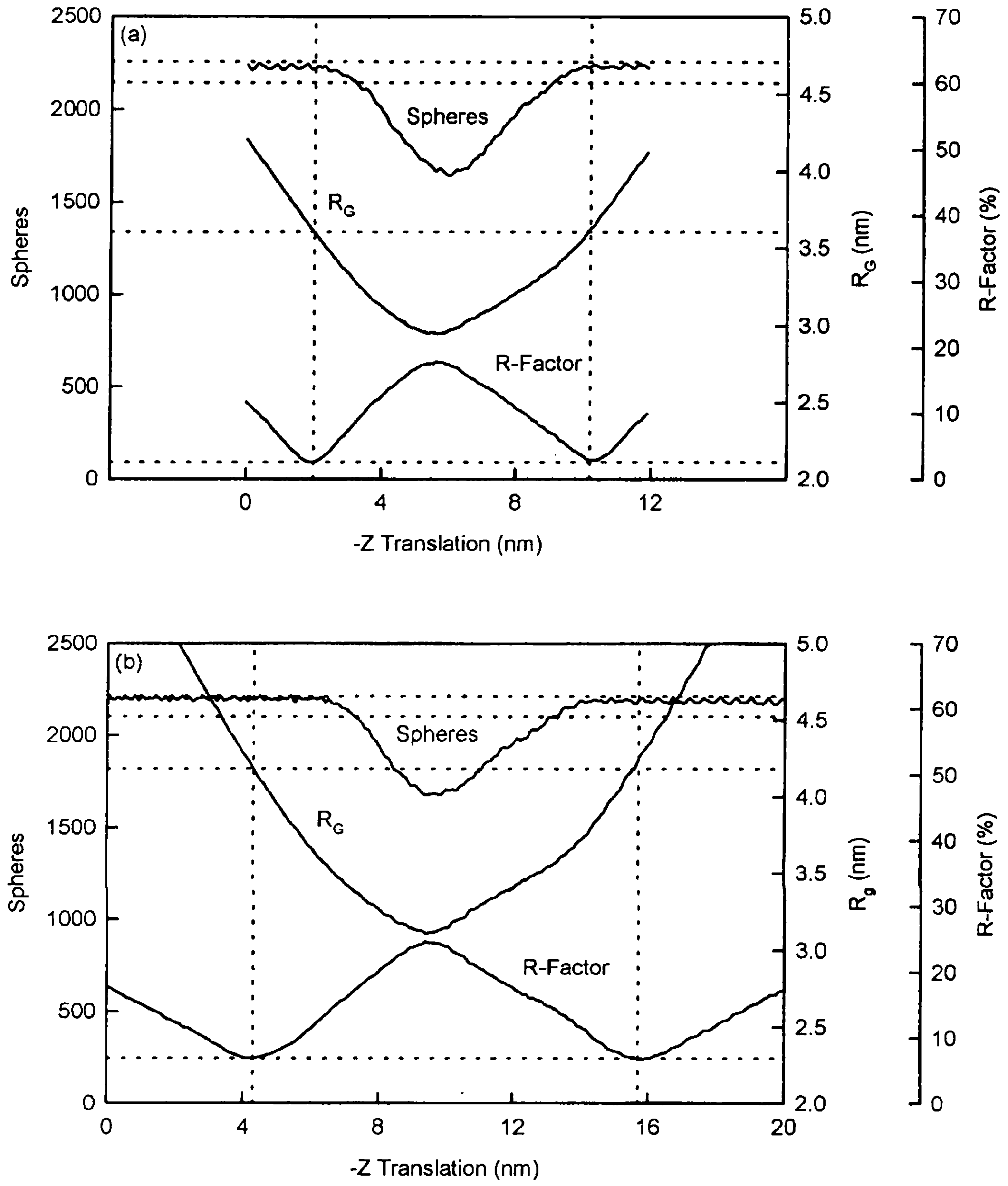


Figure 8.12: Modelling searches for the MleRuvA octamer and its complex with four-way junction. (a) The centres of two opposing MleRuvA tetramers were positioned on their common four-fold symmetry axis. One tetramer was held fixed at 6.0 nm, while the other tetramer was translated along this axis in 0.1 nm steps from 0 to 12 nm to pass through the centre of the fixed tetramer. 2212 spheres will result if there is no steric overlap between the two tetramers, and the two dashed lines indicate 0% and 5% steric overlap in the sphere model. The R_G of each model is compared with a dashed line indicating the experimental neutron R_G value of 3.65 nm. The R-factor of agreement between the neutron and calculated curves is compared with a dashed line indicating the two minima at an R-factor of 2.6%. The two vertical lines correspond to the two best fit structures for the MleRuvA octamer in which either both DNA-binding faces or convex faces are in contact with each other. (b) The corresponding search of two EcoRuvA tetramers to generate a structure for the complex with four-way junction. Here, the fixed tetramer is located at 10 nm, while the other was translated in a range of 0-20 nm in 0.1 nm steps. Other details are as in (a).

of ± 0.3 nm. The R-factor values from these fits were generally similar (Table 8.2), although it was increased to 6.6% for the 65% ^2H_2O curve fit for reason of the increased signal-noise ratio in this contrast. The R-factor of 2.6% is good compared with the range of 1.2%-8.7% found in other automated curve fit analyses (Perkins *et al.*, 1998).

Two possible structures for the octamer resulted from the analysis of Figure 8.12(a), and a choice could not be made between these on the basis of the scattering data. The DNA-binding/DNA-binding face-to-face octameric structure in Figure 8.13 was selected because this was the same as that determined for the MleRuvA-junction DNA complex below. If the reversed face-to-face octamer had been formed in unbound MleRuvA, the addition of junction DNA would result in the formation of polydisperse MleRuvA oligomers associated through both the DNA-binding faces as well as on the convex faces. Such polydispersity was not observed in the neutron data in Figures 8.7 or 8.8. The DNA-binding/DNA-binding octamer of Figure 8.13 resulted in the good D22 and LOQ curve fits in Figure 8.11(a). In these, the small I(Q) deviations at large Q are attributable to a small flat background due to incoherent scattering from the proton content in MleRuvA.

The side view of the final best fit model in Figure 8.13 provided an explanation of octamer formation. The two tetramers make limited contacts with each other through α-helix A6 of domain II which protrudes from the DNA-binding face. In particular, for reason of the four-fold symmetry of the tetramer, a pair of Lys117 and Glu121 residues on the same side of α-helix A6 from one tetramer is able to form a salt-bridge pair with an opposite pair of Glu121 and Lys117 residues from the other tetramer. In the best fit

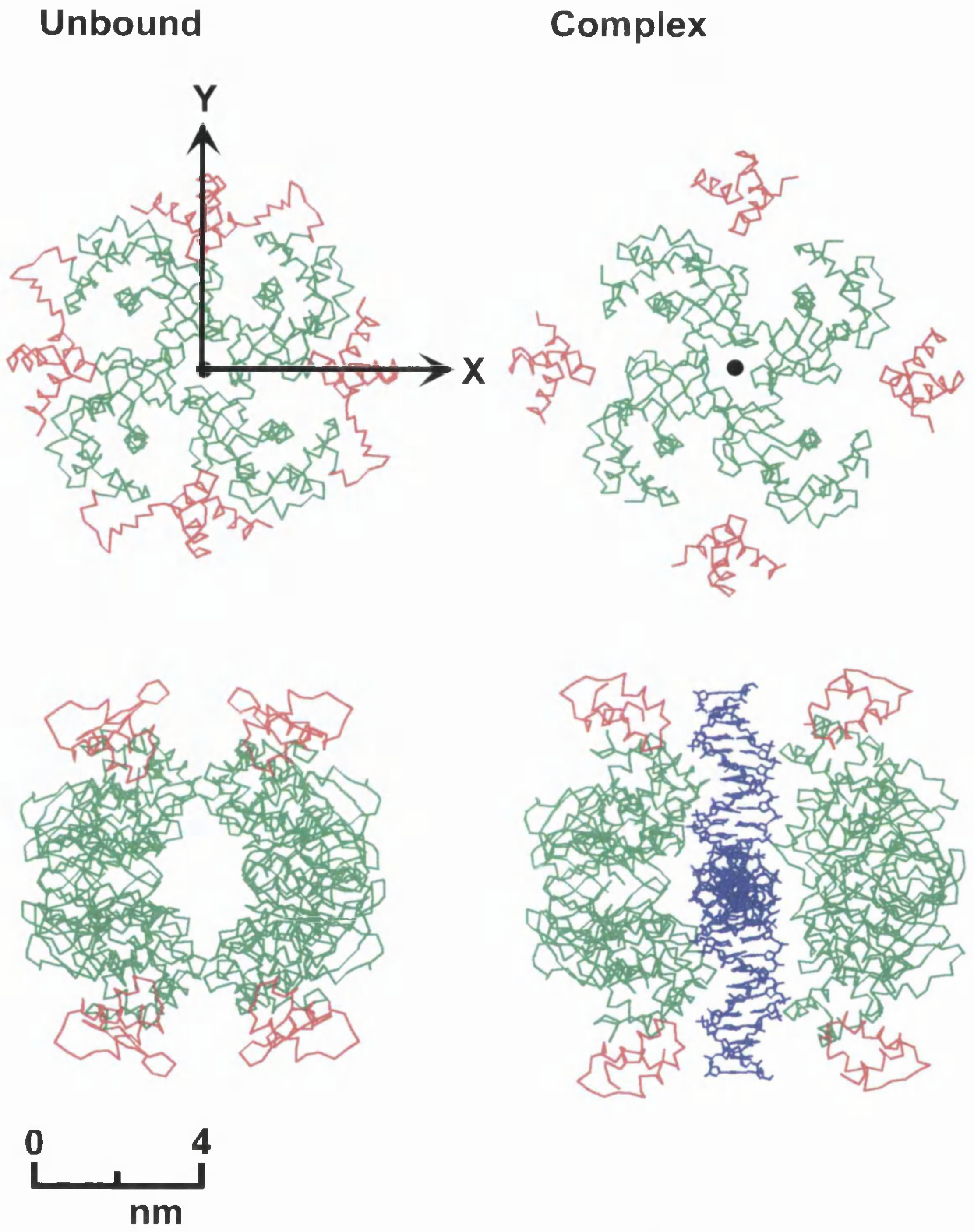


Figure 8.13: α-Carbon traces of the best-fit models for the unbound and complexed MleRuvA tetramer (upper: face-on views) and the unbound and complexed MleRuvA octamer (lower: side-on views). Domain III and the linker peptide with domain II are shown in red, while domains I and II are shown in green. The centre of each tetramer is indicated by a black symbol ● which corresponds to the Z-axis (Figure 8.12). Four B-DNA models with 12 base pairs (Brookhaven database code 1bna) were set perpendicular to each other (in blue) in the side-on view of the octameric complex to indicate how the tetramer separation determined from the best curve-fit structure was compatible with the width of the DNA double helix. For comparison with Figure 8.10, the tip-to-tip length of DNA as shown is 9 nm.

model of Figure 8.13, these pairs of α-carbon atoms were separated by 0.6 nm and their charged groups were separated by 0.4 nm, both of which are appropriate for salt bridge formation. The significance of Lys117 and Glu121 is shown by the high conservation of residues in α-helix A6 (Figure 8.5) and by the conservation of Lys117 and Glu121 in three other RuvA sequences (SwissProt codes RUVA_ECOLI, RUVA_HAEIN and RUVA_PSEAE). This implies that other RuvA molecules may form octamers in solution.

(8.3.6) Molecular Modelling of the Structure of the Complex.

The translational search of two MleRuvA tetramers for the complex in 65% ^2H_2O when the DNA component is invisible provided information on the position of the two tetramers in the complex. In this, one tetramer was moved in 200 steps of 0.1 nm relative to a second tetramer that was fixed at the half-way position of 10 nm (Figure 8.12b). The DNA-binding grooves were aligned parallel to each other in both tetramers. Here, two similar R-factor minima were obtained that corresponded to Z-axis translations close to 4 nm and 16 nm for the DNA-binding/DNA-binding and convex/convex face-to-face octamers respectively. The R_G values at both minima were in good agreement with the R_C value of 4.15 nm (Table 8.2). The total number of spheres in the two best-fit models coincided with their expected value. Comparison of Figures 8.12(a) and 8.12(b) showed that a void space existed between the two MleRuvA tetramers. Since the distance between the centres of the two tetramers had increased from 4.0 nm in free MleRuvA to 6 nm in complexed MleRuvA, the void space is 2 nm in width. This corresponds well with the diameter of the B-form DNA which is 2.2 nm (Brookhaven codes 1bna to 9bna).

The link between domains II and III in the EcoRuvA crystal structure is not visible in the electron density map (Rafferty *et al.*, 1996). This raised the possibility that either domain III may be relocated in the MleRuvA complex with junction DNA, or some other structural rearrangement may occur. This was assessed using a series of 200-model searches in which domain III was successively moved outwards from the centre of the tetramer in 20 × 0.1 nm steps. The best-fit 80 models (2%) of the 4200 that were tested showed that domain III was optimally positioned with an outward movement of 1.0 ± 0.6 nm from the position seen in the EcoRuvA crystal structure (Figure 8.13), while the position of the R-factor minimum was unchanged at 4.1 nm. A total of 146 models out of 4200 were within error of the scattering data. In 4200-model control searches that were performed for unbound MleRuvA in 65% and 100% ^2H_2O (Instrument D22), the best-fit 80 models for unbound MleRuvA corresponded to a different, smaller outward movement of domain III by 0.5 ± 0.3 nm. While the two standard deviations indicated some overlap, the overall analyses suggested that a minor conformational rearrangement of MleRuvA domains between the unbound and complexed states had occurred. The final model of Figure 8.13 is based on a Z-axis translation of 4.3 nm and a domain III translation of 0.8 nm, for which the 200-model search is shown in Figure 8.12(b). This gave an R_G value of 4.25 nm and an R-factor of 6.9% (Table 8.2). The dimensions of a standard B-DNA helix could be incorporated within this structure (Figure 8.13).

(8.4) Conclusions.

RuvA (and RuvB) homologues are universally found in prokaryotes. The sequence of RuvB is highly conserved, while RuvA homologues are more diverse, as exemplified by the 48% similarity of the MleRuvA and EcoRuvA sequences (Figure 8.5).

MleRuvA showed very similar DNA binding properties to those of EcoRuvA, and formed a functional branch migration and DNA helicase complex with *E. coli* RuvB *in vitro* (Judit Arenas-Licea, Anthony Keeley and Irina R. Tsaneva, manuscript in preparation). This neutron scattering study presents the first experimental data on the structure in solution of RuvA and its complex with a synthetic Holliday junction analogue. This demonstrated conclusively that unbound MleRuvA is a stable octamer in solution at protein concentrations of 0.4-3.7 mg/ml. It has been estimated that there are about 5600 monomers of EcoRuvA in an SOS-induced *E. coli* cell (West, 1997). If the dimensions of a single cell correspond to an ellipsoid of axes 1 μm × 0.5 μm × 0.5 μm, the concentration of RuvA would be 0.15 mg/ml, which is comparable to that used for the neutron measurements. If similar amounts of MleRuvA are present in *M. leprae* cells, then the octameric form of MleRuvA studied here is physiologically relevant. The formation of MleRuvA octamers was readily explained in terms of surface exposed Lys117 and Glu121 residues on the α-helix A6 of each monomer in the tetrameric EcoRuvA crystal structure that may form 8 salt bridges between the two tetramers in the absence of DNA (Figure 8.13). Previous gel filtration data showing EcoRuvA to be tetrameric (Tsaneva *et al.*, 1992b; Mitchell and West, 1994) are inconclusive by the nature of these experiments, and do not rule out the formation of octameric EcoRuvA in solution, even though this was crystallised as a tetramer.

In solution, the complex was also found to be an octamer of MleRuvA per DNA junction, the latter being sandwiched between two tetramers. The octameric MleRuvA-junction complex was stable in solution (Figure 8.8), and there was no evidence for other structural species in the concentration range used in the neutron experiments. This was true even in conditions that destabilised the binding of EcoRuvA to synthetic junctions, such as the presence of Mg^{2+} (Parsons and West, 1993). Electron microscopy studies of uranyl-stained and freeze-dried samples of the RuvAB branch migration complex of *E. coli in vacuo* were consistent with an octamer of EcoRuvA bound to the junction (Yu *et al.*, 1997). The present solution study of the MleRuvA-junction complex supports this stoichiometry, and shows in particular that the octameric complex does not require RuvB for its formation. A branch migration complex with two RuvA tetramers sandwiching the junction and two RuvB hexameric rings on two of the arms would constitute a well balanced molecular machine with symmetric points of contact between RuvA and RuvB.

Neutron scattering also provided information on the solution structures of unbound and complexed MleRuvA. The important outcome of this analysis was the large structural difference observed between unbound and complexed MleRuvA. This was particularly noticeable when measurements were performed in 65% ^2H_2O in which the DNA component is invisible. The scattering curve modelling of unbound MleRuvA used both the EcoRuvA crystal structure and a homology model of MleRuvA derived from it. The good curve fits obtained with a low R-factor of 2.6% showed that the unbound MleRuvA and EcoRuvA solution structures were very similar. The precision of the model was estimated as ± 0.3 nm, which was sufficient to show that the conserved residues Lys117 and Glu121 were close enough to form salt bridges between the two tetramers in the octamer. For complexed MleRuvA, the structural modelling showed that a void space was formed between the two RuvA tetramers. The modelling was precise enough to show that this void space was sufficient to accommodate the width of a B-DNA double helix. Further modelling was based on the optimisation of the position of domain III, which is required for the formation of a RuvAB complex in solution and of an active branch migration complex (Nishino *et al.,* 1998). This suggested that a minor conformational change had occurred upon binding of MleRuvA to the junction, although this analysis is not sufficiently precise to say whether this involved domain III or some other part of the structure, or an alteration of the four-fold symmetry of unbound RuvA. It is possible that such a conformational movement in RuvA would facilitate the correct formation or positioning of RuvB hexameric rings on the arms of the junction, and would imply a more complex role for RuvA in the assembly of the RuvAB molecular machine than previously suspected.

Taken together, the neutron data suggest that the MleRuvA octamer is a dynamic structure in solution, in which the buried DNA-binding surface of each tetramer is able to become transiently exposed for binding to DNA. A tetramer-octamer equilibrium would be shifted towards tetramer formation at very low RuvA concentrations, or alternatively RuvA may form tetramers through its interactions with RuvB. *E. coli* RuvA and RuvB form a complex in solution whose stoichiometry is not well defined (Shiba *et al.,* 1993; Mitchell and West, 1994; Nishino *et al.,* 1998). In addition, the interactions of RuvC with RuvA and with RuvB (Whitby *et al.,* 1996; van Gool *et al.,* 1998) could be important for the assembly of a RuvABC branch migration-resolution complex. Clearly, the detailed architecture of the RuvAB and/or RuvABC complex awaits further structural studies.

The present study provides another good illustration of the importance of automated constrained modelling methods for structural studies by solution scattering

(Perkins *et al.*, 1998). In general, scattering does not result in unique structures. However the use of tight constraints from known crystal structures when applied to an automated search of all structurally-allowed conformations will often result in a single, well-defined structural outcome. This was achieved in the present MleRuvA analysis with an estimated precision of ± 0.3 nm, which permitted a fuller interpretation of its crystal structure. The neutron scattering data demonstrated that a stable MleRuvA octamer was formed in solution, which was not expected from the tetrameric EcoRuvA structure seen by crystallography. The association of two flat disk-like structures will generally give two best-fit outcomes from scattering analyses (Figure 8.12; Ashton *et al.*, 1997), and a choice between them was required. The outcome for the unbound MleRuvA octamer was identified through the location of conserved charged residues in the best-fit model that readily explained the formation of octameric MleRuvA by putative salt bridges. The outcome for the complexed MleRuvA octamer was identified using the presumed buried location of the four DNA binding sites in each MleRuvA tetramer. The modelling enabled the structures of unbound and complexed MleRuvA to be visualised (Figure 8.13), in particular confirming the proposed mode of DNA binding to RuvA, and these will facilitate the planning of further experiments to explore function.

After this project was completed, initial reports of two crystal structures of EcoRuvA and MleRuvA complexed to junction DNA were presented at the IVth NACON meeting in Sheffield, U.K, 1998. As no crystals of unbound MleRuvA have been obtained to date, the present scattering analyses and modelling provide the only experimental evidence for its octameric structure. In confirmation of these solution structures for the MleRuvA-junction complex, both crystal structures showed that junction DNA was bound to one face of the RuvA tetramer, as proposed previously (Rafferty *et al.*, 1996). In the EcoRuvA structure, a single tetramer was bound to the junction (D. W. Rice, unpublished results), while the MleRuvA complex corresponded to an octamer of MleRuvA (L. H. Pearl, unpublished results) as seen in this solution analysis (Figure 8.13). It is highly unlikely that this represents a real difference between the two RuvA species. The final stoichiometry of the complex was obtained using a high protein-DNA ratio during its preparation. It is possible that a tetrameric complex may be assembled at a higher proportion of DNA, given that a RuvA tetramer bound to the junction is a likely intermediate in the assembly of the octameric complex. The EcoRuvA and MleRuvA structures were determined in complexes using junction DNA with 9 or 8 base pairs in each arm respectively, which are half the size of that used in the present solution study. Minor conformational changes in the MleRuvA tetramer were seen in the crystal structure of this complex. These were consistent with the present solution study, in which the larger size of the junction DNA should facilitate the full formation of the MleRuvA-DNA interactions (Figure 8.13).

CHAPTER 9

CONCLUSIONS

(9.1) 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 or models are exploited in full. 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. 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. The effect of macromolecular flexibility also needs to be taken into account for this type of 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 and can lead to a single family of related structures.

An application of curve fit analyses to analyse protein-protein complexes was developed for AmiC trimers (Chapter 5). The modelling of protein-protein complexes is less straightforward for reason of the absence of covalent links between the different subunits to constrain the models. Nonetheless the AmiC analyses were successfully performed by the use of constraints provided by symmetry considerations based on its known crystal structure, and this simplified the automated searches. The modelling study provided useful and unexpected information on the trimer formation of AmiC and its dependence on the ligand that is present. Whilst no conformation change could be detected when the ligand is changed from butyramide to acetamide, although the monomer-trimer equilibrium was shifted, it would be interesting to perform neutron scattering experiments on the AmiC-AmiR complex to see if a conformational change is detectable when using selectively deuterated AmiC.

The curve fit analyses for the multidomain protein complement factor I (Chapter 7) illustrated how an automated method for constrained modelling based on known homologous structures and the fixed connections between these structures can be easily applied. A limited family of good curve fits is filtered from a large number of possible models, and indicate molecular structures that are compatible with the scattering data. The positional precision of domains is estimated to be 1 nm by this approach. 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 factor I structure was shown to have a triangular heavy chain separated from the light chain by a linker which created a bilobal structure. The two lobes may perform different functions. Mutagenesis of the factor I sequence would allow further testing of this triangular domain arrangement by removing the serine protease domain and the linker region to see if the resulting heavy chain has a R_G of the expected size. The triangular arrangement could further be explored by mutating the 2 surface accessible cysteine residues in order to see if this causes the molecule to become more elongated and become inactive in assays.

