PHD

Molecular modelling of antibody combining sites

Pedersen, Jan T.

Award date:
1993

Awarding institution:
University of Bath

Link to publication

MOLECULAR MODELLING OF ANTIBODY COMBINING SITES

Submitted by Jan T Pedersen
for the degree of
Doctor of Philosophy
Department of Biochemistry
of the University of Bath
1993

Abstract

Molecular Modelling of Antibody Combining Sites

Jan T. Pedersen

Ph.D Thesis
April 1993

Two of the main problems in protein engineering today are the understanding of protein folding and molecular recognition. Both of these problems are embodied in the molecular structure of antibodies. The prediction of antibody variable region structures and the understanding of their function is the aim of this thesis.

A fully automated antibody modelling protocol which includes the automatic generation of framework regions, using a light chain and heavy chain variable variable region docking algorithm based on known variable region β-barrel structures is presented. This framework generation protocol gives good correlation with crystal structures (root mean square deviation values between 0.3-0.8 Å). A new method of sidechain generation has been developed, using a Monte Carlo simulated annealing protocol which includes a screening procedure based on hydrophobicity and accessibility for selection of the final conformation. With this sidechain generation algorithm sidechain conformations of surface located residues have a good correspondence with those of crystal structures. The complete modelling protocol has been implemented in the program A*b*M.

The usage of both sequence and structural data from antibodies within A*b*M is demonstrated by the development of a new method for reshaping (humanising) murine F_V sequences, resulting from an analysis of surface located residues in the framework regions of all known F_V crystal structures. An antibody has been reshaped using this protocol, termed *resurfacing*, which retains binding with a dissociation constant of $10^{-10}M^{-1}$.

Finally a method for the *ab-initio* design of antibody combining sites is presented. The design process is based on the hypothesis that, for small molecules, antigen binding is accounted for by sidechains of the antibody interacting with the antigen, thus being independent of the backbone conformation. For a given residue position all the possible conformations of 19 different residue types are generated. The sidechain generation algorithm uses a recursive torsional grid search, evaluating each of the generated sidechain conformations with a simple potential energy function. Each of the conformations generated is screened for exclusion of antigen surface area. This protocol results in a antibody combining site where electrostatic interactions and packing of the antigen are satisfied. Subsequent minimisation using a full potential energy function does not change the conformation of the combining site construct. A specific design, using morphine as the antigen, has been generated and is currently undergoing experimental test.

Acknowledgements

This is the place to thank all the people and institutions which have made this thesis a reality.

I would first of all like to thank Professor Anthony R. Rees, our esteemed leader, and my supervisor for three interesting years. Starting with a year of changes moving form the Laboratory of Molecular Biophysics in Oxford to the Depatement of Biochemistry at the University of Bath. I am grateful to Andrew C.R. Martin for spending a lot of time with me in Oxford, and getting me off to a good start.

Thanks to all the people who made the move to Bath and all the people who joined the Rees Group later, Robert Greist, David Webster, David Staunton, Alison Jones, Andrew Henry, Graham Elliott, Geoffrey Guy. A special thanks to Stephen M.J.Searle (the wiz-kid) for some good discussions and arguments about A*b*M CONGEN and the contortions of C and UNIX.

To Pnina Dauber-Osguthorpe and the master of the operating systems David Osguthorpe I give my special thanks, for letting me assimilate into the Molecular Graphics group and teaching me about Molecular Mechanics calculations, and for constructive critical reading of this manuscript.

None of this work would have been possible without financial support from two

Danish funding agencies: Carlsberg Research Foundation and the Danish Research Academy. I should also thank my friends (Ole Hvilsted and Leif Nørskov) at Novo-Nordisk for support and for donating a NeXT workstation, on which many of the calculations were performed and this thesis was prepared.

Finally I thank Mette for supporting me during the course of this thesis, and for allowing me to shout at her during stressful times.