A different application to investigate protein-protein and protein-DNA complexes was shown for MleRuvA octamers (protein-protein) and the MleRuvA-four-way junction complex (protein-DNA) (Chapter 8). Here, the automated curve fit approach was shown to work under conditions of neutron contrast matching when the DNA component became invisible. As with the AmiC analyses, the construction of MleRuvA octamers were successfully constrained by symmetry considerations based on the crystal structure of the tetramer, and this simplified the curve fitting. For reason of these stronger constraints the structural precision was estimated to be ± 0.3 nm. The modelling study provided biologically useful information on the spacing of the two tetramers in the RuvA octamer and their interactions by salt-bridges, since octamers may exist for other RuvA molecules. For the RuvA-four-way junction complex, it was possible to show that the DNA was buried within the RuvA octamer in a sandwich arrangement. Further investigations could possibly explore RuvAB interactions and RuvABC interactions in order to determine their solution structures.

All the modelling studies described here depend on the reliability of a procedure to calculate scattering curves from atomic coordinate models. The modelling studies described in Chapters 5, 7 and 8 are essentially based on a survey of electron and nuclear densities published in 1986by Perkins. This was found to work well in all the modelling analyses in this thesis (Chapters 5, 7 and 8). 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, 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 and it was concluded that this was not important for modelling purposes.

The calculation of scattering curves from coordinates also requires allowance for the instrumental geometry. This is unimportant for synchrotron X-ray cameras. The corrections are reasonably well characterised for Instruments D11 and D17 at the ILL (Ashton *et al.*, 1997), and this thesis shows that they appear to be applicable 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 although satisfactory curve fits were obtained in this thesis work, but satisfactory fits could nonetheless be obtained.

References.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Alignment Search Tool. *J. Mol. Biol.* **215**, 403-410.

Andersons, D., Engström, Å., Josephson, S., Hansson, L., and Steiner, H. (1991). Biologically-active and amidated cecropin produced in a baculovirus expression system from a fusion construct containing the antibody-binding part of protein-A. *Biochem. J.* **280**, 219-224.

Ashton, A. W., Boehm, M. K., Gallimore, J. R., Pepys, M. B. and Perkins, S. J. (1997). Pentameric and decameric structures in solution of serum amyloid P component by x-ray and neutron scattering and molecular modelling analysis. *J. Mol. Biol.* **272**, 408-422.

Ashton, A. W., Kemball-Cook, G., Johnson, D. J. D., Martin, D. M. A., O'Brien, D. P., Tuddenham, E. D. G., and Perkins, S. J. (1995). Factor VIIa and the extracellular domains of human tissue factor form a compact complex - a study by X-ray and neutron solution scattering. *FEBS Letters* **374**, 141-146.

Atkinson, A. E., Bermudez, I., Darlinson, M. G., Barnard, E. A., Earley, F. G. P., Possee, R. D., Beadle D. J., and King, L. A. (1993). Assembly of functional GABA(a) receptors in insect-cells baculovirius expression vectors. *Neuroreport* **3**, 597-600.

Bailey, M. J., McLeod, D. A., Kang, C.-L., and Bishop, D. H. L. (1989). Glycosylation is not required for the fusion activity of the G-protein of vesicular stomatitis-virus in insect cells. *Virology* **169**, 323-331.

Banner, D. W., D'Arcy, A., Chene, C., Winkler, F. D., Guha, A., Konigsberg, W. H., Nemerson, Y. and Kirchofer, D. (1996). The crystal structure of the complex of blood coagulation factor VIIa with soluble tissue factor. *Nature* **380**, 41-46.

Barlow, P. N., Baron, M., Norman, D. G., Day, A. J., Willis, A. C., Sim, R. B., and Campbell, I. D. (1991). Secondary Structure of a Complement Control protein module by 2-Dimensional H^1 NMR. *Biochemistry* **30,** 997-1004.

Barlow, P. N., Norman, D. G., Steinkasserer, A., Horne, T. J., Pearce, J., Driscoll, P. C., Sim, R. B., and Campbell, I. D. (1992). Solution Structure of the 5th Repeat of Factor-H - A 2nd Example of the Complement Control Protein Module. *Biochemistry* **31**, 3626-3634.

Baur, B., Hanselmann, K., Schlimme, W. and Jenni, B. (1996). Genetic-transformation in fresh-water - *Escherichia coli* is able to develop natural competence. *App. Env. Micro.* **62**, 3673-3678.

Beavil, A. J., Young, R. J., Sutton, B. J., and 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.

Bergh, M. L. E., Naranjo, C., Mentzer, A. F., Barsomian, G. D., Bartlett, C., Hirani, S., and Rasmussen, J.R. (1990). **IN** Welply, J. K., and Jaworski, E. (eds.). Glycobiology. *Wiley-Liss*, New York, p. 159-172.

Bilimoria, S.L. (1991). The biology of nuclear polyhedrosis viruses. **IN** Kurstak, E. (ed.). Viruses of Invertibrates. *Marcel Dekker Inc.* New York p. 1-72.

Boehm, M. K., Mayans, M. O., Thornton, J. D., Begent, R. H. J., Keep, P. A., and Perkins, S. J. (1996). Extended glycoprotein structure of the 7 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.

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett.* **286**, 47-54.

Bork, P. (1992). Mobile modules and motifs. *Curr. Opin. Struc. Biol.* **2**, 413-421.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins - mobile modules that cross phyla horizontally. *Proteins: Struct. Funct. Genet.* **17**, 363-374.

Bork, P. and Bairoch, A. (1995). Extracellular protein modules: A proposed nomenclature. *Trends Biochem. Sci.* **20**, Poster CO2.

Bork, P. and Doolittle, R. F. (1992). Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl. Acad. Sci. USA* **89**, 8990-8994.

Bottenus, R. E., Ichinose, A., and Davie, E. W. (1990). Nucleotide-sequence of the gene for the b-subunit of human factor-XIII. *Biochemistry* **29**, 11195-11209.

Bowie, J. U. and Eisenberg, D. (1993). Inverted protein structure prediction. *Curr. Opin. Struct. Biol.* **3**, 437-444.

Bradshaw, J. P. (1995). Neutrons for the Biologist. *Biologist* **42,** 178-182.

Brissett, N. (1997). Structural Studies of Domains in Aggrecan and Link Protein from Human Cartilage. *PhD Thesis*. University of London.

Brandstetter, H., Bauer, M., Huber, R., Lollar, P., and Bode, W. (1995). X-ray structure of clotting factor IXa: Active site and module structure related to Xase activity and hemophilia B. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 9796-9800.

Brown, T.A. (1991). Molecular Biology Labfax. *BIOS Scientific Publishers Ltd.*

Brown, T. and Brown, D. J. S. (1992). HPLC purification of synthetic DNA. *Methods Enzymol.* **211**, 20-35.

Buckholz, R. G. and Gleeson, M. A. G. (1991). Yeast systems for the commercial production of heterologous proteins. *Bio/tech.* **9**, 1067.

Busetta, B. and Hospital, M. (1982). An analysis of the prediction of secondary structures. *Biochim. Biophys. Acta*, **701**, 111-118.

Bustos, M. M., Luckow, V. A., Griffing, L. R., Summers, M. D., and Hall, T. C. (1988). Expression, glycosylation and secretion of phaseolin in a baculovirus system. *Plant Mol. Biol.* **10**, 475-488.

Campbell, I. D. and Downing, K. A. (1994). Building protein structure and function from modular units. *Trends Biotechnol.* **12,** 168-172.

Catterall, C. F., Lyons, A., Sim, R. B., Day, A. J., and Harris, T. J. R. (1987). Characterization of the primary amino acid sequence of human complement control protein factor I from an analysis of cDNA clones. *Biochem. J.*, **242**, 849-856.

Chamberlain, D., O'Hara, B. P., Wilson, S. A., Pearl, L. H., and Perkins, S. J. (1997). Oligomerisation of the amide sensor protein AmiC by X-ray and neutron scattering and molecular modeling. *Biochemistry* **36**, 8020-8029

Chamberlain, D., Keeley, A., Aslam,M., Judit Arenas-Licea, J., Tom Brown, T., Tsaneva, I. R., and 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. Submitted.

Chamberlain, D., Ullman, C. G., and Perkins, S. J. (1998). Structural arrangement of the five domains in human complement factor I by a combination of X-ray and neutron scattering and homology modelling. Submitted.

Chang, L. M. S., Rafter, E., Rusquet, V. R., Peterson, R. C., White, S. T., and Bollum, F. J. (1988). Expression and processing of recombinant human terminal transferase in the baculovirus system. *J. Biol. Chem.* **263**, 12509-12513.

Chang, R. (1981). Physical Chemistry with Applications to Biological Systems. Second Edition. *Macmillan* New York p. 49-63; 512-519.

Chazin, W.J., Hugli, T.E., and Wright, P.E. (1988). H^1 NMR Studies of Human C3a Anaphylatoxin in Solution - Sequential Resonance Assignments, Secondary

Structure, and Global Fold. *Biochemistry* **27**, 9139-9148.

Chothia, C. (1975) Structural invariants in protein folding. *Nature* **254**, 304-308.

Chothia, C. (1992). Proteins - 1000 families for the molecular biologist. *Nature* **357**, 543-544.

Chothia, C., and Finkelstein, A. V. (1990). The Classification and Origins of Protein Folding Patterns. *Annu. Rev. Biochem.* **59**, 1007-39.

Chou, P. Y. and 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.

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.

Craik, C. S., Choo, Q-L., Swift, G. H., Quinto, C., and Rutter W. J. (1984). Structure of 2 related rat pancreatic trypsin genes. *J. Biol. Chem.* **259**, 14255-14264.

Creighton, T. E. (1993). Proteins: structure and molecular properties (2nd Edition). *W. H. Freeman and Company,* New York.

Crook, N.E. (1991). Baculoviridae: Subgroup B. Comparative aspects of granulosis viriuses. **IN** Kurstak, E. (ed.). Viruses of Invertibrates. *Marcel Dekker Inc.* New York p. 73-109.

Crossley, L. G. (1980). C3b inactivator and β1H. *Methods Enzymol.* **80**, 112-124.

Daly, N. L., Djordjevic, J. T., Kroon, P. A., and Smith, R. (1995a). Three-dimensional structure of the second cysteine-rich repeat from the human low-density lipoprotein receptor. *Biochemistry* **34**, 14474-14481.

Daly, N. L., Scanlon, M. J., Djordjevic, J. T., Kroon, P. A., and Smith, R. (1995b). Three-dimensional structure of a cysteine-rich repeat from the low-density lipoprotein receptor. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 6334-6338.

Davidson, D. J., Fraser, M. J., and Castellino, F. J. (1990). Oligosaccharide processing in the expression of human plasminogen cDNA by lepidopteran insect (*Spodoptera frugiperda*) cells. *Biochemistry* **29**, 5584-5590.

Davis, G. T., Bedzyk, W. D., Voss, E. W., and Jacobs, T. W. (1991). Single-chain antibody (SCA) encoding genes: one step construction and expression in eukaryotic cells. *Bio/Tech.* **9**, 165.

Davis, T. R., Trotter, K. M., Granados, R.R. and Wood, H. A (1992). Baculovirus Expression of Alkaline Phosphatase as a Reporter Gene for Evaluation of Production, Glycosylation, and Secretion. *Bio/Tech.* **10**, 148-1150.

de Chateau, M. and Bjorck, L. (1994). Protein PAB, a mosaic albumin-binding bacterial protein representing the first contemporary example of module shuffling. *J. Biol. Chem.* **269**, 12147-12151.

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.

Deng, W. P., and Nickoloff, J. A. (1992). Site-directed mutagenesis of virtually any plasmid by eliminating a unique site. *Anal. Biochem.* **200**, 81-88.

DiScipio, R. G. (1992). Ultrastructures and interactions of complement factors H and I. *J. Immunol.* **149**, 2592-2599.

Delchambre, M., Gheysen, D., Thines, D., Thiriart, C., Jacobs, E., Verdin, E., Horth, A., Burny, A., and Bex, F. (1989). The gag precursor of simian immunodeficiency virus assembles into virus-like particles. *EMBO J.* **8**, 2653-2660.

Desprès, P., Dietrich, J., Girard, M., and Bouloy, M. (1991a). Recombinant baculoviruses expressing Yellow fever virus E and NS1 proteins elicit protective immunity in mice. *J. Gen. Virol.* **72**, 2811-2816.

Desprès, P., Girard, M., and Bouloy, M. (1991b). Characterization of yellow-fever virus proteins E and NS1 expressed in Vero and *Spodoptera frugiperda* cells. *J. Gen. Virol.* **72**, 1331-1342.

Devlin, J. J., Devlin, P. E., Clark, R., O'Rourke, E. C., Levenson, C., and Mark, D. F. (1989). Novel expression of chimeric plasminogen activators in insect cells. *Bio/Technology* **7**, 286-292.

Dodds, A. W. (1993). Small-scale preparation of complement components C3 and C4. *Meth. Enzymol.* **223,** 46-61.

Doi, R. H., Wong, S.-L., and Kowamura, F. (1986). Potential use of *Bacillus subtilis* for secretion and production of foreign proteins. *Trends Biotechnol.* **4**, 232.

Döller, G. (1985). The safety of insect viruses as biological control agents. **IN** Maramorosch, K. and Sherman, K.E. (eds.). Viral Insecticides for Biological Control. *Academic Press*, London p.399-439.

Domingo, D. L., and Trowbridge, I. S. (1988). Characterization of the human transferrin receptor produced in a baculovirus expression system. *J. Biol. Chem.*, **263**, 13386-13392.

Doniach, S., Bascle, J., Garel, T., and Orland H. (1995). Partially Folded States of Proteins: Characterization by X-ray Scattering. *J. Mol. Biol.* **254**, 960-967.

Donavan, J. W., and Milhalyi, E. (1974). Conformation of Fibrinogen: Calorimetric Evidence for a Three-Nodular Structure. *Proc. Natl. Acad. Sci. USA* **71**, 4125-4128.

Doolittle, R. F. (1985). The genealogy of some recently evolved vertebrate proteins. *Trends Biochem. Sci.* **10**, p. 233-237.

Doolittle, R. F., Johnson, M. S., Husain, I., van Houten, B., Thomas, D. C., and Sancar, A. (1986). Domainal evolution of a prokaryotic DNA repair protein and its relationship to active transport proteins. *Nature* **323**, 451-453.

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, W. F. (1978). Genes in pieces: where they ever together? *Nature* **272**, 581-582.

Dowbenko, D. K., Diep, A., Taylor, B. A., Luis, A. J., and Lasky, L. A. (1991). Characterization of the murine homing receptor gene reveals correspondence between protein domains and coding exons. *Genomics* **9**, 270-277.

Drake, A. F. (1994). Circular Dichroism. **IN** Jones, C., Mulloy, B., and Thomas, A.S. (eds.) Methods in Molecular Biology, Vol. 22: Microscopy, Optical Spectroscopy, and Macroscopic Techniques. *Humana Press Inc.*Totowa, N.J p.219-244.

Drew, R. E. and Wilson, S. A. (1992) **In** Galli, E., Silver, S. and Witholt, E. (eds). *Pseudomonas*: Molecular Biology and Biotechnology. *American Society of Microbiology* (Washington D.C) p. 207-213.

Dunderdale, H. J., Benson, F. E., Parsons, C. A., Sharples, G. J., Lloyd, R. G. and West, S. C. (1991). Formation and resolution of recombination intermediates by E. coli RecA and RuvC proteins. *Nature*, **354**, 506-510.

Durchschlag, H. and Zipper, P. (1988). Primary and post-irradiation inactivation of the sulfhydryl enzyme malate synthase: correlation of protective effects of additives. *FEBS Letters* **237**, 208-212.

Edwards, Y. J. K. and 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. Molec. Biol.* **260,** 277-285.

Egelman, E. H., Yu, X., Wild, R., Hingorani, M. M., and Patel, S. S. (1995). Bacteriophage T7 helicase/primase proteins form rings around single-stranded DNA that suggest a general structure for hexameric helicases. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 3869-3873.

Eggleston, A. K., Mitchell, A. H., and West, S. C. (1997). *In vitro* reconstitution of the late steps of genetic recombination in *E. coli.* *Cell*, **89**, 607-617.

Engel, A. M., Cejka, Z., Lupas, A., Lottspeich, F., and Baumeister, W. (1992). Isolation and cloning of omp-alpha, a coiled-coil protein spanning the periplasmic space of the ancestral eubacterium *Thermotoga maritima.* *EMBO J.* **11**, 4369-4378.

Evans, J. N. S (1995). Biomolecular NMR Spectroscopy. *Oxford University Press* p. 5-54; 55-75.

Fass, D., Blacklow, S., Kim, P. S. and Berger, J. M. (1997). Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module. *Nature* **388**, 691-693.

Fesik, S.W., Gampe, R.T., Zuiderweg, E.R.P., Kohlbrenner, W.E., and Weigl D. (1989). Heteronuclear 3-Dimensional NMR-Spectroscopy Applied to CMP-KDOSynthetase (27.5 KD). *Biochem. Biophys. Res. Comm.* **159**, 842-847.

Fetler, L., Tauc, P., Hervé, G., Moody, M.F., and Vachette, P. (1995). X-ray Scattering Titration of the Quaternary Structure Transition of Aspartate Transcarbamylase with a Bisubstrate Analogue: Influence of Nucleotide Effectors. *J. Mol. Biol.* **251**, 243-255.

Fitzpatrick, P. F., Chlumsky, L. J., Daubner, S. C., and O'Malley, K. L. (1990). Expression of rat tyrosine-hydroxylase in insect tissue-culture cells ad purification and characterization of the cloned enzyme. *J. Biol. Chem.*, **265**, 2042-2047.

Fong, H. K. W., Hurley, J. B., Hopkins, R. S., Miake-Lye, R., Johnson, M. S., Doolittle, R. F., and Simon, M. I (1986). Repetitive segmental structure of the transducin beta-subunit - homology with the cdc4 gene and identification of related messenger RNAs. *Proc. Natl. Acad. Sci. USA* **83**, 2162-2166.

Foster, S. J. (1993). Molecular analysis of 3 major wall-associated proteins of *Bacillus subtilis*168 - Evidence for processing of the product of a gene encoding a 258-kda precursor 2-domain ligand-binding protein. *Mol. Microbiol.* **8**, 299-310.

Fraker, P. J. and Speck, J. C. (1978). Protein and cell membrane iodinations with a sparingly soluble chloroamide 1, 3, 4, 6-tetrachloro-3a, 6a-diphenylglycoluril. *Biochem. Biophys. Res. Commun.* **80,** 849.

Francki, R.I.B., Fauquet, C.M., Knudson, D.L., and Brown, F. (eds.) (1991). Classification and Nomenclature of Viruses, Fifth report of the international comittee on taxonomy of viruses. *Archives of Virology*, **Supplementum 2,** *Springer-Verlag*, Wein, New York.

Franke, W. W., Goldschmidt, M. D., Zimbelmann, R., Mueller, H. M., Schiller, D. L., and Cowin., P. (1989). Molecular-cloning and amino-acid sequence of human plakoglobin, the common junctional plaque protein. *Proc. Natl. Acad. Sci. USA* **86**, 4027-4031.

Freeman, M, Ashkenas, J., Rees, D. J. G., Kingsley, D. M., Copeland, N. G., Jenkins, N. A., and Krieger, M. (1990). An ancient, highly conserved family of cysteine-rich protein domains revealed by cloning Type-I and Type-II Murine Macrophage Scavenger Receptors. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8810-8814

Frey, M., Sieker, L., Payan, F., Haser, R., Bruschi, M., Pepe, G., and LeGall, J. (1987). Rubredoxin from *Desulfovibrio gigas*. A molecular model of the oxidized form at 1.4 Å resolution. *J. Mol. Biol.* **197,** 325-41.

Friedman, A.M., Fischmann, T. O., and Steitz, T. A. (1995). Crystal Structure of *lac* Repressor Core Tetramer and Its Implications for DNA Looping. *Science* **268,** 1721-1727.

Fujii, T. and Miyashita, K. (1993). Multiple domain-structure in a chitinase gene (chic) of *Streptomyces lividans*. *J. Gen. Microbiol.* **139**, 677-686.

Fulton, S. and Vanderburgh, D. (1996). The Busy Researcher's Guide to Biomolecule Chromatography. *PerSeptive Biosystems.*

Garnier, J., Osguthorpe, D. J. and 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.

Garnier, J. and Robson, B. (1989). The GOR method. **IN** Fasman, G. D. (ed.) Prediction of Protein Structure and the Principles of Protein Conformation. p 417-465, *Plenum Press*, New York.

George, S. T., Arbabian, M. A., Rucho, A. E., Kiely, J., and Malbon, C. C. (1989). High-efficiency expression of mammalian beta-adrenergic receptors in baculovirus-infected insect cells. *Biochem. Biophys. Res. Comm.* **163**, 1265-1269.

Ghosh, R. E. (1989) A Computing Guide for Small Angle Scattering Experiments. Institut Laue Langevin Internal Publication 89GH02T.

Gibrat, J. F., Garnier, J., and 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.

Gilbert, W. and Glynias, M. (1993). On the ancient nature of introns. *Gene* **135**, 137-144.

Gilbert, W. (1978). Why genes in pieces? *Nature* **271**, 501.

Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller Jr, R. C., and Warren, R. A. J. (1991). Domains in Microbial ß-1,4-Glycanases: Sequence Conservation, Function, and Enzyme Families. *Microbiol. Rev.* **55**, 303-315.

Gilson, E., Higgins, C. F., Hofnung, M., Ferro-Luzzi Ames G., and Nikaido, H. (1982). Extensive homology between membrane-associated components of histidine and maltose transport systems of *Salmonella typhimurium* and *Escherichia coli.* *J. Biol. Chem.* **257**, 9915-9918.

Glatter, O., and Kratky, O. (eds.) (1982). Small Angle X-ray Scattering. *Academic Press*, New York.

Goldberger, G., Arnaout, M. A., Aden, D., Kay, R., Rits, M. and Colten, H. R. (1984). Biosynthesis and postsynthetic processing of human C3b/C4b inactivator (Factor I) in three hepatoma cell lines. *J. Biol. Chem.* **259,** 6492-6497.

Goldberger, G., Bruns, G. A. P., Rits, M., Edge, M. D. and Kwiatkowski, D. J. (1987). Human complement factor I: analysis of cDNA-derived primary structure and assignment of its gene to chromosome 4. *J. Biol. Chem.* **262**, 10065-10071.

Golding, G. B., Tsao, N., and Pearlman, R. E. (1994). Evidence for intron capture - an unusual path for the evolution of proteins. *Proc. Natl. Acad. Sci. USA* **91**, 7506-7509.

Grabowski, G. A., White, W. R, and Grace, M. E. (1989). Expression of functional human acid beta-glucosidase in COS-1 and *Spodoptera frugiperda* cells. *Enzyme* **41**, 131-142.

Granados, R.R. and Hashimoto, Y. (1989). Infectivity of baculoviruses to cultured cells. **IN** Mitsuhashi, J. (ed.) Invertibrate Cell System Applications. *CRC press, Inc.*, Boca Raton, Florida, Vol I p. 167-81.

Greenfield, C., Patel, G., Clark, S., Jones. N, and Waterfield, M. D. (1988). Expression of the human EGF receptor with ligand-stimulatable kinase-activity in insect cells using a baculovirus vector. *EMBO J.* 7, 139-146.

Greenwald, I. (1985). Lin-12, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth-factor. *Cell* **43**, 583-590.

Gregory, L., Davis, K.G., Sheth, B., Boyd, J., Jefferis, R., Nave, C., and Burton, D.R. (1987). The solution conformations of the subclassesof IgG deduced from sedimentation and small angle x-ray scattering studies. *J. Mol. Immunol.* **24**, 821-829.