Some of the work presented in this thesis has been presented or is being submitted elsewhere:

D.S. Gregory, D. Staunton, A.C.R. Martin, J. Cheetham, J. Pedersen, A.R. Rees. (1990) Antibody-combining sites - prediction and design. *Biochemical society symposium* Volume 57, Number 147, Pages 57

A.R. Rees, D. Staunton, D.S. Gregory, J. Pedersen, A. Jones, K. Hilyard, S. Roberts. (1992) Antibody Combining Sites: Prediction and Design. *Abstracts of Papers of the American Chemical Society* Volume 202 (Aug), Pages 56

V.B. Cockcroft, J.T. Pedersen, G.G. Lunt, D. Osguthorpe. (1991) BIOSITE: a program for the interactive comparison of aligned homologous protein sequences. *CABIOS* Volume 8.1, Pages 71-73

J.T. Pedersen, A.H. Henry, S.M.J. Searle, B.C. Guild, M. Roguska, and A.R. Rees. (1993) Comparison of Surface Accessible Residues in Human and Murine Immunoglobulin F_V domains: Implication for humanisation of murine antibodies. *Submitted to Journal of Molecular Biology*

J.T. Pedersen, R.R. Campbell, C.C. Carter, A.C.R. Martin, D. Rose, F. Ruker, R.K. Strong, X. He, A.R. Rees. (1992) Modelling Antibody Combining Sites : A method for prediction of the entire variable domain structure. *Document in preparation*

J.T. Pedersen, A.R. Rees. (1992) Antibodies can be reationally engineered ?. *Presented at: An International meeting of the Bichemical Society & the Royal Society of Chemistry: Engeneering Antibodies for Therapy 10th-11th of September*

J.T. Pedersen, S.M.J. Searle, A.H. Henry, A.R. Rees. (1992) Antibody Modelling: Beyond Homology. *Immunomethods* Volume 1, Number 2, Pages 126-136

A.H. Henry, J.T. Pedersen, S.M.J. Searle, B. Guild, M. Roguska, A.R. Rees. (1993) Rational reshaping of a mouse antibody *Protein Engineering* Volume 6 supp 1993, Pages 89

M.A. Roguska, J.T. Pedersen, C.A. Keddy, A.H. Henry, S.M.J. Searle, J.M. Lambert, C.A. Vater, W.A. Blattler, A.R. Rees and B.C. Guild. (1993) Humanisation of murine monoclonal antibodies through variable domain resurfacing. *Submitted to Nature*

Glossary

A*b*M Anti*b*ody Modelling algorithm.

ACS Antibody Combining Site.

Antigen Antibody binding species.

APC Antigen Presenting Cell.

C domain Constant domain.

CAMAL Combined Algorithm for Modelling Antibody Loops

CDR Complementarity Determining Region.

D segment Immunoglobulin gene Diversity segment.

dAb Single domain Antibody.

F_{AB} Antibody sub fraction consisting of *V* domain and *C* domain.

FACS Fluorescence Activated Cell Sorter

F_V Variable region of antibody, consisting of V_L and V_H domains.

F_C Constant or effector region of antibodies.

HMC Hybrid Monte Carlo simulation.

Hapten Small antigen ($M_R < 5000$).

HFR1,2,3 and 4 Heavy chain framework regions.

IgA,IgD,IgE,IgG,IgM Immunoglobulin classes.

J segment Immunoglobulin gene Joining segment.

LFR1,2,3 and 4 Light chain framework regions.

MAC Membrane Attack Complex.

MC Monte Carlo simulation.

MD Molecular Dynamics.

MHC-I,MHC-II Major Histocompatibility Complex I and II.

MM Molecular Mechanics.

MOP Maximum Overlap Procedure

MRU Minimal recognition unit

NK Natural Killer Cell

QSAR Quantitative Structure Activity Relationship

scFv Single chain F_V

SDR Structurally Determining Residue

TCR T-cell receptor

V domain Variable domain.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The prediction of protein three dimensional structure from sequence is one of the holy grails of biology. During the last twenty years it has become possible to predict the conformations of small parts of proteins when the surrounding structure is known, particularly if there already exists a family of homologous proteins where structures have been solved experimentally. This field is termed **Molecular Modelling**. The aim of this thesis is to investigate the possibilities for design and prediction of antibody structures, using the methods of molecular modelling.

This introduction will give the reader a setting for the objectives of this thesis. It will contain a basic overview of immunology, the structure of immunoglobulin superfamily proteins and an introduction to homology modelling and molecular mechanics methods that are fundamental to some of the work in the thesis. The background to antibody design is introduced separately at the beginning of Chapter 4.

1.1 The immune system

A healthy animal has several ways of defending itself against infections. First there are physicochemical barriers such as skin and mucous membranes. Second, there is a system of phagocytic cells, macrophages, natural killer cells (**NK**) and lymphocytes. Third, there exists an extensive range of blood borne molecules such as antibodies, complement, cytokines and interferons.

Some defense mechanisms are present prior to infection and are not influenced or regulated by such infections. These factors constitute what is called **natural immunity** (also called native or innate immunity). Other defense mechanisms are activated when exposure to infection occurs, and are controlled by the amount of foreign substance present. These factors constitute what is called **specific** or **acquired immunity**.

The basis of **natural immunity** can be defined by three processes. The first is the inclusion of foreign particles in neutrophils and macrophages, by a process of phagocytosis in which the particles are enclosed in a phagosome. The phagosomes then fuse inside the cell with granules containing harsh reagents such as superoxide anions, hydroxyl radicals, halide ions, and a range of proteolytic enzymes such as Cathepsin G, lysozyme, defensin etc., which will degrade any biological material. Second, the **complement** system, which is a cascade of two converging pathways of serum and membrane bound enzymes, is activated. All the components of the pathway interact in a highly regulated manner. One branch of the cascade is activated by antibody-antigen complexes and the other by direct contact with surfaces of foreign material. Both pathways lead to the activation of a final pathway which generates the **membrane attack complex** (MAC). MAC is capable of breaking down cell membranes by self insertion. Third, NK cells capable of recognising foreign cells bind to the cell surface and release the protein perforin into the inter-cell space. Perforin is then inserted into the foreign membrane, and pores are generated leading to cell death. This action is similar to the mechanism of MAC in the complement pathway.

Specific immunity is a type of immunity found predominantly in higher animals. It is a type of immunity which arises as the result of exposure to a foreign compound (**antigen**). This process is called **immunisation.** There are two classes of specific immunity. First, **Humoral immunity** which can be transferred to other individuals via cell free portions of blood (serum or plasma). This type of immunity is mediated by molecules in the blood which are specific to antigens, called **antibodies** or **immunoglobulins**. Antibodies are produced by a type of blood cells called **B-lymphocytes** (or B-cells). Second, there is **Cell-mediated immunity**, which can be transferred to other animals with cells from immunised individuals, but not with plasma or serum. This type of immunity is mediated by a second class of lymphocytes known as **T-lymphocytes** (or T-cells) that recognise specific antigens on the surface of foreign cells.

There are three phases in the immune response: 1) **Cognitive phase**, in which recognition of the antigen takes place, 2) **Activation phase**, in which specific lymphocytes are triggered and 3) **Effector phase**, in which the antigen is eliminated. The processes of the immune response are outlined in detail in Figure 1.1.