Gruenwald, S.G., and Heitz, J. (1993). Baculovirius Expression Vector System: Procedures & Methods Manual. Second Edition. *PharMingen*, San Diego, CA.

Guinier, A. and Fournet, G. (1955) Small angle scattering of X-rays. Wiley, New York.

Guss, B., Eliasson, M., Olsson, A., Uhlen, M., Frej, A. K., Jornvall, H., Flock, J. I., and Lindberg, M. (1986). Structure of the IgG-binding regions of streptococcal protein G. *EMBO J.* **5**, 1567-1575.

Hamada, K., Bethge, P.H., and Mathews, F.S (1995). Refined Structure of Cytochrome $b_{562}$from *Escherichia coli* at 1.4 Å Resolution. *J. Mol. Biol.* **247**, 947-962.

Hammarberg, B., Moks, T., Tally, M., Elmblad, A., Holmgren, E., Murby, M, Nilsson, B., Josephson, S., Uhlen, M. (1990). Differential stability of recombinant human insulin-like growth factor-II in *Escherichia coli* and *Staphylococcus aureus.* *J. Biotechnol.* **14**, 423-438.

Hammock, B. D., Bonning, B. C., Possee, R. D., Hanzlik, T. N., and Maeda, S. (1990). Expression and effects of the juvenile-hormone esterase in a baculovirus vector. *Nature* **344**, 458-461.

Hansen, C. K. (1992). Fibronectin type III-like sequences and a new domain type in prokaryotic depolymerases with insoluble substrates. *FEBS Lett.* **305**, 91-96.

Harding, S.E. (1994a). Determination of Macromolecular Homogeneity, Shape, and Interactions Using Sedimentation Velocity Analytical Ultracentrifugation. **IN** Jones, C., Mulloy, B., and Thomas, A.S. (eds.) Methods in Molecular Biology, Vol. 22: Microscopy, Optical Spectroscopy, and Macroscopic Techniques. *Humana Press Inc.*Totowa, N.J p. 61-73.

Harding, S.E. (1994b). Determination of Absolute Molecular Weights Using Sedimentation Equilibrium Analytical Ultracentrifugation. **IN** Jones, C., Mulloy, B., and Thomas, A.S. (eds.) Methods in Molecular Biology, Vol. 22: Microscopy, Optical Spectroscopy, and Macroscopic Techniques. *Humana Press Inc.*Totowa, N.J p. 75-84.

Haris, P. I. and Chapman, D. (1994). Analysis of Polypeptide and Protein Structures Using Fourier Transform Infrared Spectroscopy. **IN** Jones, C., Mulloy, B., and Thomas, A.S. (eds.) Methods in Molecular Biology, Vol. 22: Microscopy, Optical Spectroscopy, and Macroscopic Techniques. *Humana Press Inc.*Totowa, N.J p.183-202.

Haris, P. I., Lee, D. C., and Chapman, D. (1986). A Fourier transform infrared investigation of the structural differences between ribonuclease A and ribonuclease S. *Biochim. Biophys. Acta.* **874,** 255-265.

Hasemann, C. A. and Capra, J. D. (1990). High-level production of a functional immunoglobulin heterodimer in a baculovirus expression system. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 3942-3946.

Hauser, C., Fusswinkel, H., Li, J., Oellig, C., Kunze, R., Mueller-Neumann, C. A., Heinlein, M., Starlinger, P., and Doerfler, W. (1988). Overproduction of the protein encoded by the maize transposable element Ac in insect cells by a baculovirus vector. *Mol. Gen. Genet.* **214**, 373-378.

Heenan, R. K. and King, S. M. (1993) 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., and Stanley, H. B. (1989) Colette Users Guide. Internal publication RAL-89-128, Rutherford Appleton Laboratory, Didcot, U.K.

Heinderyckx, M., Jacobs, P., and Bollen, A. (1989). Secretion of glycosylated human recombinant haptoglobin in baculovirus-infected insect cells. *Mol. Biol. Rep.* **13**, 225-232.

Herrera, R., Lebwohl, D., de Herreros, A. G., Kallen, R. G., and Rosen, O. M. (1988). Synthesis, purification, and characterization of the cytoplasmic domain of the human insulin-receptor using a baculovirus expression system. *J. Biol. Chem.* **263**, 5560-5568.

Hink, W.F (1970). Established Insect Cell Line from the Cabbage Looper, *Trichoplusia ni.* *Nature* **226,** 466-467.

Hiom, K. and West, S. C. (1995). Branch migration during homologous recombination: Assembly of a RuvAB-Holliday junction complex *in vitro.* *Cell*, **80**, 787-793.

Hjelm, R . P., Jr. (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**, 4648-4652.

Hobart, M. J., Fernie, B., and DiScipio, R. G. (1993). Structure of the human C6 gene. *Biochemistry* **32**, 6198-6205.

Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science* **3**, 522-524.

Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992). Selection of representative protein data sets. *Protein Science* **1**, 409-417.

Hochuli, E., Bannwarth, W., Döbeli, H., Gentz, R., and Stüber, D. (1988). Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *Bio/Tech.* **6**, 1321-1325

Hohenester, E., Maurer, P., and Timpl, R. (1997). Crystal strcuture of a pair of follistatin-like and EF-hand calcuim-binding domains in BM-40. *EMBO J.* **16**, 3778-3786.

Holmgren, A. and Bränden, C. I. (1989). Crystal-structure of chaperone protein PapD reveals an immunoglobulin fold. *Nature* (London) **342**, 248-251.

Hsiung, L.-M., Barclay, A. N., Brandon, M. R., Sim, E., and Porter, R. R.(1982). Purification of human C3b inactivator by monoclonal antibody affinity chromatography. *Biochem. J.* **203,** 293-298.

Huang, H. -J. S., Jones, N. H., Strominger, J. L., and Herzenberg, L. A. (1987). Molecular-Cloning of Ly-1, a Membrane Glycoprotein of Mouse Lymphocytes-T and A subset of B-cells - Molecular Homology to its Human Counterpart Leu-1/T1 (CD5). *Proc. Natl. Acad. Sci. U. S. A.* **84**, 204-208.

Ibel, K. (1976) The neutron small-angle camera D11 at the high-flux reactor, Grenoble. *J. Appl.Crystallogr.* 9, 269-309.

Iwasaki, H., Takahagi, M., Nakata, A., and Shinagawa, H. (1992). *Escherichia coli* RuvA and RuvB proteins specifically interact with Holliday junctions and promote branch migration. *Genes and Dev.* **6**, 2214-2220.

Jacrot, B. (1987). Crystallography in Molecular Biology. *Plenum Publishing Corporation.*

Jacrot, B., and Zaccai, G. (1981). Determination of molecular-weight by neutron-scattering. *Biopolymers* **20**, 2413-2426.

Jarvis, D. L., and Summers, M. D. (1989). Glycosylation and secretion of human-tissue plasminogen-activator in recombinant baculovirus-infected insect cells. *Mol. Cell. Biol.* **9**, 214-223.

Jarvis, D. J., and Finn, E. E. (1995). Biochemical analysis of the N-glycosylation pathway in baculovirus-infected lepidopteran insect cells. *Virology* **212**, 500-511.

Jawetz, E., Melnick, J. L., and Adelberg, E. A. (1987) 1987 Review of Medical Microbiology. Seventeenth Edition. *Appleton and Lange*, Norwalk, Connecticut.

Jeang, K.-T., Giam, C.-Z., Nerenberg, M., and Khoury, G. (1987). Abundant synthesis of functional human t-cell leukemia-virus type-I $p40^x$ protein in eukaryotic cells by using a baculovirus expression vector. *J. Virol.* **61**, 708-713.

Jenne, D. (1991). Homology of placental protein-11 and pea seed albumin-2 with vitronectin. *Biochem. Biophys. Res. Commun.* **176**, 1000-1006.

Jenne, D. and Stanley, K. K. (1987). Nucleotide sequence and organization of the human S-protein gene - repeating peptide motifs in the pexin family and a model for their evolution. *Biochemistry* **26**, 6735-6742.

Johnson, P. W., Attia, J., Richardson, C. D., Roder, J. C., and Dunn, R. J. (1989). Synthesis of soluble myelin-associated glycoprotein in insect and mammalian cells. *Gene* **77**, 287-296.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86-89.

Joyce, K. A., Atkinson, A. E., Bermudez, I., Beadle D. J., and King, L. A. (1993). Synthesis of functional GABA(a) receptors in stable insect-cell lines. *FEBS Letters* **335**, 61-64.

Kabsch, W. and Sander, C. (1983a). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Kabsch, W. and Sander, C. (1983b). How good are predictions of protein secondary structure? *FEBS Lett.* **155**, 179-182.

Kang, C.-Y., Bishop, D. H. L., Seo, J.-S., Matsuura, Y., and Choe, M. (1987). Secretion of particles of hepatitis-b surface-antigen from insect cells using a baculovirus vector. *J. Gen. Virol.* **68**, 2607-2613.

Kaplan, G., Freistadt, M. S., and Racaniello, V. R. (1990). Neutralization of poliovirus by cell receptors expressed in insect cells. *J. Virol.* **64**, 4697-4702.

Kataoka, M., Nishii, I., Fujisawa, T., Ueki, T., Tokunaga, F., and Goto, Y. (1995). Structural Characterization of the Molten Globule and Native States of Apomyoglobin by Solution X-ray Scattering. *J. Mol. Biol.* **249**, 215-228.

Keese, P. K. and Gibbs, A. (1992). Origins of genes - big-bang or continuous creation. *Proc. Natl. Acad. Sci. USA* **89**, 9489-9493.

Kelly, M. and Clarke, P. H. (1962) An inducible amidase produced by a strain of *Pseudomonas aeruginosa.* *J. Gen. Microbiol.* **27**, 305-316.

King, L. A., Kaur, K., Mann, S. G., Lawrie, A. M., Steven, J., and Ogden, J. E. (1991). Secretion of single-chain urokinase-type plasminogen-activator from insect cells. *Gene* **106**, 151-157.

King, L.A. and Possee, R.D. (1992). The Baculovirius Expression System: A

Laboratory Guide. *Chapman and Hall*, New York.

Kiss, I., Deak, F., Mestric, S., Delius, H., Soos, J., Dekany, K., Argraves, W. S., Sparks, K. J., and Goetinck, P. F. (1987). Structure of the chicken link protein gene - exons correlate with the protein domains. *Proc. Natl. Acad. Sci* **84**, 6399-6403.

Kitts, P. A., Ayres, M. D., and Possee, R. D. (1990). Linearization of baculovirus dna enhances the recovery of recombinant virus expression vectors. *Nucleic Acids Research* **18,** 5667-5672.

Knörle, R., Schnierle, P., Koch, A., Buchholz, N.-P., Hering, F., Seiler, H., Ackermann T., and Rutishauser, G. (1994). Tamm-Horsfall glycoprotein: Role in inhibition and promotion of renal calcium oxalate stone formation studied with Fourier transform infrared spectroscopy. *Clin. Chem.* **40,** 1739-1743.

Koch, C. A., Anderson, D., Moran, M. F., Ellis, C., Pawson, T. (1991). SH2 and SH3 domains - elements that control interactions of cytoplasmic signaling proteins. *Science* **252**, 668-674.

Koch, C. and Høiby, N. (1993). Pathogenesis of Cystic Fibrosis. *Lancet* **341**, 1065-1069.

Konno T., Kataoka, M., Kamatari, Y., Kanaori, K., Nosaka, A., and Akasaka, K. (1995). Solution X-ray Scattering Analysis of Cold- Heat-, and Urea-denatured States in a protein, *Streptomyces* Subtilisin Inhibitor. *J. Mol. Biol.* **251,** 95-103.

Koonin, E. V., Bork, P., and Sander, C. (1994). A novel RNA-binding motif in omnipotent suppressors of translation termination, ribosomal-proteins and a ribosome modification enzyme. *EMBO J.* **13**, 493-503.

Kotula, L. and Curtis, P. J. (1991). Evaluation of foreign gene optimization in yeast: expression of a mouse Ig kappa chain. *Bio/tech.* **9**, 1386.

Kotula, L., Laury-Kleintop, L. D., Showe, L., Sahr, K., Linnenbach A. J., (1991). Evaluation of foreign gene codon optimization in yeast - expression of a mouse Ig kappa-chain. *Genomics* 9, 131-140.

Kratky, O. (1963). X-Ray Small Angle Scattering with Substances of Biological Interest in Diluted Solutions. *Progr. Biophys. Chem.* **13**, 105-173.

Kreis, M., Forde, B. G., Rahman, S., Miflin, B. J., and Shewry, P. R. (1985). Molecular Evolution of the Seed Storage Proteins of Barley, Rye, and Wheat. *J. Mol. Biol.* **183**, 499-502.

Krishna, S., Blacklaws, B. A., Overton, H. A., Bishop, D. H. L., and Nash, A. A. (1989). Expression of glycoprotein D of herpes-simplex virus type-1 in a recombinant baculovirus - protective responses and T-cell recognition of the recombinant-infected cell-extracts. *J. Gen. Virol.* **70**, 1805-1814.

Kunnath-Muglia, L. M., Chang, G. H., Sim, R. B., Day, A. J., and Ezekowitz, R. A. (1993). Characterization of *Xenopus laevis* complement factor I structure - conservation of modular structure except for an unusual insert not present in human factor I. *Mol. Immun.* **30**, 1249-1256.

Kuroda, K., Hauser, C., Rott, K., Klenk, H.-D., and Doerfler, W. (1986). Expression of the influenza-virus hemagglutinin in insect cells by a baculovirus vector. *EMBO J.*, **5**, 1359-1365.

Kuroda, K., Groner, A., Frese, K., Drenckhahn, D., Hauser, C., Rott, R., Doerfler, W., and Klenk, H.-D. (1989). Synthesis of biologically-active influenza-virus hemagglutinin in insect larvae. *J. Virol.* **63**, 1677-1685.

Kuroda, K., Geyer, H., Geyer, R., Doerfler, W., and Klenk, H.-D. (1990). The oligosaccharides of influenza-virus hemagglutinin expressed in insect cells by a baculovirus vector. *Virology* **174**, 418-429.

Kuryatov, A., Laube, B., Betz, H., and Kuhse, J. (1994). Mutational Analysis of the Glycine-Binding Site of the NMDA Receptor: Structural Similarity with Bacterial Amino-Acid Binding Proteins. *Neuron* **12,** 1291-1300.

Kyte, J. (1995). Structure in Protein Chemistry. *Garland Publishing Inc.* London.

Lambris, J. D., Lao, Z., Oglesby, T. J., Atkinson, J. P., Hack, C. E., and Becherer, J. D. (1996). Dissection of CR1, factor H, membrane cofactor protein and factor B binding and functional sites in the third complement component. *J. Immun.* **156,** 4821-4832.

Landschulz, W. H., Johnson, P. F., and McKnight, S. L. (1988). The leucine zipper - a hypothetical structure common to a new class of DNA-binding proteins. *Science* **240**, 1759-1764.

Lane, T. F. and Sage, E. H. (1994). The biology of SPARC, a protein that modulates cell-matrix interactions. *FASEB J.* **8**, 163-173.

Lanford, R. E. (1988). Expression of simian virus-40 t-antigen in insect cells using a baculovirus expression vector. *Virology* **167**, 72-81.

Lanford, R. E., Luckow, V., Kennedy, R. C., Dreesman, G. R., Notvall, L., and Summers, M. D. (1989). Expression and characterization of hepatitis B virus surface-antigen polypeptides in insect cells with a baculovirus expression system. *J. Virol.* **63**, 1549-1557.

Lanford, R. E. and Notvall, L. (1990). Expression of hepatitis B virus core and precore antigens in insect cells and characterization of a core-associated kinase-activity.

Virology **176**, 222-233.

Lao Z., Wang Y., Mavroidis M., Kostavasili I., and Lambris J.D. (1994). Overexpression, purification and characterization of third component of complement. *J. Immun. Meth.* **176,** 127-139.

Laskowski, R. A., McArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283-291.

Law, S. K. A. and Reid, K. B. M. (1995). Complement. 2nd edition. *IRL Press*, Oxford.

Leach, A. R. (1996). Molecular Modelling. Principles and Applications. *Longman*, Harlow.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.

Lee, D. C., Haris, P. I., Chapman, D., and Mitchell, R. C. (1990). Determination of protein secondary structure using factor analysis of infrared spectra. *Biochemistry* **29,** 9185-9193.

Lee, J-O., Rieu, P., Arnaout, M. A., and Liddington, R. (1995). Crystal Structure of the A Domain from the α Subunit of Integrin CR3 (CD11b/CD18). *Cell* **80,** 631-638.

Lilley, D. M. A. and Clegg, R. M. A. (1993). The structure of the four-way junction in DNA. *Ann. Rev. Biophy. Biomolec. Struct.* **22**, 299-328.

Lindner, P., May, R. P., and Timmins, P. A. (1992). Upgrading of the SANS Instrument-D11 at the ILL. *Physica B.* **180**, 967-972.

Little, E, Bork, P., and Doolittle, R. F. (1994). Tracing the spread of fibronectin type-III domains in bacterialglycohydrolases. *J. Mol. Evol.* **39**, 631-643.

Ljungquist, C., Brietholtz, A., Brink-Nilsson, H., Moks, T., Uhlén, M., and Nilsson, B. (1989). Immobilization and affinity purification of recombinant proteins using histidine peptide fusions. *Eur. J. Biochem* **186**, 563-569.

Luecke, H., and Quiocho, F. A. (1990). High specificity of a phosphate transport protein determined by hydrogen bonds. *Nature* **347**, 402-406.

Lupas, A., Engelhardt, H., Peters, J., Santarius, U, Volker, S., and Baumeister, W. (1994). Domain structure of the *Acetogenium kivui* surface layer revealed by electron crystallography and sequence analysis. *J. Bacteriol.* **176**, 1124-1233.

MacArthur, M. W., Driscoll, P. C., and Thornton, J. M. (1994). NMR and crystallography - complementary approaches to structure determination. *Trends Biotechnol.* **12**, 149-158.

Maeda, S. (1989). Increased insecticidal effect by a recombinant baculovirus carrying a synthetic diuretic hormone gene. *Biochem. Biophys. Res. Comm.* **165**, 1177-1183.

Mahdi, A. A., Sharples, G. J., Mandal, T. N., and Lloyd, R. G. (1996). Holliday junction resolvases encoded by homologous rusA genes in *Escherichia coli* K-12 and phage 82. *J. Mol. Biol.* **257**, 561-573.

Mancuso, D.J., Tuley, E.A., Westfield, L.A., Worrall, N.K., Shelton-Inloes, B.B. (1989). Structure of the gene for human von Willebrand factor. *J. Biol. Chem.* **264**, 19514-19527.

Mandal, T. N., Mahdi, A. A., Sharples, G. J., and Lloyd, R. G. (1993). Resolution of Holliday intermediates in recombination and DNA repair: indirect suppression of ruvA, ruvB, and ruvC mutations. *J. Bacteriol.* **175**, 4325-4334.

Martin, B. M., Tsuji, S., Lamarea, M. E., Maysak, K., Eliason, W., and Ginna, E. I. (1988). Glycosylation and processing of high-levels of active human glucocerebrosidase in invertebrate cells using a baculovirus expression vector. *DNA* 7, 99-106.

Mastrangelo, I. A., Hough, P. V. C., Wall, J. S., Dobson, M., Dean, F. B., and Hurwitz, J. (1989). ATP-dependent assembly of double hexamers of SV40 T antigen at the viral origin of DNA replication. *Nature*, **338**, 658-662.

Matsuura, Y., Miyamoto, M., Sato, T., Morita, C., and Yasui, K. (1989). Characterization of Japanese encephalitis-virus envelope protein expressed by recombinant baculoviruses. *Virology* **173**, 674-682.

Mayans, M. O., Coadwell, W. J., Beale, D., Symons, D. B. A., and Perkins, S. J. (1995) The arrangementof the Fab and Fc fragments in the overall structures of bovine IgG1 and IgG2 in solution is similar by pulsed neutron scattering. *Biochem. J.* **311**, 283-291.

McBride, A. A., Bolen, J. B., and Howley, P. M. (1989). Phosphorylation sites of the E2 transcriptional regulatory proteins of bovine papillomavirus type-1. *J. Virol.* **63**, 5076-5085.

McRorie, D. K., and Voelker, P. J. (1993) Self-associating systems in the analytical ultracentrifuge. Beckmann Instruments Inc.

Meinke, A., Gilkes, N. R., Kilburn, D. G., Miller, R. C., and Warren, R. A. J. (1991). Multiple domains in endoglucanase-B (CENB) from *Cellulomonas fimi* - functions and relatedness to domains in other polypeptides. *J. Bacteriol.* **173**, 7126-7135.

Mendelson, R.A., Schneider, D.K., and Stone, D.B. (1996). Conformations of Myosin Subfragment 1 ATPase Intermediates from Neutron and X-ray Scattering. *J. Mol. Biol.* **256**, 1-7.

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.

Minta J. O., Wong M. J., Kozak C. A., Kunnath-Muglia L. M., and Goldberger G. (1996). cDNA cloning, sequencing and chromosomal assignment of the gene for mouse complement factor I (C3b/C4b inactivator): identification of a species specific divergent segment in factor I. *Mol. Immunol.* **33**, 101-112.

Mitchell, A. H. and West, S. C. (1994). Hexameric rings of *Escherichia coli* RuvB protein: Cooperative assembly, processivity and ATPase activity. *J. Mol. Biol.* **243**, 208-215.

Miyamoto, C., Smith, G. E., Farrell-Towt, J., Chizzonite, R., Summers, M. D. and Ju, G. (1985). Production of human c-*myc* protein in insect cells infected with a baculovirus expression vector. *Mol. Cell. Biol.* **5**, 2860-2865.

Moore, P.B. (1985). Applications of Neutron Scattering to Biology. *Phys. Today* **38**, 63-72.

Mowbray, S. L., and Petsko, G. A. (1983). The X-ray Structure of the Periplasmic Binding Protein from *Salmonella typhimurium* at 3.0-Å Resolution. *J. Biol. Chem.* **258**, 7991-7997.

Muller, B., Tsaneva, I. R., and West, S. C. (1993). Branch migration of Holliday junctions promoted by the *Escherichia coli* RuvA and RuvB proteins: II. Interaction of RuvB with DNA. *J. Biol. Chem.* **268**, 17185-17189.

Mullis, K., and Faloona, F. (1987). Specific synthesis of DNA *in vitro* via a polymerase catalysed chain reaction. *Meth. Enzymol.* **155**, 335-350.

Murphy, C. I., Weiner, B., Bikel, I., Piwnica-Worms, H., Bradley, M. K., and Livingston, D. M. (1988). Purification and functional properties of simian virus-40 large and small-t antigens overproduced in insect cells. *J. Virol.* **62**, 2951-2959.

Murzin, A. G., Lesk, A. M., and Chothia, C. (1992). Beta trefoil fold patterns of structure and sequence in the kunitz inhibitors interleukins-1-beta and 1-alpha and fibroblast growth factors. *J. Mol. Biol.* **62**, 531-543.

Musacchio, A., Gibson, T., Rice, P., Thompson, J., and Saraste, M. (1993). The PH domain - a common piece in the structural patchwork of signaling proteins. *Trends Biochem. Sci.* **18**, 343-348.

Nave, C., Helliwell, J.R., Moore, P.R., Thompson, A.W., Worgan, J.S., Greenall, R.J., Miller, A., Burley, S.K., Bradshaw, J., Pigram, W.J., Fuller, W., Siddons, D.P., Deutsch, M., and Tregear, R.T. (1985). Facilities for solution scattering and fibre diffraction at the Daresbury SRS. *J. Appl. Crystallogr.* **18,** 396-403.