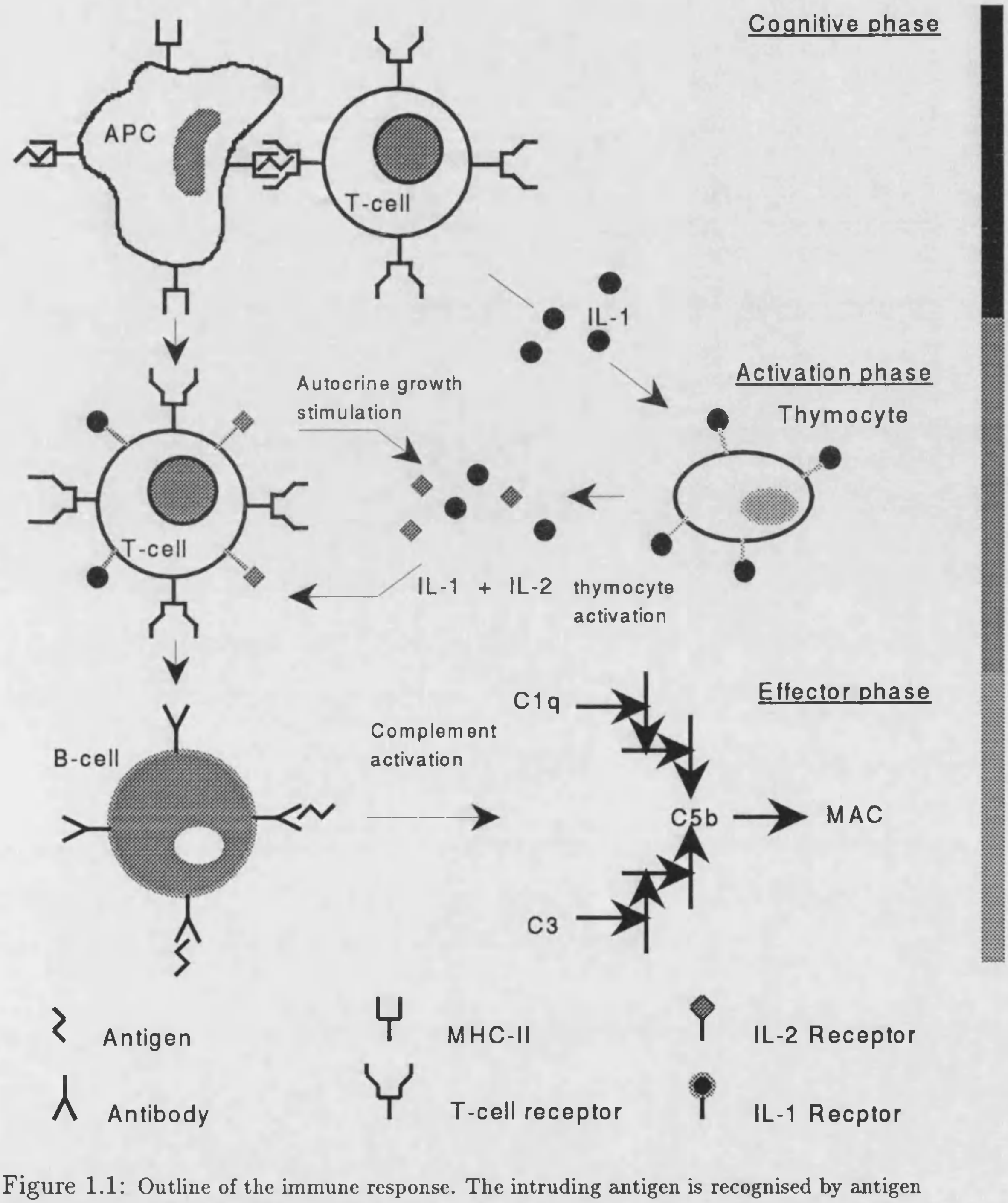


Figure 1.1: Outline of the immune response. The intruding antigen is recognised by antigen presenting cells (**APC**'s) (macrophages etc), degraded and presented on the cell surface via class II major histocompatibility complex (MHC II). The presented antigen is recognised by CD4$^+$T cells (T helper cells), which start the production of lymphokine interleukin-1 (IL-1). IL-1 stimulates thymocytes to produce both IL-1 and IL-2. IL-1 and IL-2 simulate the proliferation of CD4$^+$T-helper cells, other T-cells, thymocytes, and B-cells. The proliferation of B-cells producing antigen specific antibodies is enhanced. This membrane bound antibody in turn activates the first protein in the complement cascade, leading to neutralisation of the antigen.

1.2 The immunoglobulin protein superfamily

The Immunoglobulin superfamily of proteins is one of the best described and characterised families of proteins to date. There are approximately 50 (Abbas *et al.*, 1991) different sub-families within the superfamily all which are encoded by independent gene complexes. Common to all these proteins is the structural motif, the β-sheet sandwich which, when present in this superfamily is also known as the **Immunoglobulin fold** or **domain**, see Figure 1.2. The structure of Ig domains has been reviewed by (Amzel and Poljak, 1979), and are classified as either variable (V-type), constant (C-type) or primitive (P-type) domains. All the proteins of the family are constructed by one or several units of this motif linked together. Figure 1.3 shows examples of some of these structures.

1.3 Immunoglobulin structure

Clues to the antibody or immunoglobulin structure were first discovered by Porter by performing proteolytic digestion of antibody isolates (Porter, 1958; Porter, 1959). It was determined that all antibody structures have the same overall structure, consisting of four chains: two identical light chains, of molecular weight 24 kilodaltons (kD), and two identical heavy chains of about 55 or 70 kD depending on the antibody class. The fragments isolated by proteolytic digestion were called F_C (crystallisable fraction), and F_{AB} (antigen binding). Later F_V (variable domain) was obtained by cleavage of F_{AB}. The structural significance of these fragments is outlined in Figure 1.2.

Although all antibodies are similar they can be subdivided into classes called

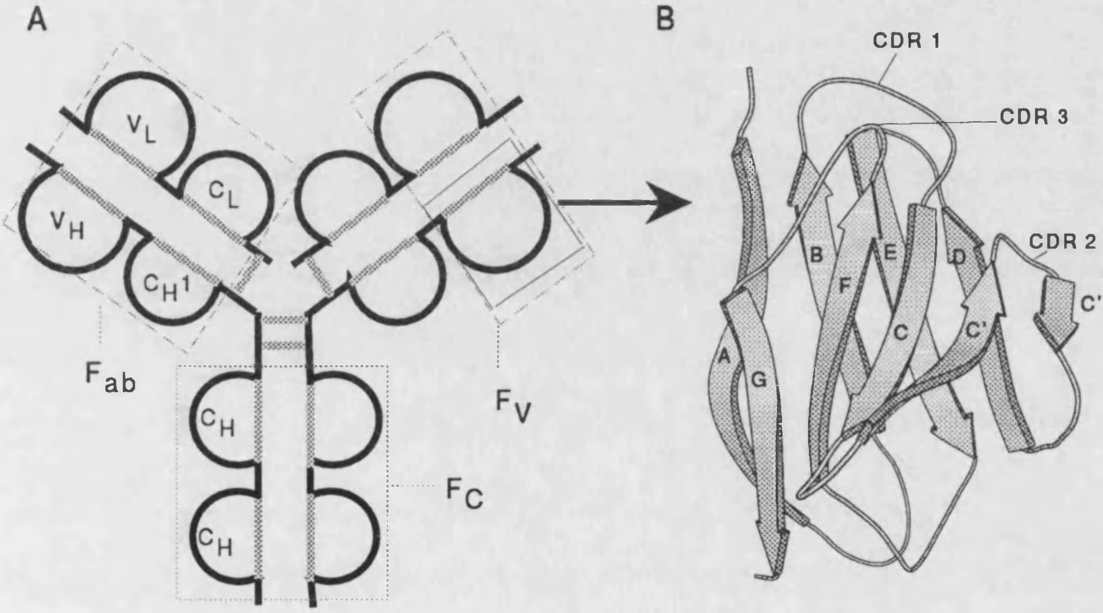


Figure 1.2: Outline of antibody structure (IgG). A) The overall domain composition of the antibody. B) the β-sheet sandwich, the building block of the immunoglobulin superfamily. Each of the two halfs of the antibody are identical, consisting of one light and one heavy chain. Each loop in Figure A) is equivalent to an Ig-domain β-sheet shown in Figure B (indicated by an arrow). The naming of the strands is shown on the figure. Strands **C'** and **C"** are found in the **V** domain but not the **C** domain of an antibody. The boxes enclosing different parts of the antibody show the fragments obtained by proteolytic digestion (explained in text). Grey bonds in A indicate disulphide links.