Neer, E. J., Schmidt, C. J., Manbudripad, R., and Smith, T. F. (1994). The ancient regulatory-protein family of WD repeat proteins. *Nature* **371**, 297-300.

Nettsheim, D.G., Edalji, R.P., Mollison, K.W., Greer, J. and Zuiderweg, E.R.P. (1988). Secondary Structure of Complement Component C3a Anaphylatoxin in Solution as Determined by NMR-Spectroscopy - Differences between Crystal and Solution Conformations. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5036-5040.

Newcomer, M. E., Lewis, B. A., and Quiocho, F. A. (1981) The radius of gyration of L-Arabinose Binding Protein decreases upon binding of ligand. *J. Biol. Chem.* **256**, 13218-13222.

Nishino, T., Ariyoshi, M., Iwasaki, H., Shinagawa, H., and Morikawa, K. (1998). Functional analysis of the domain structure in the Holliday junction binding protein RuvA. *Structure*, **6**, 11-21.

Norman, D.G., Barlow, P.N., Baron, M., Day, A.J., Sim, R.B., and Campbell, I.D. (1991). 3-Dimensional Structure of a Complement Control Module in Solution. *J. Mol. Biol.* **219,** 717-725.

Noteborn, M. H. M., de Boer, G. F., Kant, A., Koch, G., Bos, J. L., Zantema, A., and van der Eb, A. J. (1990). Expression of avian leukemia-virus env-gp85 in *Spodoptera frugiperda* cells by use of a baculovirus expression vector. *J. Gen. Vir.* **71**, 2641-2648

Ny, T., Elgh, F., and Lund, B. (1984). The structure of the human tissue-type plasminogen-activator gene - correlation of intron and exon structures to functional and structural domains. *Proc. Natl. Acad. Sci. USA* **81**, 5355-5359.

Nygren, P-Å., Ståhl, S., and Uhlén, M. (1994). Engineering proteins to facilitate bioprocessing. *Trends Biotechnol.* **12**, 184-188.

Nyunoya, H., Akagi, T., Ogura, T., Maeda, S., and Shimotohno, K. (1988). Evidence for phosphorylation of trans-activator $p40^x$ of human T-cell leukemia-virus type-1 produced in insect cells with a baculovirus expression vector. *Virology* **167**, 538-544.

Nyyssonen, E., Penttila, M., Harkki, A., Saloheimo, A., Knowles, J. K. C., and Keranen, S. (1993). Efficient production of antibody fragments by the filamentous fungus

Trichoderma reesei. *Bio'tech.* **11**, 591.

Oh, B. -H., Pandit, J., Kang, C. -H., Nikaido, K., Gokcen, S., Ames, G. F. -L., and Kim, S. -H. (1993). Three-dimensional Structures of the Periplasmic Lysine/Argine/Ornithine-binding Protein with and without a Ligand. *J. Biol. Chem.* **268**, 11348-11355.

Oh, B. -H., Kang, C. -H., De Bondt, H., Kim, S. -H., Nikaido, K., Joshi, A. K., and Ames, G. F. -L. (1994). The Bacterial Periplasmic Histidine-binding Protein. *J. Biol. Chem.* **269**, 4135-4143.

O'Hara, P. J., Sheppard, P. O., Thøgersen, H., Venezia, D., Haldeman, B. A., McGrane, V., Houamed, K. M., Thomsen, C., Gilbert, T. L., and Mulvihill, E. R. (1993). The Ligand-Binding Domain in Metabotropic Glutamate Receptors Is related to Bacterial Periplasmic Binding Proteins. *Neuron* **11**, 41-52.

Olah, G.A, Gray, D.M., Gray, C.W., Kergil, D.L., Sosnick, T.R., Mark, B.L., Vaughan, M.R., and Trewhella, J. (1995). Structures of fd Gene 5 Protein-Nucleic Acid Complexes: A Combined Solution Scattering and Electron Microscopy Study. *J. Mol. Biol.* **249,** 576-594.

Ollo, R. and Maniatis, T. (1987). *Drosophila* krüppel gene-product produced in a baculovirus expression system is a nuclear phosphoprotein that binds to DNA. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 5700-5704.

Oram, M., Keeley, A., and Tsaneva, I. R. (1988). Holliday junction resolvase in *Schizosaccharomyces pombe* has identical endonuclease activity to the CCE1 homologue YDC2. *Nucl Acid Res.* **26**, 594-601.

O'Reilly, D. R. and Miller, L. K. (1988). Expression and complex-formation of simian virus-40 large T-antigen and mouse p53 in insect cells. *J. Virol.* **62**, 3109-3119.

O'Reilly, D., Miller, L.K., and Luckow, V.A. (1992). Baculovirius Expression Vectors, A Laboratory Manual. *W.H. Freeman and Company*, New York, NY.

Ouzounis, C., Bork, P., and Sander, C. (1994). The modular structure of NifU proteins. *Trends Biochem. Sci.* **19**, 199-200.

Overton, H. A., Fujii, Y., Price, I. R., and Jones, I. M. (1989). The protease and gag gene-products of the human immunodeficiency virus - authentic cleavage and post-translational modification in an insect cell expression system. *Virology* **170**, 107-116.

Parsons, C. A. and West, S. C. (1993). Formation of a RuvAB-Holliday junction complex *in vitro*. *J. Mol. Biol.* **232**, 397-405.

Parsons, C. A., Tsaneva, I., Lloyd, R. G., and West, S. C. (1992). Interaction of

Escherichia coli RuvA and RuvB proteins with synthetic Holliday juctions. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 5452-5456.

Parsons, C. A., Stasiak, A., Bennet, R. J., and West, S. C. (1995). Structure of a multisubunit complex that promotes DNA branch migration. *Nature*, **374**, 375-378.

Patel, G. and Stabel, S. (1989). Expression of a functional protein kinase-C-gamma using a baculovirus vector - purification and characterization of a single protein kinase-C iso-enzyme. *Cellular Signalling* **1**, 227.

Patel, G., Greenfield, C., Stabel, S., Waterfield, M. D., Parker, P. J., and Jones, N. C. (1988). **IN** Gluzman, Y., and Hughes, S. H. (eds.). Current Communication in Molecular Biology: Viral Vectors. *Cold Spring Harbour*, New York, p. 98-103.

Patel, R. S., Odermatt, E., Schwarzbauer, J. E., and Hynes, R. O. (1987). Organization of the fibronectin gene provides evidence for exon shuffling during evolution. *EMBO J.* **6**, 2565-2572.

Patthy, L. (1987). Intron-dependent evolution - preferred types of exons and introns. *FEBS Lett.* **214**, p. 1-7.

Paul, J. I., Tavare, J., Denton, R. M., and Steiner, D. F. (1990). Baculovirus-directed expression of the human insulin-receptor and an insulin-binding ectodomain. *J. Biol. Chem.* **265**, 13074-13083.

Pearl, L. H., O'Hara, B. P., Drew, R. E., and Wilson, S. A. (1994). Crystal structure of AmiC: the controller of transcription antitermination in the amidase operon of *Pseudomonas aeruginosa.* *EMBO J.* **13**, 5810-5817.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444-2448.

Peifer, M., McCrea, P. D., Green, K. J., Wieschaus, E., and Gumbiner, B. M. (1992). The vertebrate adhesive junction proteins beta-catenin and plakoglobin and the drosophila segment polarity gene *armadillo* form a multigene family with similar properties. *J. Cell Biol.* **118**, 681-691.

Peifer, M. and Weischaus, E. (1990). The segment polarity gene *armadillo* encodes a functionally modular protein that is the drosophila homolog of human plakoglobin. *Cell* **63**, 1167-1178.

Pendergast, A. M., Clark, R., Kawasaki, E. S., McCormick, F. P., and Witte, O. N. (1989). Baculovirus expression of functional P^{210} BCR-ABL oncogene product. *Oncogene* **4**, 759-766.

Perkins, S. J. (1986). Protein volumes and hydration effects - the calculations 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. (1988a). Structural studies of proteins by high-flux X-ray and neutron solution scattering. *Biochem. J.* **254**, 313-327.

Perkins, S. J. (1988b). X-ray and neutron solution scattering. **IN** Neuberger, A. and Van Deenen, L. L. M., (eds.) New Comprehensive Biochemistry. Volume 11B Part II. *Elsevier*, Amsterdam, p. 143-264.

Perkins, S. J. (1989). Hydrodynamic properties of macromolecular assemblies. **IN** Harding, S.E., and Rowe, A.J. (eds.) Dynamic Properties of Biomolecular Assemblies. *Royal Society of Chemistry*, Cambridge, U.K p. 226-245.

Perkins, S. J. (1994). High-Flux X-Ray and Neutron Solution Scattering. **IN** Jones, C., Mulloy, B., and Thomas, A.S. (eds.) Methods in Molecular Biology, Vol. 22: Microscopy, Optical Spectroscopy, and Macroscopic Techniques. *Humana Press Inc.*Totowa, N.J p. 39-60.

Perkins, S. J., Ashton, A. W., Boehm, M. K., and Chamberlain, D. (1998). Molecular structures from low angle X-ray and neutron scattering studies. *Int. J. Biolog. Macromol.* **22**, 1-16.

Perkins, S. J., Nealis, A. S., Sutton, B. J., and Feinstein, A. Solution Structure of Human and Mouse Immunoglobulin-M by Synchrotron X-ray-scattering and Molecular Graphics Modeling - a Possible Mechanism for Complement Activation. *J. Mol. Biol.* 1991, **221**, 1345-1366

Perkins, S. J. and Smith, K. F. (1993). Identity of the putative serine-proteinase fold in proteins of the complement system with nine relevant crystal structures. *Biochem. J.*, **295**, 109-114.

Perkins, S. J., Smith, K. F., and Sim, R. B. (1993a). Molecular modelling of the domain structure of factor I of human complement by X-ray and neutron solution scattering. *Biochem. J.*, **295**, 101-108.

Perkins, S. J., Smith, K. F., Kilpatrick, J. M., Volanakis, J. E., and Sim, R. B. (1993b) 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. and 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.

Pflugrath, J. W. and Quiocho, F. A. (1988). The 2 Å Resolution Structure of the Sulfate-binding Protein Involved in Active Transport in *Salmonella typhimurium*. *J. Mol. Biol.* **200**, 163-180.

Possee, R. D. (1986). Cell-surface expression of influenza-virus hemagglutinin in insect cells using a baculovirus vector. *Virus Research* **5**, 43-59.

Préhaud, C., Harris, R. D., Fulop, V., Koh, C.-L., Wong, J., Flamand, A., and Bishop, D. H. L. (1989). Expression, characterization, and purification of a phosphorylated rabies nucleoprotein synthesized in insect cells by baculovirus vectors. *Virology* **178**, 486-497.

Préhaud, C., Takehara, K., Flamand, A., and Bishop, D. H. L. (1989). Immunogenic and protective properties of rabies virus glycoprotein expressed by baculovirus vectors. *Virology* **173**, 390-399.

Provencher, S. W. (1982). A constrained regularization method for inverting data represented by linear algebraic or integral-equations. CONTIN - a general-purpose constrained regularization program for inverting noisy linear algebraic and integral-equations. *Comput. Phys. Commun.* **27**, 213-227 and 229-242.

Pschorr, J., Bieseler, B., and Fritz, H. -J. (1994). Production of the immunoglobulin variable domain REI$_v$ via a fusion protein synthesized by *Staphylococcus carnosus*. *Biol. Chem. Hoppe-Seyler* **375**, 271.

Quiocho, F. A., Wilson, D. K., and Vyas, N. K. (1989). Substrate specificity and affinity of a protein modulated by bound water molecules. *Nature* **340**, 404-407.

Rafferty, J. B., Sedelnikova, S. E., Hargreaves, D., Artymiuk, P. J., Baker, P. J., Sharples, G. J., Mahdi, A. A., Lloyd, R. G., and Rice, D. W. (1996). Crystal structure of DNA recombination protein RuvA and a model for its binding to the Holliday junction. *Science*, **274**, 415-420.

Ralston, G. (1993). Introduction to Analytical Ultracentrifugation. Beckman Instruments Inc.

Ramalingam, R., Blume, J. E., and Ennis, H. L. (1992). The *Dictyostelium discoideum* spore germination-specific cellulase is organized into functional domains. *J. Bacteriol.* **174**, 7834-7838.

Ray, M. V. L., van Duyne, P., Bertelsen, A. H., Jackson-Matthews, D. E., Sturmer, A. M., Merkler, D. J., Consalvo, A. P., Young, S. D., Gilligan, J. P., and Shields, P. P. (1993). Production of recombinant salmon-calcitonin by *in-vitro* amidation of an *Escherichia coli* produced precursor peptide. *Bio/Tech.* **11**, 64-70

Ray, R., Galinski, M. S., and Compans, R. W. (1989). Expression of the fusion glycoprotein of human para-influenza type-3 virus in insect cells by a recombinant baculovirus and analysis of its immunogenic property. *Virius Research* **12**, 169-180.

Resnick, D., Pearson, A., and Krieger, M. (1994). The SRCR superfamily: a family reminiscent of the Ig superfamily. *Trends Biochem. Sci.* **19**, 5-8.

Resnick, D., Chatterton, J. E., Schwatz, K., Slayter, H., and Krieger M. (1996). Structures of Class A macrophage scavenger receptors. *J. Biol. Chem.***271**, 26924-26930.

Richarme, G. (1982). Associative Properties of the *Escherichia coli* Galactose Binding Protein and Maltose Binding Protein. *Biochem. Biophys. Res. Commun.* **105**, 476-481.

Robson, B. and Susuki, E. (1976). Conformational properties of amino acids residues in globuar proteins. *J. Mol. Biol.* **107**, 327-356.

Robson, B. and 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.

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. and Sander, C. (1993a). Improved prediction of protein secondary structure by the use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. USA.* **90**, 7558-7562.

Rost, B. and Sander, C. (1993b). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599.

Rost, B. and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216-226.

Rost, B. and Sander, C. (1996). Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 113-136.

Rusche, J. R., Lynn, D. L., Robert-Guroff, M., Langlois, A. J., Lyerly, H. K., Carson, H., Krohn, K., Ranki, A., Gallo, R. C., Bolognesi, D. P., Putney, S. D., and Matthews, T. J. (1987). Humoral immune-response to the entire human immunodeficiency virus envelope glycoprotein made in insect cells. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6924-6928.

Sack, J. S., Saper, M. A., and Quiocho, F. A. (1989a). Periplasmic binding protein structure and function: Refined X-ray structures of the Leucine/Isoleucine/Valine-binding protein and its complex with leucine. *J. Mol.*

Biol. **206**, 171-191.

Sack, J. S., Trakhanov, S. D., Tsigannik, I. H., and Quiocho, F. A. (1989b). Structure of the L-Leucine-binding Protein Refined at 2.4 Å Resolution and Comparison with the Leu/Ile/Val-binding Protein Structure. *J. Mol. Biol.* **206**, 193-207.

Saier, M. H. and Reizer, J. (1990). Domain shuffling during evolution of the proteins of the bacterial phosphotransferase system. *Res. Micro.* **141**, 1033-1038.

Šali, A. and Blundell, T. L. (1990). The definition of topological equivalence in homologous and analogous structures: A procedure involving a comparison of local properties and relationships. *J. Mol. Biol.* **212**, 403-428.

Sambrook, E. F., Fritsch, E. F., and Maniatis, T. (1989). Molecular cloning: A laboratory manual. Second edition. *Cold Spring Harbor Laboratory Press*, New York.

Schatz, P. J. (1993). Use of peptide libraries to map the substrate-specificity of a peptide-modifying enzyme - a 13 residue consensus peptide specifies biotinylation in *Escherichia coli*. *Bio Tech.* **11**, 1138-1143.

Schmidt, T. G. M. and Skerra, A. (1993). The random peptide library-assisted engineering of a c-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. *Protein. Eng.* **6**, 109-122.

Seidel, H. M., Pompliano, D. L., and Knowles, J. R. (1992). Exons as microgenes. *Science* **257**, 1489-1490.

Semenyuk, A. V. and Svergun, D. I. (1991). GNOM- A Program Package for Small-Angle Scattering Data-Processing. *J. Appl. Crystallogr.* **24**, 537-540.

Sharff, A. J., Rodseth, L. E., Spurlino, J. C., and Quiocho, F. A. (1992). Crystallographic Evidence of a Large Ligand-Induced Hinge-Twist Motion between the Two Domains of the Maltodextrin Binding Protein Involved in Active Transport and Chemotaxis. *Biochemistry* **31**, 10657-10663.

Sharma, A. K. and Pangburn, M. K. (1994). Biologically active recombinant human complement factor H: synthesis and secretion by the baculovirus system. *Gene* **143,** 301-302.

Sharma, A. K. and Pangburn, M. K. (1996). Identification of three physically and functionally distinct binding sites for C3b in human complement factor H by deletion mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 10996-11001.

Sharma, A. K. and Pangburn, M. K. (1997). Localization by site-directed mutagenesis of the site in human complement factor H that binds to *Streptococcus pyogenes* M protein. *Infect. Immun.* **65**, 484-487.

Sharples, G. J., Chan, S. N., Mahdi, A. A., Whitby, M. C., and Lloyd, R. G. (1994). Processing of intermediates in recombination and DNA repair: identification of a new endonuclease that specifically cleaves Holliday junctions. *EMBO J.* **13**, 6133-42.

Shiba, T., Iwasaki, H., Nakata, A., and Shinagawa, H. (1991). SOS-inducible DNA repair proteins, RuvA and RuvB, of *Escherichia coli*: Functional interactions between RuvA and RuvB for ATP hydrolysis and renaturation of the cruciform structure in supercoiled DNA. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8445-8449.

Shiba, T., Iwasaki, H., Nakata, A., and Shinagawa, H. (1993). *Escherichia coli* RuvA and RuvB proteins involved in recombination repair: Physical properties and interactions with DNA. *Mol. Gen. Genet.* **237**, 395-399.

Shinagawa, H. and Iwasaki, H. (1996). Processing the Holliday juction in homologous recombination. *Trends Biochem. Sci.* **21**, 107-111.

Sikorski, R. S., Boguski, M. S., Goebl, M., and Hieter, P. (1990). A repeating amino-acid motif in cdc23 defines a family of proteins and a new relationship among genes required for mitosis and RNA-synthesis. *Cell* **60**, 307-317.

Sim, R. B., Day, A. J., Moffatt, B. E., and Fontaine, M. (1993). Complement factor I and cofactors in control of complement system convertase enzymes. *Meth. Enzymol.* **223**, 13-35.

Sim, E. and Sim, R. B. (1983). Enzymic assay of C3b receptor on intact cells and solubilized cells. *Biochem. J.* **210**, 567-576.

Singleton, P. and Sainsbury, D. (1987). Dictionary of Microbiology and Molecular Biology. Second Edition. *Wiley*, Chichester.

Smas, C. M., Green, D., and Sul, H. S. (1994). Structural characterization and alternate splicing of the gene encoding the preadipocyte EGF-like protein PREF-1. *Biochemistry* **33**, 9257-9265.

Smith, C. G. M., Tew, D. G., and Wolf, C. R. (1994). Dissection of NADPH-cytochrome p450 oxidoreductase into distinct functional domains. *Proc. Natl. Acad. Sci. USA* **87**, 8710-8714.

Smith, D. B. and Johnson, K. S. (1988). Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* **67**, 31-40

Smith, G. E., Fraser, M. J., and Summers, M. D. (1983a). Molecular engineering of the *Autographa californica* nuclear polyhedrosis virius genome: deletion mutants within the polyhedrin gene. *J. Virol.* **46**, 584-593.

Smith, G. E., Summers, M. D., and Fraser, M. J. (1983b). Production of human beta interferon in insect cells infected with a baculovirus expression vector. *Mol. Cell. Biol*, **3**, 2156-65.

Smith, K. F. (1992). Structural Investigations of the Components of the Complement Cascade. *Ph.D Thesis*, University of London.

Smith, K. F., Harrison, R. A., and 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.

Soames, C. J. and Sim, R. B. (1997). Interactions between human complement components factor H, factor I and C3b. *Biochem. J.* **326**, 553-561.

Spraggon, G., Phillips, C., Nowak, U. K., Ponting, C. P., Saunders, D., Dobson, C. M., Stuart, D. I., and Jones, E. Y. (1995). The crystal structure of the catalytic domain of human urokinase-type plasminogen activator. *Structure*, **3** 681-691.

Spurlino, J. C., Lu, G.- Y., and Quicho, F. A. (1991). The 2.3-Å Resolution Structure of the Maltose- or Maltodextrin-binding Protein, A Primary Receptor of Bacterial Active Transport and Chemotaxis. *J. Biol. Chem.* **266**, 5202-5219.

Stanbury, P. F., Whitaker, A., and Hall, S. J. (1995). Principles of Fermentation Technology. Second Edition. *Pergamon*.

Stanier, R.Y., Palleroni, N.J., and Doudoroff, M. (1966). The anaerobic pseudomonads: a taxonomic study. *J. Gen. Microbiol.* **43**, 159-271.

Stasiak, A., Tsaneva, I. R., West, S. C., Benson, C. J., Yu, X., and Egelman, E. H. (1994). The *Escherichia coli* RuvB branch migration protein forms double hexameric rings around DNA. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 7618-7622.

Stern-Bach, Y., Bettler, B., Hartley, M., Sheppard, P. O., O'Hara, P. J., and Heinemann, S. F. (1994). Agonist Selectivity of Glutamate Receptors is Specified by Two Domains Structurally Related to Bacterial Amino Acid-Binding Proteins. *Neuron* **13**, 1345-1357.

Stewart, C. N. Jr., Adang, M. J., All, J. N., Boerma, R., Cardineau, G., Tucker, D., and Parrott, W. A. (1996). Genetic Transformation, Recovery, and Characterization of Fertile Soybean Transgenic for a Synthetic *Bacillus thuringensis cryIAc* Gene. *Plant Physiol.* **112**, 121-129.

Stirling, D. A., Petrie, A., Pulford, D. J., Paterson, D. T. W., and Stark, M. J. R. (1992). Protein A calmodulin fusions - a novel approach for investigating calmodulin function in yeast. *Mol. Microbiol.* **6**, 703-713.

Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. Jr, and Doolittle, W. F. (1994).

Introns and the origin of protein-coding genes - Reply. *Science* **265**, 202-207.

Stoppa-Lyonnet, D., Carter, P. E., Meo, T., and Tosi, M. (1990). Clusters of intragenic Alu repeats predispose the human C1-inhibitor locus to deleterious rearrangements. *Proc. Natl. Sci. USA* **87,** 1551-1555.

Sudhof, T. C., Russell, D. W., Goldstein, J. L., Brown, M. S., Sanchez-Pescador, R., and Bell, G. I. (1985). Cassette of 8 exons shared by genes for LDL receptor and EGF precursor. *Science* **228**, 893-895.

Summers, M. D. and Smith, G. E. (1978). Baculovirius structural polypeptides. *Virol.* **84,** 390-402.

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., and Feigin, L. A. (1988). Small-Angle Scattering Data Treatment by the Regularization Method. *Acta Crystallogr.* **A44**, 244-250.

Takehara, K., Ireland, D., and Bishop, D. H. L. (1988). Co-expression of the hepatitis-B surface and core antigens using baculovirus multiple expression vectors. *J. Gen. Virol.* **69**, 2763-2777.