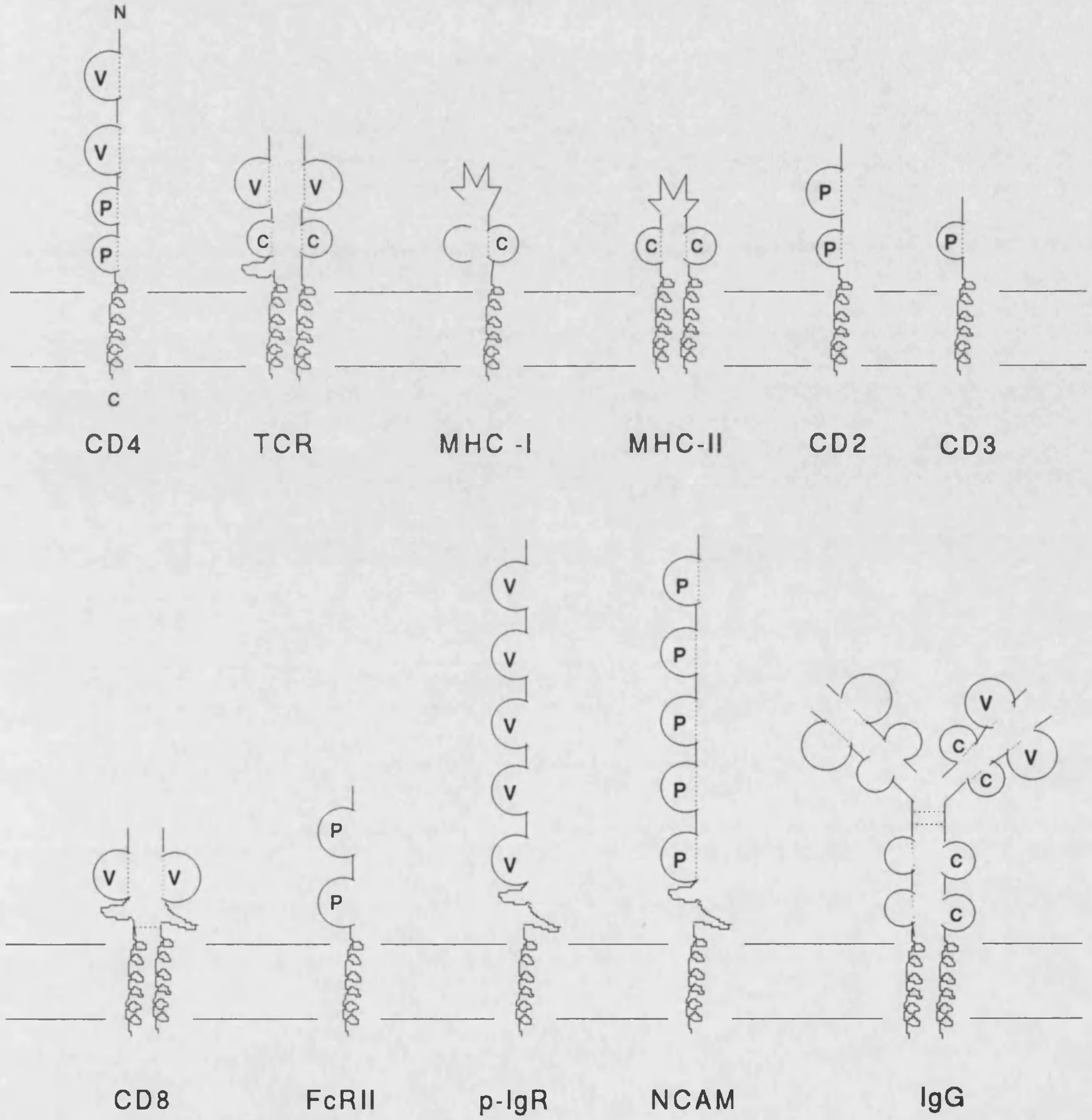


Figure 1.3: Some immunoglobulin superfamily proteins - all presented on the cell surface. The basic domain building block is the β-**sheet sandwich**. V: Variable type domain, C: Constant domain, P: Distantly related Ig-Domain. Broken bonds indicate disulphide bonds. The direction of the protein chain is indicated on CD4. The proteins are: **CD2,3,4,8**: Cell Differentiation antigen 4, **TRC**: T-cell receptor, **MHC-I,II**: Major Hisocompatibility complex, **FcRII**: Fc receptor, **p-IgR**: poly-Ig recptor, **NCAM**: Neural Cell Adhesion Molecule, **IgG**: Immunoglobulin G.

Immunoglobulin type	Heavy chains	Light chains
IgA	$\alpha 1, \alpha 2$	κ, λ
IgD	δ	
IgE	ϵ	
IgG	$\gamma 1, \gamma 2, \gamma 3, \gamma 4$	
IgM	μ	

Table 1.1: Ig subtypes and chain classes, showing that the **H** chains are much more diverse than the **L** chains

IgA, IgD, IgE, IgG and **IgM**. The basis of this classification is historical, structural and physiological at the same time. IgA's form a primary defense barrier, as they are secreted through the mucus membrane. IgD is structurally identical to IgG in humans, is expressed on the B-cell surface and is thought to have a role in tolerance. IgE, which has an extra constant Ig-domain, is involved in allergy reactions, and binds to specialised cells (mast cells) that express F_C-receptors specific to IgE. The binding of IgE-allergen complexes to these F_C-receptors promotes histamine release. IgM is the first antibody to be synthesised in an immune response. Both IgM and IgG are blood borne. IgG is probably the most important of the immunoglobulins, and has the highest blood concentration. The classes IgA and IgG are further subdivided into subclasses: **IgA1, IgA2, IgG1, IgG2, IgG3** and **IgG4**. The heavy chains are classified as α(A), γ(G), δ(D), ϵ(E), or μ(M). There are two classes of light chains, κ and λ. Table 1.1 outlines the classes of antibodies and the chain denomination.

All the antibody types have the same basic **Y** or better **T** (Figure 1.2) shape, but members of the IgA class are dimers and IgMs are pentamers. In both instances the multimeric form is stabilised by an extra chain, the **J chain** (or joining chain).

The basic structure is outlined in Figure 1.2. The heavy and light chains pair, using both covalent (disulphide bonds) and non-covalent interactions (hydrophobic domain packing). The pairing of the light and heavy chain will be discussed further in Chapter 2. The $\mathbf{C_H1}$ and $\mathbf{C_L}$ domains pair as do the $\mathbf{V_H}$ and $\mathbf{V_L}$ domains. Sequence alignments show that the $\mathbf{C_L/C_H1}$ domain (**C domain**) and the F_C portion of the antibody are highly conserved, whereas the $\mathbf{V_L/V_H}$ (**V domain**) contains hypervariable (Wu and Kabat, 1970) regions. There are three hypervariable regions in each of the chains. Each region is between 3 to 20 residues long and is called a Complememmunentarity Determining Region or **CDR**. The CDR's are numbered **CDR L1, CDR L2** and **CDR L3** in the light chain and **CDR H1, CDR H2** and **CDR H3** in the heavy chain. The CDR's are exposed loop regions situated in three-dimensionally contiguous regions of the antibody, and constitute the antibody combining site (or **ACS**) of the antibody. The ACS is responsible for the recognition of antigens.

The core of the F_V domain is conserved and is termed the **framework**, and consists of a β-barrel formed by contributions from the V_L and V_H chain. The framework is described in more detail in Section 2.3.

1.4 Immunoglobulin diversity and gene organisation

One of the most intriguing question in molecular immunology today is the precise size of the **immunoglobulin repertoire**. Since each antibody is specific to a single, or very few, antigenic determinants there should exist a large number of different antibodies in an organism in order for the immune system to be able to recognise any new antigen, although the antibodies may be generated partly according to need (The Instructive Hypothesis of (Jerne, 1973)).