Tam, R., and Saier M. H. Jr. (1993). Structural, Functional, and Evolutionary Relationships among Extracellular Solute-Binding Receptors of Bacteria. *Microbiol. Rev.* **57**, 320-346.

Tame, J. R. H., Murshudov, G. N., Dodson, E. J., Neil, T. K., Dodson, G. G., Higgins, C. F., and Wilkinson, A. J. (1994). The Structural Basis of Sequence-Independent Peptide Binding by OppA Protein. *Science* **264**, 1578-1581.

Taylor, K. M., Morgan, B. P., and Campbell, A. K. (1994). Altered glycosylation and selected mutation in recombinant human complement component C9: effects on haemolytic activity. *Immunology* **83**, 501-506.

Towns-Andrews, E., Berry, A., Bordas, J., Mant, G. R., Murray, P. K., Roberts, K., Sumner, I., Worgan, J. S., Lewis, R., and Gabriel, A. (1989). A time-resolved X-ray difraction station: X-ray optics, detectors and data acquistion. *Rev. Scient. Instrum.* **60**, 2346-2349.

Tsaneva, I. R., Müller, B., and West, S. C. (1992a). ATP-dependent branch migration of Holliday junctions promoted by the RuvA and RuvB proteins of *E. coli*. *Cell*, **69**, 1171-1180.

Tsaneva, I. R., Illing, G. T., Lloyd, R. G., and West, S. C. (1992b). Purification and properties of the RuvA and RuvB proteins of *Escherichia coli*. *Mol. Gen. Genet.* **235**, 1-10.

Tsaneva, I. R., Müller B., and West, S. C. (1993). RuvA and RuvB proteins of *Escherichia coli* exhibit DNA helicase activity *in vitro*. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 1315-1319.

Ueda, Y., Tsumoto, K., Watanabe, K., and Kumagai, I. (1993). Synthesis and expression of a DNA encoding the Fv domain of an anti-lysozyme monoclonal antibody, HyHEL10, in *Streptomyces lividans*. *Gene* **129**, 129.

Uhlén, M., Nilsson, B., Guss, B., Lindberg, M., Gatenbeck, S., and Philipson, L. (1983). Gene fusion vectors based on the gene for staphylococcal protein A. *Gene* **23**, 369-378.

Ullman, C. G. (1994). Expression Systems for Structural Studies of Medically Important Protein Domains in Papillomaviruses and Complement. *PhD Thesis*. University of London.

Ullman, C. G., Haris, P. I., Smith, K. F., Sim, R. B., Emery, V. C., and Perkins, S. J. (1995). β-sheet secondary structure of an LDL receptor domain from complement factor I by consensus structure predictions and spectroscopy. *FEBS Lett.* **371**, 199-203.

Ullman, C. G., Sim, R. B. and Perkins, S. J. (1996). β-sheet structures in the five domains of human complement factor I. *Molec. Immun.* **33**, 80-80 (Abstract)

Ullman, C. G., Chamberlain, D., Sim, R.B., and Perkins, S.J. (1997). Expression, secretion and characterisation of active human factor I by insect cells. *Exp. Clin. Immunogenet.* **14**, 38-38 (Abstract)

Ullman, C. G. and Perkins, S. J. (1997). The factor I and follistatin domain families: the return of a prodigal son. *Biochem. J.* **326**, 939-941.

Ullman, C. G., Chamberlain, D., Ansari, A., Emery, V. C., Haris, P. I., Sim, R. B., and Perkins, S. J. (1998). Human complement factor I: its expression and secretion by insect cells and its immunological and structural characterisation. Submitted for publication.

van der Logt, C. P. E., Reitsma, P. H., and Bertina, R. M. (1991). Intron exon organization of the human gene coding for the lipoprotein-associated coagulation inhibitor - the factor Xa dependent inhibitor of the extrinsic pathway of coagulation. *Biochemistry* **30**, 1571-1577.

van Gelder, R., Roberts, K.J., and Rossi, A. (1995). Using Synchrotron Radiation to Examine the *in-situ* Processing of Long-Chain Hydrocarbons. *The CCP13 Newsletter* **4**, 28.

van Gool, A. J., Shah, R., Mezard, C., and West, S. C. (1997). Functional interactions between the Holliday junction resolvase and the branch migration motor of

Escherichia coli. *EMBO J.* **17**, 1838-1845.

Van Wyke Coelingh, K. L., Murphy, B. R., Collins, P. L., Lebacq-Verheyden, A.-M., and Battey, J. F. (1987). Expression of biologically-active and antigenically authentic para-influenza type-3 virus hemagglutinin-neuraminidase glycoprotein by a recombinant baculovirus. *Virology* **160**, 465-472.

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.

Van de Velde, H., von Hoegen, I., Luo, W., Parnes, J. R., and Thielemans, K. (1991). The B-cell surface protein CD72/LYB-2 is the ligand for CD5. *Nature* (London) **351**, 662-664

van Holde, K. and Zlatanova, J. (1995). Chromatin Higher Order Structure: Chasing a Mirage. *J. Biol. Chem.* **270**, 8373-8376.

Varki, A. (1993). Biological role of oligosaccharides: all of the theories are correct. *Glycobiology*, **3**, 97-130.

Vaughn, J. L., Goodwin, R. H., Tompkins, G. J., and McCawley, P. (1977). The establishment of Two Cell Lines from the Insect *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *In Vitro* **13**, 213-217.

Verhoeyen, M. E. and Windust, J. H. C. (1996). Advances in antibody engineering. **IN** Hames, B. D. and Glover, D. M. (eds). Molecular Immunology. Second Edition. *IRL Press* p. 283-325.

Vik, D. P., Amiguet, P., Moffat, G. J., Fey, M., Amiguet-Barras, F., Wetsel, R. A, and Tack, B. F. (1991). Complete organization of the human-complement C3 gene and analysis of its 5' promoter region. *Biochemistry* **30**, 1080-1085.

Vik, T. A., Sweet, L. J., and Erikson, R. L. (1990). Coinfection of insect cells with recombinant baculovirus expressing $pp60^{v\text{-}src}$ results in the activation of a serine-specific protein-kinase $pp90^{rsk}$. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2685-2689.

Vogt, P. K., Bos, T. J., and Doolittle, R. F. (1987). Homology between the DNA-binding domain of the GCN4 regulatory protein of yeast and the carboxyl-terminal region of a protein coded for by the oncogene *jun*. *Proc. Natl. Acad. Sci.* **84**, 3316-3319.

Vyse, T. J., Morley, B. J., Bartók, I., Theodoridis, E. L., Davies, K. A., Webster, D. B., and Walport, M. J. (1996). The molecular basis of hereditary complement factor I deficiency. *J. Clin. Investig.*, **97**, 925-933.

Wako, H. and Blundell, T. L. (1994a). 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. and Blundell, T. L. (1994b). 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., and Kuriyan, J. (1992). Crystal-structure of the phosphotyrosine recognition domain SH2 of V-SRC complexed with tyrosine-phosphorylated peptides. *Nature* **358**, 646-653.

Wathen, M. W., Brideau, R. J., and Thomson, D. R. (1989). Immunization of cotton rats with the human respiratory syncytial virus-F glycoprotein produced using a baculovirus vector. *J. Infect. Disea.* **159**, 255-264.

Webb, N. R., Madoulet, C., Tossi, P.-F., Broussard, D. R., Sneed, L., Summers, M. D., and Nicolau, C. (1989). Cell-surface expression and purification of human cd4 produced in baculovirus-infected insect cells - (human immunodeficiency virus flow-cytometry epitope mapping glycosylation immunoaffinity purification). *Proc. Natl. Acad. Sci. U.S.A.* **86**, 7731-7735.

Weber, K. and Kabsch, W. (1994). Intron positions in actin genes seem unrelated to the secondary structure of the protein. *EMBO J.* **13**, 1280-1286.

Weis, W.I, Kahn, R., Fourme, R., Drickamer, K, and Hendrickson, W.A. (1991). Structure of the Calcium-Dependant Lectin Domain from a Rat Mannose-Binding Protein Determined by *MAD* Phasing. *Science* **254,** 1608-

Weis, W.I, Crichlow, G.V, Murthy, H.M.K, Hendrickson, W.A, and Drickamer, K. (1991). Physical Characterization and Crystallization of the Carbohydrate-Recognition Domain of a Mannose-Binding Protein from Rat. *J. Biol. Chem.***266** 20678-

West, S. C. (1996). The RuvABC proteins and Holliday junction processing in *Escherichia coli.* *J. Bacteriol.* **178**, 1237-1241.

West, S. C. (1997). Processing of recombination intermediates by the RuvABC proteins. *Annu. Rev. Genet.* **31**, 213-244.

Wetlaufer, D. B. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proc. Natl. Acad. Sci. USA* **70,** 697-701.

Wharton, K. A., Johansen, K. M., Xu, T., and Artavanis-Tsakonas, S. (1985). Nucleotide-sequence from the neurogenic locus *notch* implies a gene-product that shares homology with proteins containing EGF-like repeats. *Cell* **43**, 567-581.

Whickham, T.J., Davis, T., Granados, R.R., Schuler, M.L., and Wood, H.A. (1992). Screening of Insect Cell Lines for the Production of Recombinant Proteins and Infectious Virius in the Baculovirius Expression System. *Biotechnol. Prog.* **8**, 391-396.

Whickham,T.J. and Nemerow, G.R. (1993). Optimization of Growth Methods and Recombinant Protein Production in BTI Tn-5B1-4 Insect Cells using the Baculovirius Expression Vector. *Biotechnol. Prog.* **9**, 25-30.

Whitby, M. C., Bolt, E. L., Chan, S. N., and Lloyd, R. G. (1996). Interactions between RuvA and RuvC at Holliday junctions: Inhibition of junction cleavage and formation of a RuvC-RuvC-DNA complex. *J. Mol. Biol.* **264**, 878-890.

White, M. F., Giraud-Panis, M. -J. E., Pöhler, R. G., and Lilley, D. J. M. (1997). Recognition and manipulation of branched DNA structure by junction-resolving enzymes. *J. Mol. Biol.* **269**, 647-664.

Whitefleet-Smith, J., Rosen, E., McLinden, J., Ploplis, V. A., and Fraser, M. J. (1989). Expression of human-plasminogen cDNA in a baculovirus vector-infected insect cell system. *Arch. Biochem. Biophys.* **271**, 390-399.

Whitford, M., Stewart, S., Kuzio, J., and Faulkner, P. (1989). Identification and Sequence Analysis of a Gene Encoding gp67, an Abundant Envelope Glycoprotein of the *Autographa californica* Nuclear polyhedrosis Virius. *J. Virology* **63**, 1393-1399.

Whitney, G. S., Starling, G. C., Bowen, M. A., Modrell, B., Siadak, A. W., and Aruffo, A. (1995). The membrane-proximal scavenger receptor cysteine-rich domain of CD6 contains the activated leukocyte cell-adhesion molecule-binding site. *J. Biol. Chem.* **270**, 18187-18190.

Wignall, G. D. and Bates, F. S. (1987). Absolute Calibration of Small-Angle Neutron-Scattering Data. *J. Appl. Crystallogr.* **20**, 28-40.

Willard, H. H., Merritt, Jr, L. L., Dean, J. A., and Settle, Jr, F. A. (1981). Instrumental Methods of Analysis. Seventh Edition. *Wadsworth* California p.422-464; 614-655.

Williams, A. F. and Barclay, A. N. (1988). The immunoglobulin superfamily - domains for cell-surface recognition. *Annu. Rev. Immunol.* **6**, 381-405.

Wilson, S.A., Chayen, N.E., Hemmings, A.M., Drew, R.E., and Pearl, L.H. (1991) Crystallization of and Preliminary X-ray Data for the Negative Regulator (AmiC) of the Amidase Operon of *Pseudomonas aeruginosa.* *J. Mol. Biol.* **222**, 869-871.

Wilson, S. A. and Drew, R. E. (1991) Cloning and DNA sequence of *amiC*, a new gene regulating expression of the *Pseudomonas aeruginosa* aliphatic amidase and purification of the *amiC* product. *J. Bacteriol.* **173**, 4914-4921.

Wilson, S. A., Wachira, S. J., Drew, R. E., Jones, D., and Pearl, L. H. (1993). Antitermination of amidase expression in *Pseudomonas aeruginosa* is controlled by a novel cytoplasmic amide-binding protein. *EMBO J.* **12**, 3637-3642.

Wilson, S. A., Williams, R. J., Pearl, L. H., and Drew, R. E. (1995). Identification of two new genes in the Pseudomonas aeruginosa amidase operon, encoding an ATPase (AmiB) and a putative integral membrane protein (AmiS). *J.Biol.Chem.* **270**, 18818-18824.

Wilson, S. A., Wachira, S. J. M., Norman, R. A., Pearl, L. H., and Drew, R. E. (1996). Transcription antitermination regulation of the *Pseudomonas aeruginosa* amidase operon. *EMBO J.* **15**, 5907-5916.

Winstanley, D. and Rovesti, L. (1993). Insect viruses as biocontrol agents. **IN** Jones, D.G. (ed.). Exploitation of Microorganisms. *Chapman and Hall,* London, p. 105-136.

Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708-713.

Wong, M. J., Goldberger, G., Isenman, D. E., and Minta, J. O. (1995). Processing of human factor I in COS-1 cells co-transfected with factor I and paired basic amino acid cleaving enzyme (PACE) cDNA. *Molec. Immun.* **32**, 379-387.

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., and 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.

Wright, H. T. (1993). Introns and higher-order structure in the evolution of serpins. *J. Mol. Evol.* **36**, 136-143.

Wright, P. E. (1989). What Can 2-Dimensional NMR Tell Us About Proteins. *Trends Biochem. Sci.* **14**, 255-260.

Wu, X. -C., Ng, S. -C., Near, R., and Wong, S. L. (1993). Efficient production of a functional single-chain antidigoxin antibody via an engineered *Bacillus subtilis* expression-secretion system. *Bio/tech.* **11**, 71.

Yano, R., Oakes, M. L., Tabb, M. M., and Nomura, M. (1994). Yeast SRP1P has homology to armadillo/plakoglobin/beta-catenin and participates in apparently multiple nuclear functions including the maintenance of the nucleolar structure. *Proc. Natl. Acad. Sci. USA* **91**, 6880-6884.

Yao, N., Trakhanov, S., and Quiocho, F. A. (1994). Refined 1.89-Å Structure of the Histidine-Binding Protein Complexed with Histidine and its Relationship with many Other Active Transport/Chemosensory Proteins. *Biochemistry*, **33**, 4769-4779.

Yoden, S., Kikuchi, T., Siddell, S. G., and Taguchi, F. (1989). Expression of the peplomer glycoprotein of murine coronavirus JHM using a baculovirus vector. *Virology*, **173**, 615-623.

Yoo, D. W., Parker, M. D., Song, J., Cox, G. J., Deregt, D., and Babiuk, L. A. (1991). Structural-analysis of the conformational domains involved in neutralization of bovine coronavirus using deletion mutants of the spike glycoprotein S1-subunit expressed by recombinant baculoviruses. *Virology* **183**, 91-98.

Yu, X. and Egelman, E. H. (1997). The RecA hexamer is a structural homologue of ring helicases. *Nature Struct. Biol.* **4**, 101-104.

Yu, X., West, S. C., and Egelman, E. H. (1997). Structure and subunit composition of the RuvAB-Holliday junction complex. *J. Mol. Biol.* **266**, 217-222.

Zipper, P., Wilfing, R., Kriechbaum, M., and Durchschlag, H. (1985). A small-angle X-ray scattreing study on pre-irradiated malate synthase. The influence of formate, superoxide dismutase and catalase on the X-ray induced aggregation of the enzyme. *Z. Naturforsch.* **40c**, 364-372.

Zubay, G. (1987). Genetics. *Benjamin/Cummings*, Menlo Park, California.

Zueco, J. and Boyd, A. (1992). Protein A fusion proteins as molecular-weight markers for use in immunoblotting. *Anal. Biochem.* **207**, 348-349.

Publications.

Chamberlain, D., O'Hara, B. P., Wilson, S. A., Pearl, L. H., and Perkins, S. J. (1997). Oligomerisation of the amide sensor protein AmiC by X-ray and neutron scattering and molecular modeling. *Biochemistry* **36**, 8020-8029.

Ullman, C. G., Chamberlain, D., Sim, R. B., Ansari, A., Emery, V. C., and Perkins, S. J. (1998). Human complement factor I: its expression and secretion by insect cells and its immunological and structural characterisation. *Molecular Immunology*, in press.

Chamberlain, D., Ullman, C. G., and Perkins, S. J. (1998). Unusual structural arrangement of the domains in human complement factor I by a combination of X-ray and neutron scattering data and homology modelling. Submitted for Publication.

Perkins, S. J., Ashton, A. W., Boehm, M. K., and Chamberlain, D. (1998). Molecular structures from low angle X-ray and neutron scattering studies. *International Journal of Biological Macromolecules* **22**, 1-16.

Chamberlain, D., Keeley, A., Aslam, M., Arenas-Licea, J., Brown, T., Tsaneva, I. R., and Stephen J. 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. Submitted for Publication.

Chamberlain, D., Hinshelwood, J., Ullman, C. G., and Perkins, S. J. (1998). Molecular modelling of the compact domain structure of factor B of human complement by synchrotron X-ray and neutron solution scattering. Manuscript in Preparation.

Hinshelwood, J., Chamberlain, D., Edwards, Y. J. K., Sim, R. B., and Perkins, S. J (1998). Investigation of the Conformational Stability of the Domains and Fragments of Factor B Reveals that the Structure of the Serine Protease Domain is pH-Labile: A Study by ^{1}H Nuclear Magnetic Resonance. Manuscript in Preparation.

Abstracts and Poster Presentations.

Chamberlain, D., O'Hara, B. P., Wilson, S. A., and Perkins, S. J. (1995). Ligand Mediated Conformational Changes in the Structure of AmiC: a Preliminary Study with X-rays and Neutrons. Poster presented at Neutron Scattering 1995, Hulme Hall, Manchester University 3-4 April 1995.

Chamberlain, D., O'Hara, B. P., Wilson, S. A., and Perkins, S. J. (1995). Conformations of AmiC, the Amide Sensor Protein of *Pseudomonas aeruginosa:* A Preliminary Study with SAXS and SANS. Poster and abstract presented at CCP13/NCD Workshop, Daresbury Laboratory, Warrington 9-11 May 1995 . **Received a commendation from the judges.**

Chamberlain, D., Hinshelwood, J., Ullman, C. G., Smith, K. F., Sim, R. B., and Perkins, S. J. (1995). Compact domain structure of complement Factor B by X-ray and neutron scattering. Poster and Abstract presented at the BSI 3rd Annual Congress, Joint Meeting with the NVVI, The Brighton Centre, Brighton 6-8 December 1995.

Chamberlain, D., Hinshelwood, J., Ullman, C. G., Smith, R. B. Sim, and Perkins S. J. (1996). Molecular modelling of the compact five-domain structure of complement Factor B by X-ray and neutron scattering. *Molecular Immunology* Supplement **1** p. 80. Poster and abstract presented at the XVI International Complement Workshop, Boston USA 16-21 June 1996.

Ullman, C. G., Chamberlain, D., Sim, R. B., and Perkins, S. J. (1997). Expression, secretion and characterisation of active human factor I by insect cells. *Experimental and Clinical Immunogenetics* **14**, p. 38. Poster and Abstract presented at the 6th European Meeting on Complement in Human Disease, Innsbruck, Austria, 12-15 March 1997.

Chamberlain, D., Ullman, C. G., and Perkins, S. J. (1997). Molecular modelling of the compact domain structure of Factor I of human complement 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.

Oligomerization of the Amide Sensor Protein AmiC by X-ray and Neutron Scattering and Molecular Modeling

Dean Chamberlain, Bernard P. O'Hara, Stuart A. Wilson,
Laurence H. Pearl, and Stephen J. Perkins
Department of Biochemistry and Molecular Biology,
Royal Free Hospital School of Medicine, Rowland Hill Street,
London NW3 2PF, U.K., and Department of Biochemistry and
Molecular Biology, University of College London,
Gower Street, London WC1E 6BT, U.K.

Biochemistry®

Reprinted from
Volume 36, Number 26, Pages 8020–8029

Oligomerization of the Amide Sensor Protein AmiC by X-ray and Neutron Scattering and Molecular Modeling

Dean Chamberlain,[‡] Bernard P. O'Hara,[§] Stuart A. Wilson,[§,∥] Laurence H. Pearl,[§] and Stephen J. Perkins*,[‡]

Department of Biochemistry and Molecular Biology, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, U.K., and Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, U.K.

Received February 12, 1997; Revised Manuscript Received April 22, 1997[⊗]

ABSTRACT: AmiC is the negative regulator of the amidase operon which is involved in amide metabolism in the cytosol of *Pseudomonas aeruginosa.* Crystal structures show that AmiC contains two large domains that are very similar to the periplasmic leucine–isoleucine–valine binding protein (LivJ) of *Escherichia coli.* Synchrotron X-ray and neutron (in 100% ^2H_2O buffer) scattering data were obtained for AmiC in the presence of its substrate acetamide and its anti-inducer butyramide which binds more weakly to AmiC than acetamide. Guinier analyses to obtain radius of gyration R_G and molecular weight M_r values showed that AmiC formed trimers whose formation was favored in the presence of acetamide and which exhibited concentration-dependent properties at concentrations between 0.4 and 2 mg/mL. Above 2 mg/mL, where trimers predominated, the R_G data were identical within 0.05 nm for AmiC–acetamide and AmiC–butyramide with mean X-ray and neutron R_G values of 3.35 and 3.28 nm, respectively. Scattering curve fits constrained by the crystal structure of AmiC–acetamide were evaluated in order to describe a model for trimeric AmiC. A translational search of parallel alignments of three monomers to form a symmetric AmiC homotrimer gave a good X-ray curve fit. Combinations of calculated curves for monomeric, dimeric, trimeric, and tetrameric AmiC as seen in the crystal structure of AmiC gave reasonable but weaker X-ray curve fits which did not favor the existence of tetrameric AmiC. It is concluded that AmiC exhibits novel ligand-dependent oligomerization properties in solution when these are compared to other members of the periplasmic binding protein superfamily, where AmiC exists in monomeric and trimeric forms, the proportions of which depend on the presence of acetamide or butyramide.

Pseudomonas aeruginosa is a ubiquitous gram negative rod-shaped bacterium that is an important opportunistic pathogen in man and other animals (Singleton & Sainsbury, 1987; Jawetz et al., 1987; Koch & Høiby, 1992). It can be isolated from infected burns, urinary tract infections, and the lungs of patients with cystic fibrosis. It can occasionally be pathogenic in stressed plants. *P. aeruginosa* can utilize short-chain aliphatic amides such as acetamide $CH_3{\cdot}CO{\cdot}NH_2$ as sole carbon and nitrogen sources, and these are hydrolyzed to ammonia and acetic acid. The enzyme system is induced by the presence of amides (Kelly & Clarke, 1962; Stanier et al., 1966). The amidase operon consists of five genes, namely, *amiE, amiB, amiC, amiR,* and *amiS*, in that order. AmiC is a soluble cytoplasmic protein that functions as an amide sensor and negative regulator of the amidase operon. AmiC controls the activity of the transcription antitermination factor AmiR, which in turn regulates expression of the amidase enzyme system. *amiE* is the gene which corresponds to the amidase enzyme, and *amiB* and *amiS* appear to form a membrane transport system for the importation of amide into the bacteria (Drew & Wilson, 1992; Wilson et al., 1995).

The combination of secondary structure predictions and fold recognition analyses indicated that, despite only 17% amino acid sequence identity, AmiC had the same protein fold as the leucine–isoleucine–valine binding protein (LivJ) of *Escherichia coli* (Sack et al., 1989a; Wilson et al., 1993). LivJ corresponds to the Cluster 4 subclass of periplasmic binding proteins (Tam & Saier, 1993). The prediction was confirmed by the crystal structure of AmiC bound to its substrate acetamide (Pearl et al., 1994). The similarity of the AmiC structure to that of periplasmic binding proteins is of interest in that these proteins form a large family of related structures that are involved with the transport of small molecules into bacteria (Tam & Saier, 1993). A total of eight different prokaryotic subclasses that bind to sugars, amino acids, and anions have been identified, and crystal structures have been determined for six of these (Table 1). Nonetheless AmiC exhibits distinct functional properties in that it controls AmiR in response to a signal from acetamide, while the periplasmic binding proteins transport small molecules within the inner bacterial membrane. A similar relationship with LivJ has also been identified for the extracellular domain of the eukaryotic protein glutamate receptor, which is involved in neurotransmitter activity (O'Hara et al., 1993; Stern-Bach et al., 1994). AmiC is constructed from two nonequivalent α-helix/β-sheet domains joined by three polypeptide links which flank a ligand-binding site in a large cleft between them (Figure 1). Interestingly, the binding site cleft in LivJ is opened by a domain movement of approximately 35° compared to that in AmiC (Figure 1). In the AmiC–acetamide crystal

* To whom correspondence and requests for reprints should be addressed.

[‡] Royal Free Hospital School of Medicine.

[§] University College London.

[∥] Present address: Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.

[⊗] Abstract published in *Advance ACS Abstracts*, June 15, 1997.

Table 1: R_G Analyses of Representative Crystal Structures for Periplasmic Binding Proteins[a]

periplasmic binding protein	Brookhaven PDB code	R_G (nm)	reference
Cluster 1 (molecular weight 40 600)			
maltodextrin-binding protein	1omp	2.48	Sharff et al. (1992)
maltodextrin-binding protein + maltose	2mbp	2.39	Spurlino et al. (1991)
Cluster 2 (molecular weight 33 000)			
arabinose binding protein + ligand	1abe, 1abf, 5abp	2.24−2.26	Quiocho et al. (1989)
Cluster 3 (molecular weight 26 100)			
histidine-binding protein + histidine	1hpb, 1hsl	1.99−2.02	Oh et al. (1994); Yao et al. (1994)
Lys−Arg−Orn-binding protein	2lao	2.12	Oh et al. (1993)
Lys−Arg−Orn-binding protein + ligand	1lst, 1laf, 1lah, 1lag	1.99	Oh et al. (1993)
Cluster 4 (molecular weight 36 800 and 41 200)			
Leu−Ile−Val-binding protein LivJ	2liv	2.43	Sack et al. (1989a)
leucine-binding protein	2lbp	2.44	Sack et al. (1989b)
acetamide-binding protein AmiC + acetamide	1pea	2.23	Pearl et al. (1994)
acetamide-binding protein AmiC + butyramide	-	2.25	O'Hara et al. (1997)
Cluster 5 (molecular weight 59 100)			
oligopeptide-binding protein + tri-/tetrapeptide	1ola, 1olb	2.54−2.56	Tame et al. (1994)
Cluster 6 (molecular weight 34 300)			
phosphate-binding protein + phosphate	1abh	2.23	Luecke & Quiocho (1990)
sulfate-binding protein + sulfate	1sbp	2.15	Pflugrath & Quiocho (1988)

[a] Crystal structures for at least 35 periplasmic binding proteins are available in the Brookhaven database.

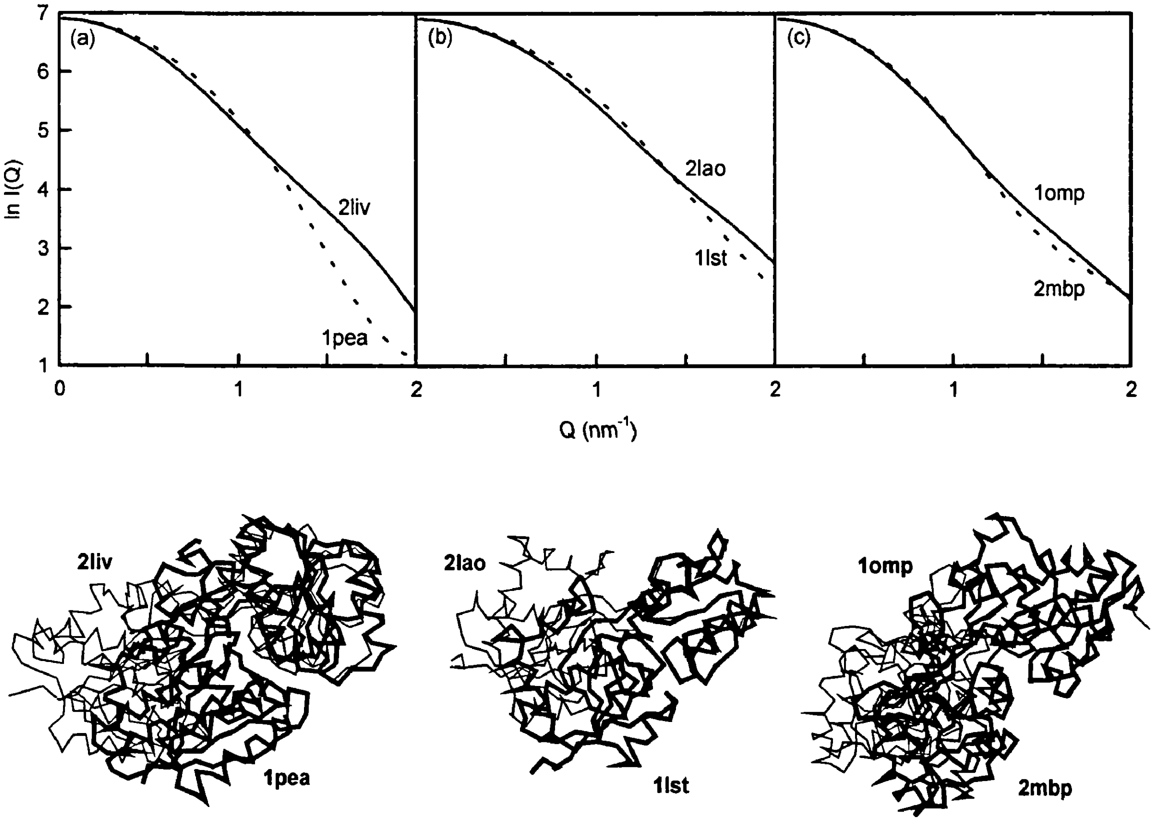


FIGURE 1: X-ray scattering curve simulations for Clusters 4, 3, and 1 of the periplasmic binding proteins. The dashed scattering curves correspond to the closed conformations. The α-carbon coordinates of the two proteins are shown with the C-terminal domains superimposed, and the closed conformations are represented in bold outline. The views of the α-carbon traces are shown to maximise the domain movement seen between the open and closed forms. (a) AmiC (1pea) and LivJ (2liv) in Cluster 4 were represented by 623 and 559 spheres, respectively, of radius 0.220 nm. (b) Lysine−argine−ornithine-binding protein (1lst and 2lao) in Cluster 3 was represented by 398 spheres of radius 0.220 nm. (c) Maltodextrin-binding protein (1omp and 2mbp) in Cluster 1 were represented by 94 and 96 spheres, respectively, of radius 0.410 and 0.405 nm.

structure, the cleft is substantially closed. The AmiC amide binding site is extremely specific for acetamide with a dissociation constant of 3.7 μM. Butyramide $CH_3{\cdot}CH_2{\cdot}CH_2{\cdot}CO{\cdot}NH_2$ is an anti-inducer of AmiC and has a 100-fold larger dissociation constant. It is possible that AmiC−acetamide and AmiC−butyramide may possess alternative conformations.

Small-angle X-ray and neutron scattering are powerful low-resolution methods for studies of the arrangement of domains in multidomain proteins and their degree of oligomerization (Perkins, 1988). They have advantages in that the data are obtained in solution. The utility of the methods has been much improved by the development of calibrated procedures for the calculation of scattering curves from known crystal structures (Smith et al., 1990; Perkins et al., 1993). Automated scattering curve fit procedures constrained by known atomic structures can now be used to assess the unknown structure of a multidomain protein (Mayans et al., 1995; Beavil et al., 1995; Boehm et al., 1996). The previous application of X-ray scattering to periplasmic binding proteins showed that L-arabinose binding protein was monomeric, and that on the addition of L-arabinose its R_G value of 2.12 nm decreased by 0.094 ± 0.033 nm (Newcomer et al., 1981). This decrease corresponded to a rotation of the two domains closer to one another by 18° ± 4°. Other periplasmic binding proteins showed a 52° domain rotation between ligated and unligated lysine–arginine–ornithine binding protein (Oh et al., 1993), and a 35° domain rotation between ligated and unligated maltodextrin binding protein (Sharff et al., 1992) (Figure 1). Here, X-ray and neutron scattering methods are applied to determine the solution structure of AmiC. Unlike the classical periplasmic binding proteins which are monomeric, we show that AmiC exists in monomeric and trimeric forms, the proportions of which depend on the presence of acetamide or butyramide. We assess whether a ligand-dependent conformational change may occur and describe how automated curve fit methods can be applied to interpret the scattering curves in terms of a structure for trimeric AmiC.

MATERIALS AND METHODS

(a) Expression and Purification of AmiC for Solution Scattering

The expression system consisted of a 1.3 × 10^3 base pair fragment of the amidase system containing the *amiC* open reading frame, cloned into a broad host range vector pMMB66HE, and transformed into a *P. aeruginosa* amidase deletion strain as described and characterized by Wilson and Drew (1991). The bacteria were fermented in a modified Oxoid No. 2 broth, and protein expression was started by the addition of isopropyl β-D-thiogalactopyranoside. Cells were harvested by low-speed centrifugation and lysed immediately by sonication in AmiC buffer (20 mM Tris-HCl, pH 8.0, 1 mM dithiothreitol, 1 mM EDTA, 1 mM phenylmethylsulfonyl fluoride) (Wilson et al., 1991). The supernatant after sonication was clarified by centrifugation at 25000*g* for 30 min, and AmiC was precipitated using 40–60% saturated $(NH_4)_2SO_4$. The AmiC fractions were pooled, resuspended in AmiC buffer and dialyzed overnight to remove $(NH_4)_2SO_4$. AmiC was purified further by ion exchange (Q-Sepharose, Pharmacia) when it was eluted in the range 450–550 mM NaCl using a 0–1 M NaCl gradient. These fractions were pooled, made up to 1.2 M $(NH_4)_2SO_4$, loaded onto a phenyl-Sepharose hydrophobic interaction column, and eluted using a 0–1 M $(NH_4)_2SO_4$ gradient. The pooled AmiC fractions were dialyzed for several days against AmiC buffer containing 10 mM butyramide to remove acetamide. AmiC was then concentrated in an Amicon pressure cell and purified by gel filtration as a single peak to give concentrations of up to 17 mg/mL. The absorption coefficient of AmiC (1%, 1 cm, 280 nm) was calculated as 13.6 (Perkins, 1986).

AmiC samples were stored at 4 °C and used within a few days for scattering experiments. For X-ray scattering and neutron scattering in H_2O buffers, the AmiC–butyramide samples were used as prepared above, and AmiC–acetamide was generated by adding 10 mM acetamide immediately prior to data collection to displace butyramide. For neutron scattering in ^2H_2O buffers, the AmiC–butyramide samples were dialyzed for 36 h with four buffer changes into AmiC buffer prepared in ^2H_2O and containing either 10 mM acetamide or butyramide. Alternatively, AmiC–butyramide was dialyzed into its ^2H_2O buffer containing 10 mM butyramide, and AmiC–acetamide was generated by adding 10 mM-acetamide in AmiC buffer in ^2H_2O immediately prior to data collection.

(b) X-ray and Neutron Scattering Data Collection

X-ray scattering curves were obtained in two beam sessions using a camera with a quadrant detector at Station 2.1 at the Synchrotron Radiation Source, Daresbury, U.K. (Towns-Andrews et al., 1989; Worgan et al., 1990). Sample–detector distances of 3.14 or 3.17 m were used, with beam currents of 122–173 mA and a storage ring energy of 2.0 GeV. This resulted in a usable Q range of 0.1–2.3 nm^{-1} ($Q = 4\pi \sin \theta/\lambda$; scattering angle $= 2\theta$; wavelength $= \lambda$). Data acquisition times were 10 min, obtained as 10 time frames of 1 min each as a control for radiation damage. Other details of data collection and analyses are described elsewhere [e.g., Beavil et al. (1995)]. Sample temperatures were set at 15 °C.

Neutron scattering data were obtained in one session on Instrument D11 at the Institut Laue–Langevin, Grenoble, France (Lindley et al., 1992). Sample–detector distances of 2.00 and 5.00 m were used. With the monochromator set for λ of 1.00 nm, and using a 64 × 64 cm^2 detector, the two detector settings resulted in a usable Q range of 0.06–1.1 nm^{-1}. When a rectangular beam aperture of 7 × 10 mm^2 was used, data acquisition times were typically 5 min in ^2H_2O buffers and 30 min in H_2O buffers. Samples were measured at 15 °C in rectangular quartz Hellma cuvettes of path length 2 mm for samples in ^2H_2O buffers and 1 mm for samples in H_2O buffers, and absorbances at 280 nm for AmiC concentrations were measured directly in the same cells. Sample and buffer transmissions were measured relative to an empty cell transmission for use in data reduction. Data were processed using standard Grenoble software (RNILS, SPOLLY, RGUIM, and RPLOT; Ghosh, 1989). A cadmium run for electronic and neutron background was first subtracted from each scattering curve. The buffer background run was subtracted from that of the sample run, and the result was normalized for the detector response by using a water run from which an empty cell background, corrected for the transmission of water, had been subtracted.

Neutron scattering data were also obtained in one session on the LOQ instrument at the pulsed neutron source ISIS at the Rutherford Appleton Laboratory, Didcot, U.K. (Heenan & King, 1993). The moderated pulsed neutron beam was derived from a tantalum target after proton bombardment at 50 Hz (proton beam current of 171 μA). Based on a fixed sample–detector distance of 4.3 m, the usable Q range was 0.1–2.0 nm^{-1}. The data acquisition time was 1 h at a sample temperaure of 15 °C. Other details of data collection and analyses are described elsewhere [e.g., Mayans et al. (1995) and Beavil et al. (1995)].

(c) Guinier and Distance Distribution Function Analyses of Reduced Scattering 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):

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

This expression is valid in a QR_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 weights M_r of the proteins when referenced against a suitable standard (Kratky, 1963; Jacrot & Zaccai, 1981; Wignall & Bates, 1987). 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 in reciprocal space $I(Q)$ into that in real space $P(r)$ was carried out using GNOM (Svergun et al., 1988; Semenyuk & Svergun, 1991; Svergun, 1992).

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

$P(r)$ corresponds to the distribution of distances r between 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. For this, the X-ray $I(Q)$ curve in the range between 0.3 and 2.0 nm^{-1} contained 345 data points, which were reduced to 255 points by GNOM for the transformation. The LOQ neutron $I(Q)$ contained 76 data points in the Q range between 0.1 and 2.1 nm^{-1}. GNOM employs a regularization procedure with an automatic choice of the transformation parameter α to stabilize the $P(r)$ calculation (Svergun, 1992). The $P(r)$ curve contains 61 points. A range of maximum assumed dimensions D_{max} 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; and (iii) the $P(r)$ curve should be stable as D_{max} is increased beyond the estimated macromolecular length.

(d) Automated Procedure for Debye Sphere Modeling of AmiC

The X-ray and neutron scattering curves were modeled using small single-density spheres to represent the AmiC 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 (Perkins & Weiss, 1983):

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

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

where $g(Q)$ is the squared form factor for the sphere of radius R, n is the number of spheres 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 calibrated with known crystal coordinates (Smith et al., 1990; Perkins et al., 1993).

The monomeric AmiC–acetamide coordinates (Brookhaven PDB accession code: 1pea) formed the asymmetric unit in space group $P4_22_12$, and were used for all calculations. The coordinates were converted to spheres by placing all residue atoms within a three-dimensional grid of cubes of side 0.457 nm. A cube was included in the model if it contained sufficient atoms above a specified cutoff such that the total volume of the 580 cubes equaled that of the dry protein of 55.0 nm^3 calculated from the sequence (SWISSPROT name, AMIC_PEASE; accession code, P27017) (Chothia, 1975; Perkins, 1986). As AmiC contains 384 residues, while only 369 residues were visible in the crystal structure for reason of crystallographic disorder at the N- and C-termini, this procedure compensated for the 4% smaller volume present in the crystal structure. X-ray curve fits were based on a rescaled hydrated model, whose volume is the sum of the dry model and that of a hydration shell of 0.3 g of H_2O/g of AmiC. The latter corresponds to an electrostricted volume of 0.0245 nm^3 per bound H_2O (Perkins, 1986). The rescaled cube coordinates have sides of 0.496 nm and correspond to spheres of radius 0.308 nm. The sphere sizes are much less than the nominal resolution of $2\pi/Q_{max}$ of the scattering curves. No corrections were applied for X-ray wavelength spread or beam divergence as these are considered to be negligible. For both LOQ and D11 data, a 16% spread in λ for a nominal λ of 1.0 nm and a beam divergence of 0.016 radians were used to correct the calculated neutron scattering curve for the reasons discussed in Mayans et al. (1995). Neutron curve fits were used after X-ray curve fitting to confirm that possible solute–solvent contrast effects were not significant. The R_G value of the model was calculated from the Guinier fit of the calculated curve in the same Q range used for experimental data. The quality of the curve fit was defined using an R-factor $R_{2.0}$ to measure the agreement between the experimental and calculated X-ray curves in the Q range between 0.1 and 2.0 nm^{-1} (Smith et al., 1990; Beavil et al., 1995). For a given set of models and curve fits, the R_G and $R_{2.0}$ values were imported into a spreadsheet for filtering and sorting to identify the best fit. Models for oligomers were not retained if they contained less than 95% of the required total of spheres in order to exclude models with significant steric overlap between the monomers.

In application to the comparative simulations of Figure 1, two changes were made: (i) The Brookhaven database files themselves were used directly in the simulations without correction for residues not observed in the electron density maps. (ii) As only α-carbon coordinates were reported in the 2mbp structure, only the α-carbon coordinates in the 1omp structure were used for reason of consistency.

In application to automated X-ray curve-fitting analyses, INSIGHT II 95.0 (Biosym/MSI, San Diego, CA) was used for all manipulations. Three approaches were developed: (i) A symmetric trimer was considered by orientating arbitrarily three monomers parallel to each other along their long axes such that they were related by 120° rotations about a 3-fold Z-axis of symmetry. Starting from a model in which the centers of the three monomers were close to the central 3-fold axis of symmetry and the monomers were sterically overlapped, further models for curve calculations were generated using INSIGHT II macros by moving the monomers outward from the central Z-axis in 0.2 nm steps in a total range of 4 nm. (ii) Mixtures of monomeric, dimeric, trimeric, and tetrameric AmiC were considered by calculating the scattering curves for each of the crystallographic monomer, dimer, trimer, and tetramer. The putative dimer was generated using the symmetry-related transformation x, y, z to y, x, $-z$ by application of the crystallographic dyad at x, x, $^1/_2$ (Pearl et al., 1994). A putative tetramer was then

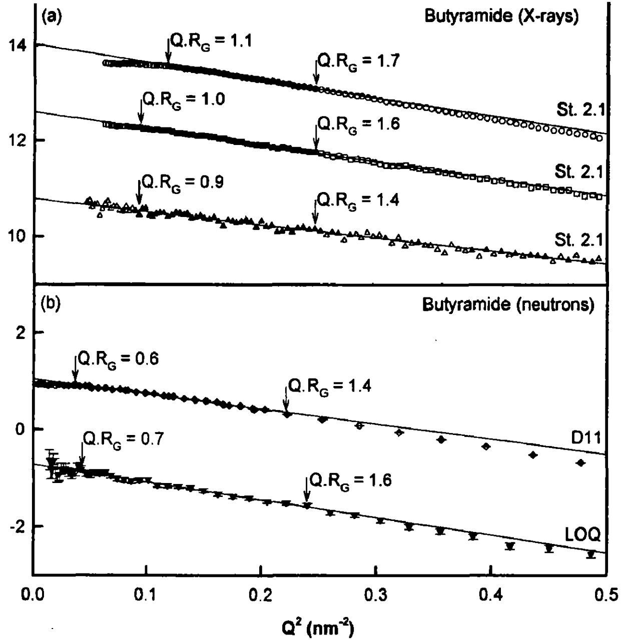


FIGURE 2: Guinier R_G plots for AmiC−butyramide. The filled symbols between the QR_G values as arrowed denote the range used to determine $I(0)$ and R_G. Statistical error bars are only shown when these are large enough to be visible. (a) Dilution series studied by synchrotron X-ray scattering for AmiC−butyramide at concentrations of 13.7 (○), 4.1 (□), and 1.0 mg/mL (△). The Q ranges for Guinier fits were 0.3−0.5 nm^{-1} for the 1 mg/mL curve and 0.35−0.5 nm^{-1} for the 13.7 mg/mL curve. (b) Neutron scattering data for AmiC−butyramide concentrations of 5.1 mg/mL used for D11 (◇) and 2.6 mg/mL used for LOQ (▽).

generated by application of the crystallographic dyad at $^1/_2$, $^1/_2$, z to this dimer. The putative trimer was formed by deleting any one of the four monomers in the tetramer. All combinations of these four curves in 1% steps were summed for fits with the experimental data. (iii) A putative asymmetric trimer model was considered using the crystallographic dimer and monomer. These were aligned manually so that their centers were close to each other without steric overlap. Cartesian axes were defined by reference to the center of mass of the AmiC monomer. The monomer was then moved −6 nm along its major translational Z-axis and −3 nm along the X- and Y-axes. The two structures were then translated in +0.2 nm steps relative to each other in the X, Y, and Z directions for distances of up to 12 nm using INSIGHT II macros, and the scattering curve was then calculated from each model for comparison with experimental data.

RESULTS AND DISCUSSION

(a) AmiC Oligomers by Synchrotron X-Ray Scattering

X-ray scattering data for AmiC in AmiC buffer containing 10 mM butyramide or 10 mM acetamide as appropriate (Methods) were obtained in the concentration range between 0.4 and 16.4 mg/mL. These are denoted as AmiC-buytramide and AmiC−acetamide respectively. Figure 2a shows that linear Guinier plots in satisfactory QR_G ranges were obtained. Guinier analyses of the 10 time frames used during data acquisition indicated the absence of radiation damage effects that are commonly seen with other proteins. However pronounced concentration effects were observed at above 10 mg/mL, where the Guinier plots exhibited diminished intensities at the lowest Q values. These are typical of interparticle interference effects when each protein molecule senses the presence of its neighbors (Guinier & Fournet, 1955). At these higher concentrations, a reduced QR_G range of fit corresponding to 0.35−0.5 nm^{-1} was required in order to obtain linear Guinier analyses.

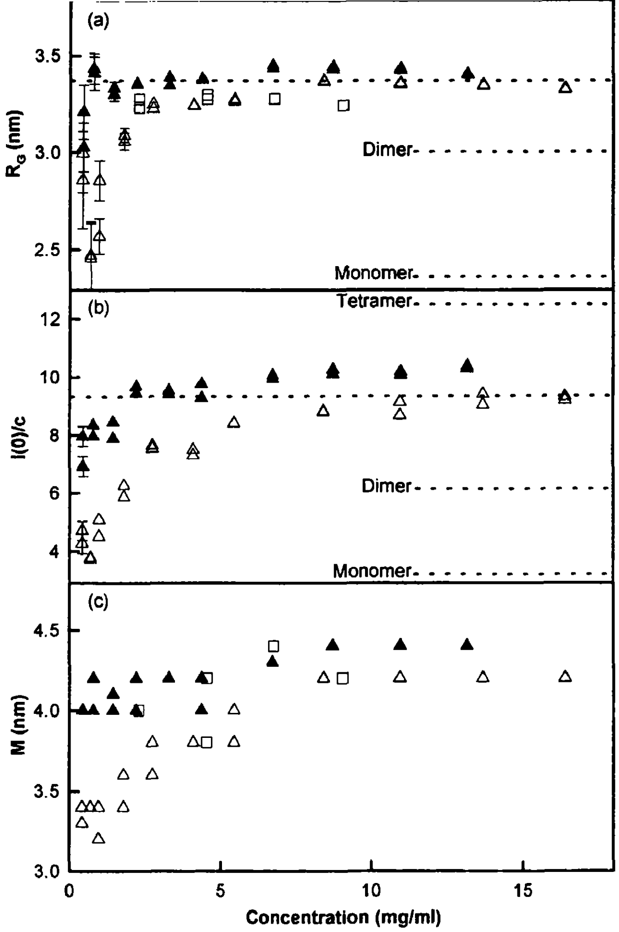


FIGURE 3: Concentration dependence of the AmiC X-ray Guinier and $P(r)$ parameters. Statistical error bars are only shown when these are large enough to be visible. (a) Those for the R_G values for AmiC−butyramide in two different beam-time sessions (□ and △) and for AmiC−acetamide (▲) is summarized. The dashed lines denote the R_G values of 1, 2, and 3 subunits of AmiC from Table 2 for comparison. (b) The corresponding $I(0)/c$ values for AmiC−butyramide and AmiC−acetamide are shown, using the same symbols as in a. The dashed lines denote the $I(0)/c$ values corresponding to the molecular weights for 1, 2, 3, and 4 subunits of AmiC. (c) The most frequently occurring distances M in $P(r)$ curves for AmiC−butyramide and AmiC−acetamide are shown, also using the same symbols as in a.

The Guinier analyses showed that, at concentrations below 5 mg/mL, both the R_G and $I(0)/c$ values decreased with decrease in AmiC concentration (Figures 3a and 3b). This is typical of the dissociation of an oligomeric protein. The R_G and $I(0)/c$ values were consistently higher for AmiC−acetamide when compared with AmiC−butyramide, in particular at AmiC concentrations below 2 mg/mL, and again at above 10 mg/mL. This suggested that the presence of acetamide induced a higher degree of oligomer formation in AmiC compared to butyramide. In contrast, L-arabinose binding protein behaved as a monomeric protein in the concentration range 6−36 mg/mL of protein (Newcomer et al., 1981).

To determine whether a conformational change could be detected in AmiC when the ligand was changed from butyramide to acetamide, the R_G values for the two forms were compared for curves with identical $I(0)/c$ values (i.e., similar degrees of oligomerization). For an $I(0)/c$ value of

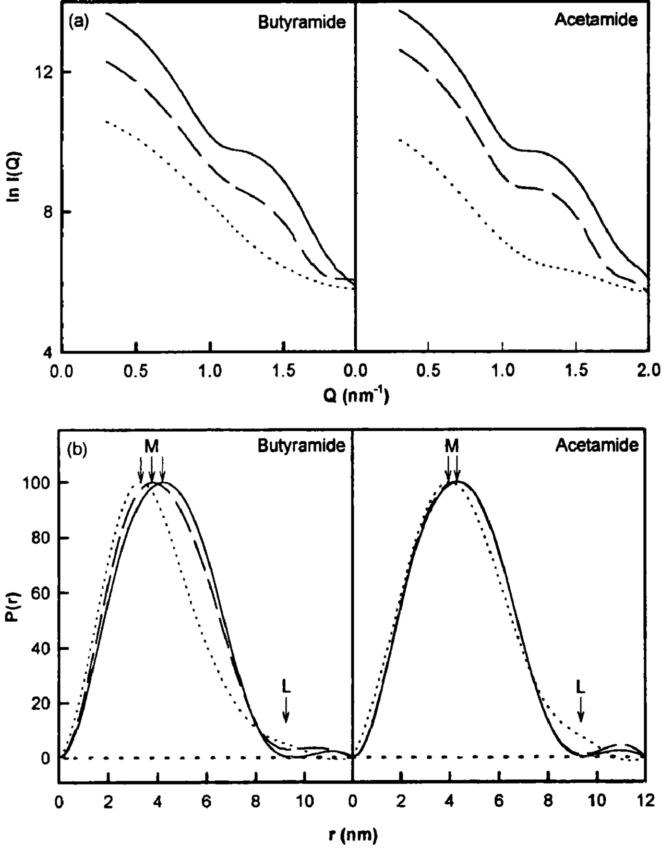


FIGURE 4: Comparison of the X-ray scattering curves $I(Q)$ and distance distribution functions $P(r)$ for AmiC−butyramide and AmiC−acetamide. (a) Concentration dependence of the X-ray $I(Q)$ curves for AmiC−butyramide and AmiC−acetamide. For clarity, the $I(Q)$ curves were smoothed using GNOM. Continuous, 13.7 mg/mL; dashed, 4.1 mg/mL; dotted, 1.0 mg/mL. (b) Corresponding distance distribution functions $P(r)$ for AmiC−butyramide and AmiC−acetamide. Each concentration is denoted as in a.

9.3, the full dashed lines in Figure 3a showed that the R_G values of AmiC−butyramide at 8−13 mg/mL were the same at 3.35 nm (within a range of 0.05 nm) to those for AmiC−acetamide at 3−5 mg/mL. This showed that no large conformational change had occurred within these limits. The errors in R_G values are larger at low concentrations, and it was not possible to consider this question for AmiC below 2 mg/mL. If the binding of acetamide had induced a rotational closure of the cleft in AmiC, the R_G value would be expected to be smaller by about 0.1 nm by analogy with L-arabinose binding protein (Newcomer et al., 1981).

Changes in AmiC oligomerization were also visible in the full scattering curves $I(Q)$ out to $Q = 2$ nm^{-1} (Figure 4a), in which a submaximum at Q of 1.2 nm^{-1} at high AmiC concentration disappeared at low AmiC concentration. Satisfactory distance distribution functions $P(r)$ were calculated from $I(Q)$ curves on the basis of a presumed maximum dimension D_{max} of 12 nm (Figure 4b). The $P(r)$ curves offered an alternative determination of R_G and $I(0)/c$, and these corroborated the Guinier analyses of Figures 3a and 3b (data not shown). The $P(r)$ curves also demonstrated a concentration dependence which is larger for AmiC−butyramide. At all concentrations, the maximum dimension L of AmiC is close to 9 nm and shows that the different oligomers are similar in overall length. The peak maximum M of the $P(r)$ curves corresponds to the most commonly occurring distance in AmiC. The concentration dependence of M in Figure 3c exhibited similar trends to those already observed with the R_G and $I(0)/c$ values, in that its position demonstrated a greater concentration dependence with AmiC−butyramide and ranged from 4.2 to 3.2 nm.

(b) Identification of AmiC Trimers by Neutron Scattering

Neutron scattering for AmiC in ^2H_2O buffers provided molecular weights as well as acting as a control for the absence of X-ray radiation damage effects and internal scattering density inhomogeneity effects in different solvents. AmiC is now visualized in a high negative solute−solvent contrast in place of the high positive contrast seen by X-rays. Linear Guinier R_G plots were obtained from both the neutron cameras D11 and LOQ (Figure 2b). The higher neutron flux on D11 permitted a dilution series of AmiC−butyramide and AmiC−acetamide to be measured between 0.8−5.1 mg/mL (data not shown). The neutron Guinier $I(0)/c$ values confirmed the X-ray concentration dependence in Figure 3b. The mean D11 R_G value for AmiC−butyramide was 3.26 ± 0.08 nm (five determinations between 1.2 and 5.1 mg/mL), and that for AmiC−acetamide was 3.30 ± 0.06 nm (four determinations between 0.8 and 1.9 mg/mL). The corresponding LOQ R_G values were in agreement at 3.30 ± 0.05 nm for AmiC−butyramide (one determination at 2.6 mg/mL) and 3.26 ± 0.06 nm for AmiC−acetamide (one determination at 2.8 mg/mL). The neutron R_G values were close to but slightly less than the X-ray value of 3.35 nm as expected (Table 2). The small decrease of up to 0.1 nm in the neutron R_G values is attributable to the surface location of hydrophilic amino acids and the core location of hydrophobic amino acids in AmiC, since hydrophilic residues have a higher scattering density than hydrophobic residues (Perkins, 1986, 1988).

M_r calculations were performed from the neutron $I(0)/c$ values, as $I(0)/c$ is measured relative to known standards. For D11 data, $I(0)/c$ for AmiC−butyramide at 3.9 mg/mL in H_2O buffer was determined to be 0.072 ± 0.005 relative to the incoherent scattering of H_2O at a wavelength of 1.0 nm, and this gave an M_r of 127 000 ± 10 000. Since monomeric AmiC has an M_r of 42 600, this is equivalent to 3.0 ± 0.2 subunits. For LOQ data, the mean $I(0)/c$ value of 0.176 observed for AmiC−butyramide and AmiC−acetamide referenced to a known polymer standard and other $I(0)/c$ values determined for five proteins of known M_r between 51 000 and 144 000 (Mayans et al., 1995; Ashton et al., 1995; Beavil et al., 1995) gave an M_r of 150 000 ± 25 000, which corresponds to 3.6 ± 0.6 subunits. The full X-ray $I(0)/c$ concentration series in Figure 3b shows that AmiC is predominantly trimeric between 5 and 10 mg/mL and undergoes significant dissociation at concentrations below 5 mg/mL. As an $I(0)/c$ value of 9.3 can be assigned to 3 AmiC subunits in Figure 3b, an $I(0)/c$ value of 3.1 will correspond to monomeric AmiC. Figure 3a shows that the R_G of the AmiC monomer is less than 2.5 nm and that AmiC dissociates into monomers at low concentrations.

(c) X-ray Scattering Curve Simulations for Three Periplasmic Binding Proteins

Curve simulations were performed using known crystal structures for free and complexed forms of the periplasmic binding proteins in order to assess whether solution scattering will monitor domain movements between their open and closed conformations.

(i) The periplasmic binding proteins from six clusters (Tam & Saier, 1993) exhibited R_G values between 1.99 and 2.56

Table 2: X-ray and Neutron Scattering Parameters for AmiC Samples and Models

technique (instrument)	protein	concentration (mg/mL)	experimental R_G (nm)
synchrotron X-ray (St 2.1)	AmiC−butyramide	7−16	3.35 ± 0.05
	AmiC−acetamide[a]	2−5	3.35 ± 0.05
	AmiC−acetamide[a]	0.4	3.12 ± 0.13
neutron (D11)	AmiC−butyramide	1.2−5.1	3.26 ± 0.08
	AmiC−acetamide[a]	0.8−1.9	3.30 ± 0.06
neutron (LOQ)	AmiC−butyramide	2.6	3.30 ± 0.05
	AmiC−acetamide[a]	2.8	3.26 ± 0.06

AmiC models	concentration (mg/mL)	R factor (2.0 nm^{-1})	modeled R_G (nm)
AmiC scattering (symmetric trimer)	6.8[b]	4.7	3.39
AmiC crystallographic monomer			2.34
AmiC crystallographic dimer			2.97
AmiC crystallographic trimer[c]			3.53
AmiC crystallographic tetramer			3.66
AmiC scattering (dimer + tetramer)	6.8[b]	6.3	3.38
AmiC scattering (asymmetric trimer 1)	6.8[b]	4.1	3.34
AmiC scattering (asymmetric trimer 2)	6.8[b]	3.9	3.32

[a] Not shown in Figure 2. [b] Shown in Figures 5 and 6. [c] Generated by deleting any one of the four AmiC structures in the tetramer.

nm in an M_r range between 26 100 and 59 100 (Table 1). The Cluster 4 proteins AmiC and LivJ showed a decrease of 0.21 nm in R_G values on going from the unbound to the complexed form. The Cluster 3 proteins gave a smaller decrease of 0.13 nm, and that for the Cluster 1 proteins gave a decrease of 0.08 nm (Table 1).

(ii) Corresponding changes were seen in the full scattering curves out to Q of 2.0 nm^{-1} for these three groups of proteins (Figure 1). The scattering curve at low Q exhibited small changes which corresponded to the changes in the Guinier region. More noticeable intensity changes between the free and complexed forms were visible in the Q range beyond 1 nm^{-1}.

Figure 1 also indicates the domain movements between the free and complexed forms of these proteins when the C-terminal domains were superimposed upon each other. While large domain movements of the order of 30−40° are observed and are detectable by solution scattering, Figure 1 and Table 1 show that accurate measurements will be required. In the case of trimeric AmiC, no domain movements could be detected within a precision in R_G values of 0.05 nm.

(d) X-ray Scattering Curve Simulations for Trimeric AmiC

To extend the data interpretation, scattering curve simulations were performed in three different analyses for trimeric AmiC, starting from the crystal structure for AmiC−acetamide. The trimer will have a 3-fold axis of symmetry, as observed crystallographically for proteins such as tumor necrosis factor α, deoxyUTPase, and chloramphenicol transferase. A structure based on the asymmetric association of a monomer with a dimer is most unlikely on the grounds of symmetry. If a monomer is bound to one face of a dimer in such a trimer, a symmetry-related site for a second monomer will exist on the other side of the dimer, and AmiC would be tetrameric.

A symmetric AmiC homotrimer was constructed from three monomers whose longest axes were aligned parallel to each other with their ligand clefts arbitrarily set to face outward and with a 3-fold axis of symmetry between them along the Z-axis (Figure 5). Based on the scattering curve for AmiC−butyramide at 6.8 mg/mL, a one-parameter translational search explored the effect of varying the separation between the monomers in the XY-plane while retaining 3-fold symmetry. The best model by this approach in Figure 5 gave a good curve fit, 3s in Figure 6, with a low $R_{2.0}$ value of 4.7%. This model also resulted in a good fit (not shown) to the experimental curve at 1 mg/mL in Figure 4a with a satisfactory $R_{2.0}$ value of 7.3%, using a scattering curve constructed from 40% monomer and 60% homotrimer (curves 1 and 3s in Figure 6). This monomer:trimer ratio resulted in an estimated association constant K_{a3} of 2 × 10^{10} M^{-2}, where $K_{a3} = c_{\text{trimer}}/(c_{\text{monomer}})^3$ (McRorie & Voelker, 1993).

A second analysis was based on the monomer in the crystal structure of AmiC−acetamide, together with the putative dimer, trimer, and tetramer (curves 1, 2, 3, and 4, respectively, in Figure 6: see Materials and Methods). The curves changed in the Q range between 0.0 and 0.5 nm^{-1} to correspond to the increase in R_G with oligomerization (Table 2). The dimer model (Figure 5) gave a reasonable curve fit for AmiC−butyramide at 1.0 mg/mL with an $R_{2.0}$ value of 8.0%, but this fit was visibly not as good as that for the monomer−homotrimer mixture above. In terms of the AmiC−butyramide scattering curve at 6.8 mg/mL, the four models gave poor curve fits with $R_{2.0}$ values of 15.3%, 12.1%, 9.7%, and 39.3%, respectively. In particular, the experimental curve at 6.8 mg/mL showed a subminimum at Q = 1.12 nm^{-1}, which is different from that at Q = 0.98 nm^{-1} calculated from the tetramer model (curve 4). It was postulated that the observed curve may represent a combination of the four curves. Analysis of 5151 combinations of any three curves stepped in 1% increments gave a best fit with 0% monomer, 51% dimer, and 49% tetramer. Analysis of all 176 851 combinations of four curves gave a best fit with 0% monomer, 51% dimer, 0% trimer, and 49% tetramer (curves 2 and 4 in Figure 6). Although the $R_{2.0}$ value of 6.3% for this fit is reasonable, the curve fit is seen to deviate at Q values above 0.8 nm^{-1}. The limited success of these fits showed that the putative tetramer does not exist, and this supports the modeling based on a monomer−trimer equilibrium.

A third curve fit search assumed that AmiC at 6.8 mg/mL corresponded to an asymmetric trimer formed from the crystallographic monomer and dimer. In three-parameter X-,

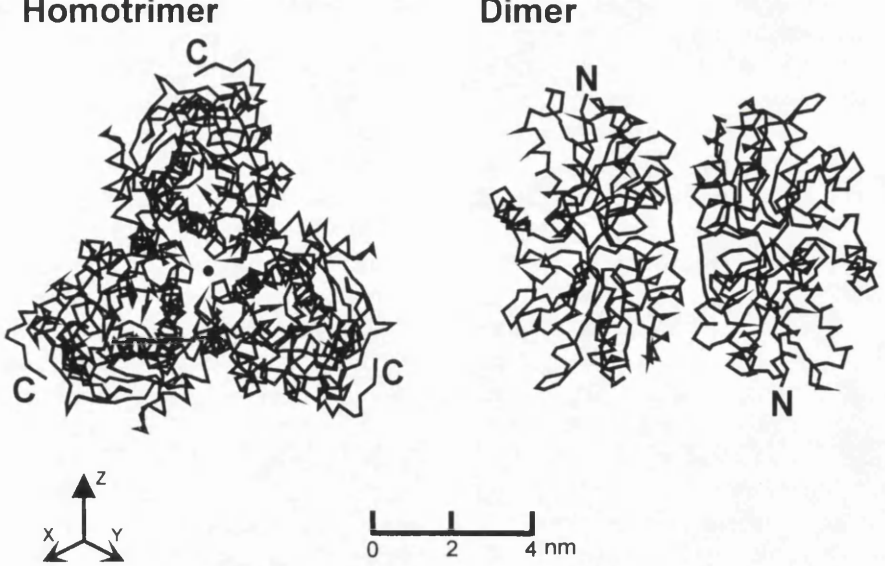


FIGURE 5: α-Carbon outlines of the homotrimer and dimer models for AmiC, based on the AmiC−acetamide crystal structure. The homotrimer of AmiC is viewed down the Z-axis which is indicated by the dot at the center of the structure. The C-termini in this model are denoted by C. In the final model, the center of mass of the monomer is 2.2 nm from the center of mass of the trimer on its 3-fold axis of symmetry. The putative crystallographic dimer is depicted as an antiparallel association of two monomers, in which the N-termini are denoted by N. In the putative crystallographic tetramer, the second dimer is rotated clockwise by 135° which is then positioned in front of the first dimer as shown.

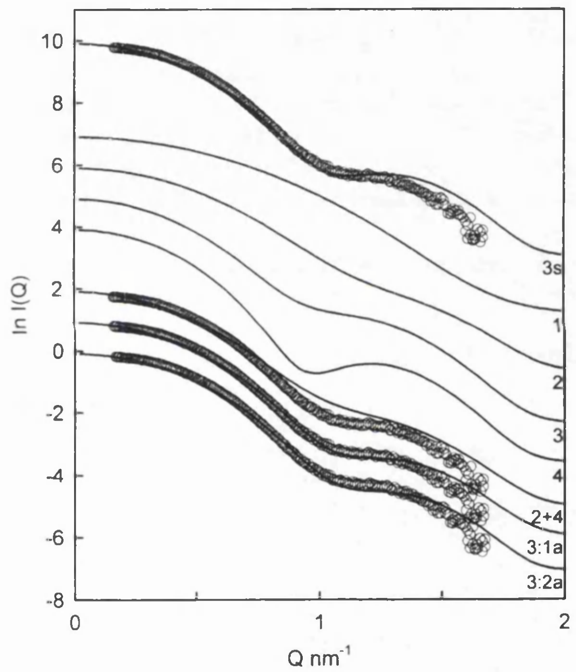


FIGURE 6: Curve fits of experimental scattering curves for AmiC−butyramide. Symmetric homotrimer, 3s; monomer, 1; dimer, 2; trimer, 3; tetramer, 4; asymmetric trimers, 3:1a and 3:2a (Table 2). The curves were compared with X-ray data for AmiC−butyramide at 6.8 mg/mL. The asymmetric trimer 1 corresponds to an AmiC monomer with its long axis perpendicular to that of the dimer (Figure 5). The monomer coordinates are superimposed on one of the two monomers in the dimer by translations of $X = 1.7$ nm, $Y = 2.7$ nm, $Z = -1.5$ nm and rotations of $X = 95°$, $Y = 12°$, $Z = -90°$. Trimer 2 was generated from a 90° reorientation of the AmiC monomer such that the long axes of the monomer and dimer are parallel. The monomer coordinates are superimposed on one of the two monomers in the dimer by translations of $X = -0.02$ nm, $Y = -3.0$ nm, $Z = -2.0$ nm and rotations of $X = -14°$, $Y = -20°$, $Z = 195°$.

Y-, and Z-axis translational searches, the long axis of the monomer was set perpendicular (curve 3:1a) or parallel (curve 3:2a) to that of the dimer. A clear minimum in the $R_{2.0}$ values was obtained for each of the X-, Y-, and Z-axes in trial translational searches and showed that a global minimum could be defined. The full systematic search was based on 31 × 31 × 41 steps of 0.2 nm along the X-, Y-, and Z-axes and gave 39 401 coordinate models. Models were selected if they contained at least 1650 of the expected total of 1752 spheres (i.e., only those models without monomer−dimer steric overlap were retained) and had a calculated R_G value between 3.0 and 3.5 nm. From these searches, the best curve fits, 3:1a and 3:2a, had similar $R_{2.0}$ values of 3.9% and 4.1%, respectively (Figure 6 and Table 2). While these $R_{2.0}$ values are now better than those above, the importance of this search is that it showed that a good fit can be obtained from a model with incorrect symmetry. As these models gave the best fits, they were also used for neutron curve fits as a check of consistency. The fits (not shown) gave good R factors in both H_2O and ^2H_2O buffers, and the calculated curve deviated slightly from the two observed curves in opposite directions at larger Q values as expected from the two opposite solute−solvent contrasts in use.

CONCLUSIONS

The classical view of periplasmic binding proteins is that they are monomeric with two domains that undergo large conformational change during ligand binding. While AmiC is most closely related to the Cluster 4 protein LivJ in sequence and structure, AmiC is a cytoplasmic protein that controls the activity of AmiR which is directly involved with *ami* gene expression, while LivJ is a perisplasmic protein that binds aliphatic amino acids. Both exhibit similar interactions with membrane-bound proteins (Wilson et al., 1995). Unexpectedly AmiC exhibits oligomeric properties at high concentrations. Other periplasmic binding proteins are generally monomeric, although the galactose binding protein from *Salmonella typhimurium* and *E. coli* and the maltose binding protein from *E. coli* are dimeric, and the addition of ligand causes them to become monomeric (Mowbray & Petsko, 1983; Richarme, 1982). The crystal structures of histidine-binding protein (1hsl) and a malto-dextrin-binding protein mutant (1mdp) reveal dimeric structures; however, these are attributable to repeated lattice contacts between the same domain in a pair of monomers. In distinction to these examples, AmiC forms antiparallel dimers in its crystal structure (Figure 5), most probably the consequence of the crystal lattice, and it participates in a monomer−trimer association in solution. These results show that AmiC behaves differently from the classical periplasmic binding proteins.

Above 2 mg/mL, AmiC is predominantly trimeric. That trimer formation is promoted in the presence of its ligand acetamide rather than the anti-inducer butyramide may be of biological interest. The concentration dependence of the scattering curves is clear from Figure 4, and the $I(0)/c$ graphs rise with c to a value corresponding to trimers in Figure 3. Initial analytical ultracentrifugation data by sedimentation equilibrium methods using a Beckman XLA ultracentrifuge also showed this concentration dependence and yielded a comparable estimate of the association constant K_{a3} (O. Byron, D. Chamberlain, and S. J. Perkins, unpublished data; Ralston, 1993; McRorie & Voelker, 1993). Molecular modeling based on the AmiC−acetamide crystal structure showed that trimeric structures gave R_G values that corresponded to the observed values. The possibility of an imposter model based on a mixture of dimers and tetramers was also considered. This was discounted for two reasons: (i) A single peak and not a double peak was routinely observed during AmiC purifications by gel filtration (Methods). (ii) Modeling of the X-ray data showed that it was not possible to optimize a curve fit based on a mixture of dimer and tetramer that was equivalent to or better than one based on a trimer (Figure 6).

Unlike periplasmic binding proteins in general, scattering showed that no conformational changes between AmiC−acetamide and AmiC−butyramide trimers could be detected within a precision of 0.05 nm in R_G value. Butyramide contains an extra pair of $CH_2 \cdot CH_2$ carbon atoms. As butyramide binds 100-fold more weakly to AmiC than acetamide (Wilson et al., 1993), this weaker binding may reflect the energy required to accommodate butyramide within its binding site in AmiC without a large cleft opening. Trimer formation in AmiC would involve extensive contacts between the monomers and may hinder the free movement of the cleft on change of ligand. The question of whether this domain movement occurs or not on ligand binding will require data collection on monomeric AmiC or on the complex between AmiC and AmiR. Nonetheless the larger size of the butyramide ligand has clearly reduced the stability of the AmiC trimer.

Solution scattering will monitor domain movements in the periplasmic binding proteins, even though the changes are small (Figure 1 and Table 1). Here, even though L-arabinose binding protein was monomeric and well-behaved between concentrations of 4 and 36 mg/mL (Newcomer *et al.*, 1981), AmiC demonstrated interparticle interference effects above 10 mg/mL as well as oligomer formation in its scattering curves. Dilution series and absolute M_r calculations were key controls of the present scattering data. Data analyses were enhanced by the use of automated constrained curve fitting procedures. Hypotheses could be stated which could then be tested in detail with relatively little effort, although expensive in terms of CPU time, and enabled a choice to be made between a monomer−trimer or a monomer−dimer−tetramer association. Automated curve fits had previously been used to assess domain structures in single large multidomain proteins (Mayans et al., 1995; Beavil et al., 1995; Boehm et al., 1996). The present study shows that the method is applicable to the study of protein−protein complexes. It should be noted that a good curve fit is only a test of consistency and will not constitute a unique low-resolution structure determination. The good curve fit obtained for the symmetry-forbidden asymmetric trimer of AmiC is an illustration of this limitation.

Following the scattering studies, thermal denaturation experiments (B. P. O'Hara, G. Siligardi, S. A. Wilson, R. E. Drew, and L. H. Pearl, 1997, manuscript in preparation) have shown that AmiC−butyramide is less stable than AmiC−acetamide. The crystal structure of AmiC−butyramide has been determined and shows that no major conformational changes in AmiC were observed. The root mean square difference between the α-carbon coordinates of both forms of AmiC was 0.040 nm. This is supported by circular dichroism spectroscopy of AmiC−butyramide and AmiC−acetamide which also suggested no major conformational changes. These results are consistent with the present scattering data.

ACKNOWLEDGMENT

We thank the Biotechnology and Biological Sciences Research Council for an Earmarked studentship to D.C. and for grant support. We thank Dr. R. E. Drew for useful discussions, Mr. A. W. Ashton for computational support, the Engineering and Physical Sciences Research Council for X-ray and neutron 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.

REFERENCES

Ashton, A. W., Kemball-Cook, G., Johnson, D. J. D., Martin, D. M. A., O'Brien, D. P., Tuddenham, E. D. G., & Perkins, S. J. (1995) *FEBS Lett. 374*, 141−146.

Beavil, A. J., Young, R. J., Sutton, B. J., & Perkins, S. J. (1995) *Biochemistry 34*, 14449−14461.

Boehm, M. K., Mayans, M. O., Thornton, J. D., Begent, R. H. J., Keep, P. A., & Perkins, S. J. (1996) *J. Mol. Biol. 259*, 718−736.

Chothia, C. (1975) *Nature (London) 254*, 304−308.

Drew, R. E., & Wilson, S. A. (1992) in *Pseudomonas: Molecular Biology and Biotechnology* (Galli, E., Silver, S., & Witholt, E., Eds) pp 207−213, American Society of Microbiology, Washington, DC.

Ghosh, R. E. (1989) Institut Laue-Langevin Internal Publication No. 89GH02T, Institut Laue-Langevin, Grenoble, France.

Glatter, O., & Kratky, O., Eds. (1982) *Small-Angle X-ray Scattering*, Academic Press, New York.

Guinier, A., & Fournet, G. (1955) *Small Angle Scattering of X-rays*, Wiley, New York.

Heenan, R. K., & King, S. M. (1993) Proceedings of an International Seminar on Structural Investigations at Pulsed Neutron Sources, Dubna, Russia, September 1−4, 1992, Report E3-93-65, Joint Institute for Nuclear Research, Dubna, Russia.

Heenan, R. K., King, S. M., Osborn, R., & Stanley, H. B. (1989) Colette Users Guide, Internal Publication No. RAL-89-128, Rutherford Appleton Laboratory, Didcot, U.K.

Jacrot, B., & Zaccai, G. (1981) *Biopolymers 20*, 2413−2426.

Jawetz, E., Melnick, J. L., & Adelberg, E. A. (1987) *Review of Medical Microbiology*, 17th ed., Appleton & Lange, Norwalk, CT.

Kelly, M., & Clarke, P. H. (1962) *J. Gen. Microbiol. 27*, 305−316.

Koch, C., & Høiby, N. (1993) *Lancet 341*, 1065−1069.

Kratky, O. (1963) *Progr. Biophys. Chem. 13*, 105−173.

Lindley, P., May, R. P., & Timmins, P. A. (1992) *Physica B 180*, 967−972.

Luecke, H., & Quiocho, F. A. (1990) *Nature (London) 347*, 402−406.

Mayans, M. O., Coadwell, W. J., Beale, D., Symons, D. B. A., & Perkins, S. J. (1995) *Biochem. J. 311*, 283−291.

McRorie, D. K., & Voelker, P. J. (1993) *Self-Associating Systems in the Analytical Ultracentrifuge*, Beckman Instruments Inc., Fullerton, CA.

Mowbray, S. L., & Petsko, G. A. (1983) *J. Biol. Chem. 258*, 7991−7997.

Newcomer, M. E., Lewis, B. A., & Quiocho, F. A. (1981) *J. Biol. Chem. 256*, 13218−13222.

Oh, B.-H., Pandit, J., Kang, C.-H., Nikaido, K., Gokcen, S., Ames, G. F.-L., & Kim, S.-H. (1993) *J. Biol. Chem. 268*, 11348−11355.

Oh, B.-H., Kang, C.-H., De Bondt, H., Kim, S.-H., Nikaido, K., Joshi, A. K., & Ames, G. F.-L. (1994) *J. Biol. Chem. 269*, 4135−4143.

O'Hara, P. J., Sheppard, P. O., Thøgersen, H., Venezia, D., Haldeman, B. A., McGrane, V., Houamed, K. M., Thomsen, C., Gilbert, T. L., & Mulvihill, E. R. (1993) *Neuron 11*, 41−52.

Pearl, L. H., O'Hara, B., Drew, R. E., & Wilson, S. A. (1994) *EMBO J. 13*, 5810−5817.

Perkins, S. J. (1986) *Eur. J. Biochem. 157*, 169−180.

Perkins, S. J. (1988) in *New Comprehensive Biochemistry* (Neuberger, A., & Van Deenen, L. L. M., Eds.) Vol. 11B, Part 2, pp 143−264, Elsevier, Amsterdam.

Perkins, S. J., & Weiss, H. (1983) *J. Mol. Biol. 168*, 847−866.

Perkins, S. J., Smith, K. F., Kilpatrick, J. M., Volanakis, J. E., & Sim, R. B. (1993) *Biochem. J. 295*, 87−99.

Pflugrath, J. W., & Quiocho, F. A. (1988) *J. Mol. Biol. 200*, 163−180.

Quiocho, F. A., Wilson, D. K., & Vyas, N. K. (1989) *Nature (London) 340*, 404−407.

Ralston, G. (1993) *Introduction to Analytical Ultracentrifugation*, Beckman Instruments Inc., Fullerton, CA.

Richarme, G. (1982) *Biochem. Biophys. Res. Commun. 105*, 476−481.

Sack, J. S., Saper, M. A., & Quiocho, F. A. (1989a) *J. Mol. Biol. 206*, 171−191.

Sack, J. S., Trakhanov, S. D., Tsigannik, I. H., & Quiocho, F. A. (1989b) *J. Mol. Biol. 206*, 193−207.

Semenyuk, A. V., & Svergun, D. I. (1991) *J. Appl. Crystallogr. 24*, 537−540.

Sharff, A. J., Rodseth, L. E., Spurlino, J. C., & Quiocho, F. A. (1992) *Biochemistry 31*, 10657−10663.

Singleton, P., & Sainsbury, D. (1987) *Dictionary of Microbiology and Molecular Biology*, 2nd ed., pp 719−720, Wiley, Chichester, U.K.

Smith, K. F., Harrison, R. A., & Perkins, S. J. (1990) *Biochem. J. 267*, 203−212.

Spurlino, J. C., Lu, G.-Y., & Quiocho, F. A. (1991) *J. Biol. Chem. 266*, 5202−5219.

Stanier, R. Y., Palleroni, N. J., & Doudoroff, M. (1966) *J. Gen. Microbiol. 43*, 159−271.

Stern-Bach, Y., Bettler, B., Hartley, M., Sheppard, P. O., O'Hara, P. J., & Heinemann, S. F. (1994) *Neuron 13*, 1345−1357.

Svergun, D. I. (1992) *J. Appl. Crystallogr. 25*, 495−503.

Svergun, D. I., Semenyuk, A. V., & Feigin, L. A. (1988) *Acta Crystallogr. A44*, 244−250.

Tam, R., & Saier, M. H., Jr. (1993) *Microbiol. Rev. 57*, 320−346.

Tame, J. R. H., Murshudov, G. N., Dodson, E. J., Neil, T. K., Dodson, G. G., Higgins, C. F., & Wilkinson, A. J. (1994) *Science 264*, 1578−1581.

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) *Rev. Sci. Instrum. 60*, 2346−2349.

Wignall, G. D., & Bates, F. S. (1987) *J. Appl. Crystallogr. 20*, 28−40.

Wilson, S. A., & Drew, R. E. (1991) *J. Bacteriol. 173*, 4914−4921.

Wilson, S. A., Chayen, N. E., Hemmings, A. M., Drew, R. E., & Pearl, L. H. (1991) *J. Mol. Biol. 222*, 869−871.

Wilson, S. A., Wachira, S. J., Drew, R. E., Jones, D., & Pearl, L. H. (1993) *EMBO J. 12*, 3637−3642.

Wilson, S. A., Williams, R. J., Pearl, L. H., & Drew, R. E. (1995) *J. Biol. Chem. 270*, 18818−18824.

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) *Nucl. Instrum. Methods Phys. Res. A291*, 447−454.

Yao, N., Trakhanov, S., & Quiocho, F. A. (1994) *Biochemistry 33*, 4769−4779.

BI9703251

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 tc 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 × 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 $2\pi/Q_{\mathrm{max}}$ of the scattering curves (2.7 nm for $Q_{\mathrm{max}} = 2.3$ 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.

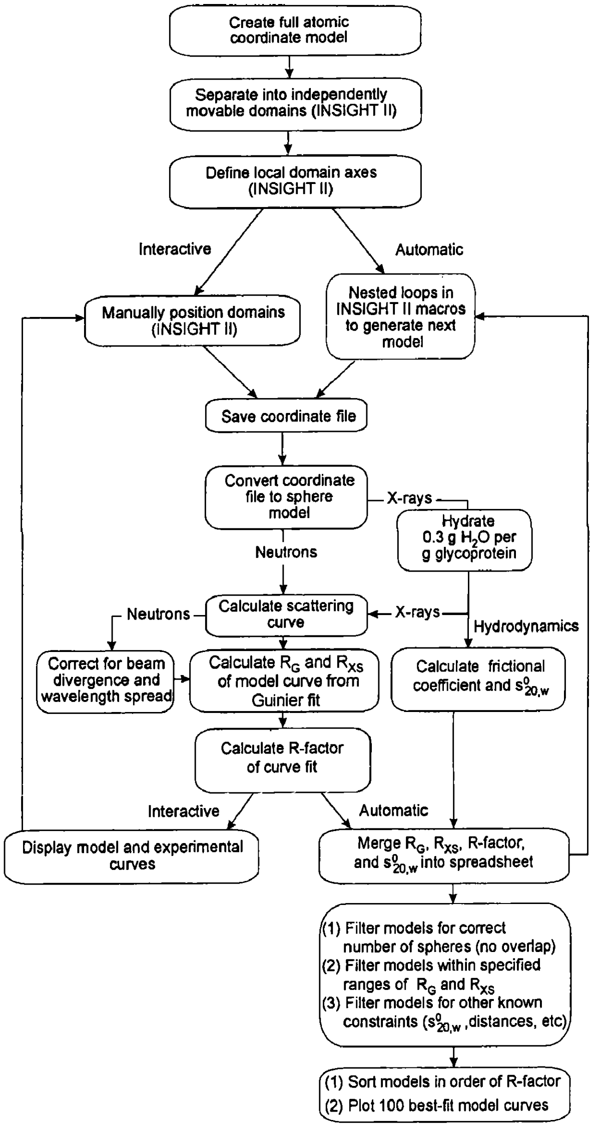

Create full atomic coordinate model
Separate into independently movable domains (INSIGHT II)
Define local domain axes (INSIGHT II)
Interactive
Automatic
Manually position domains (INSIGHT II)
Nested loops in INSIGHT II macros to generate next model
Save coordinate file
Convert coordinate file to sphere model
X-rays
Hydrate 0.3 g H_2O per g glycoprotein
Neutrons
Calculate scattering curve
Neutrons
X-rays
Hydrodynamics
Correct for beam divergence and wavelength spread
Calculate R_G and R_{XS} of model curve from Guinier fit
Calculate frictional coefficient and $s^0_{20,w}$
Calculate R-factor of curve fit
Interactive
Automatic
Display model and experimental curves
Merge R_G, R_{XS}, R-factor, and $s^0_{20,w}$ into spreadsheet
(1) Filter models for correct number of spheres (no overlap)
(2) Filter models within specified ranges of R_G and R_{XS}
(3) Filter models for other known constraints ($s^0_{20,w}$,distances, etc)
(1) Sort models in order of R-factor
(2) Plot 100 best-fit model curves

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.

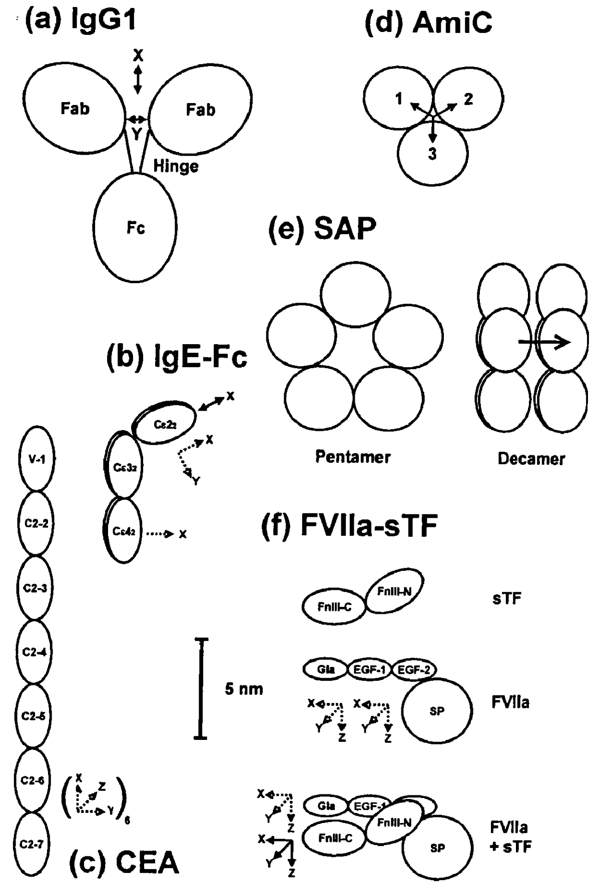


Fig. 2. Schematic outlines of six multidomain or oligomer structures to show how domain or subunit translations and rotations were implemented during the curve fit analyses. Translations are denoted by solid arrows, and rotations by dashed arrows: (a) For IgG1 and IgG2, the pair of Fab fragments were moved together in two-parameter translational searches along the X- and Y-axes relative to the Fc fragment; (b) For IgE-Fc, the C$\varepsilon 2_2$ domain pair were translated along the X-axis twice, and rotated about the X- and Y-axes relative to the Cε3 and Cε4 domains. A further X-axis rotation involving the C$\varepsilon 4_2$ 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 around 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 ± 0.3 nm from the experimental values; and (3) Models were then assessed using a goodness-of-fit $R\text{-factor} = 100^*\Sigma|I(Q)_{exp}\text{-}I(Q)_{cal})|/\Sigma|\mathrm{I}(Q)_{exp}|$ 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 $(C\varepsilon 2)_2$ 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 $(C\varepsilon 3)_2$ and $(C\varepsilon 4)_2$ by molecular graphics using the crystal structure of the corresponding four domains in IgG-Fc. The solution structure of the $(C\varepsilon 2)_2$ domain pair relative to those of the $(C\varepsilon 3)_2$ and $(C\varepsilon 4)_2$ 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.

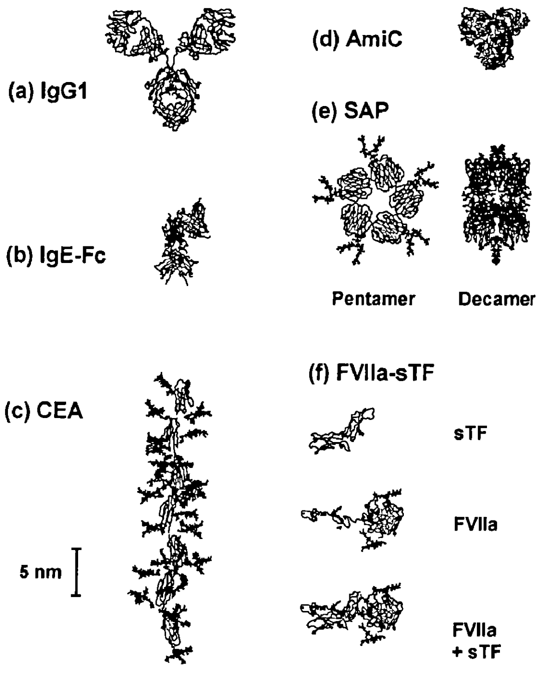


Fig. 3. The best-fit model from each curve fitting analysis to follow that of Fig. 2. The protein structure is denoted by an α-carbon trace, while oligosaccharides are shown in full if present.

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 Cε3 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 Cε3 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 Cε2 Y-axis rotation of 70° and an unchanged Cε4 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 Cε2–Cε3 bend angle and ±50° for the Cε3–Cε4 bend angle. In conclusion, the modelling showed that IgE-Fc must adopt a bent structure at either the Cε2–Cε3 or Cε3–Cε4 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,

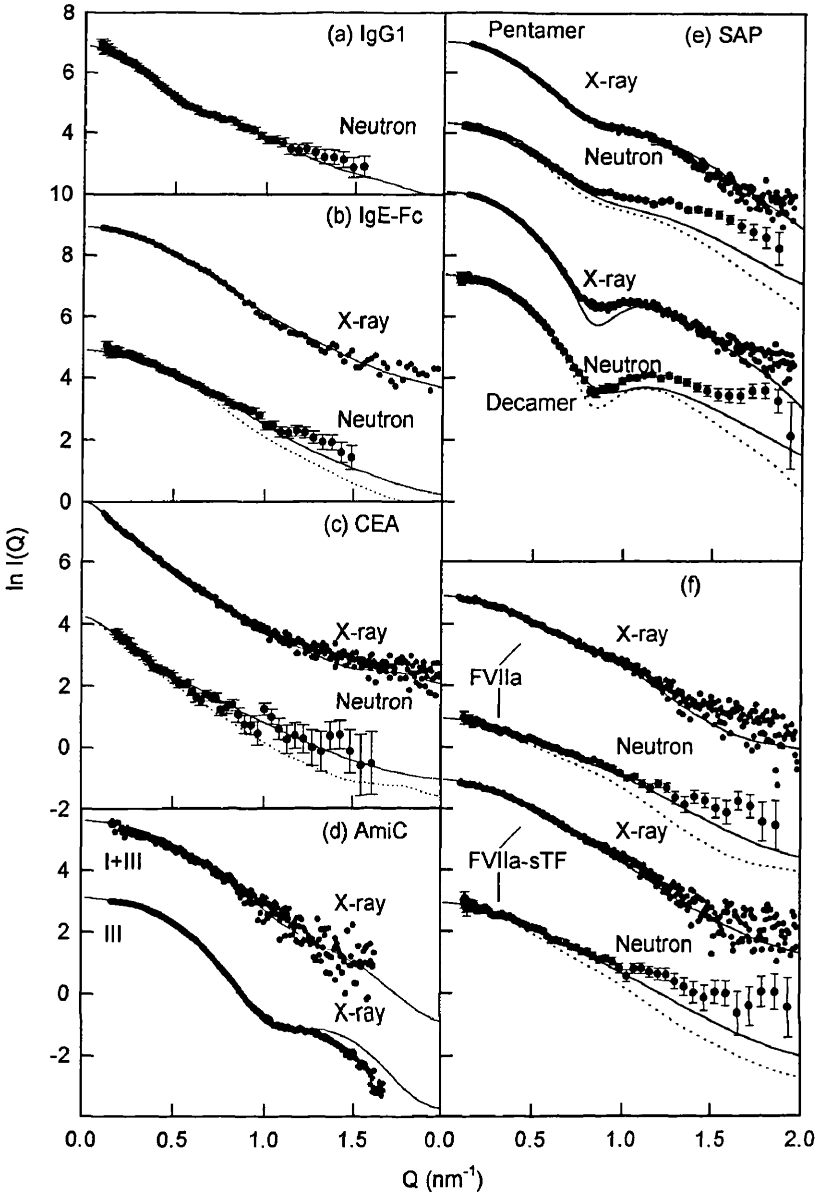


Fig. 4. Final X-ray and neutron curve fits based on the best-fit models from Fig. 3. The X-ray data were obtained from Stations 8.2 for IgE-Fc and CEA, and from Station 2.1 for AmiC, SAP, FVIIa and the FVIIa-sTF complex. Neutron data using 100% ^2H_2O buffer systems were obtained from LOQ. The continuous 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 $^1H-^2H$ 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. Bochemistry 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.