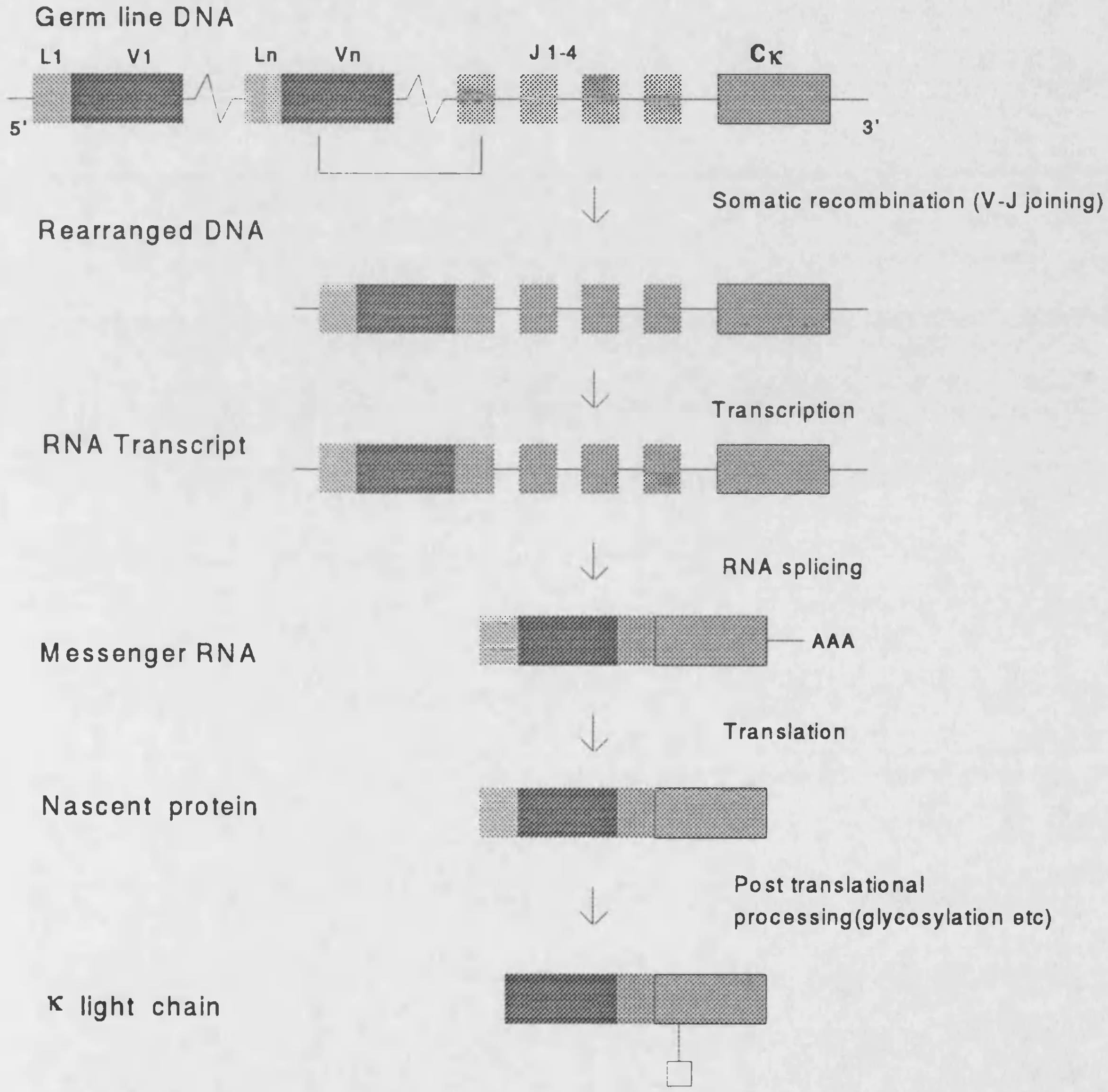


Figure 1.4: The sequence of events which lead to the generation of a mature κ-chain. Only one round of somatic recombination occurs (V-J joining), to form the rearranged light chain gene. □ indicates glycosylation site

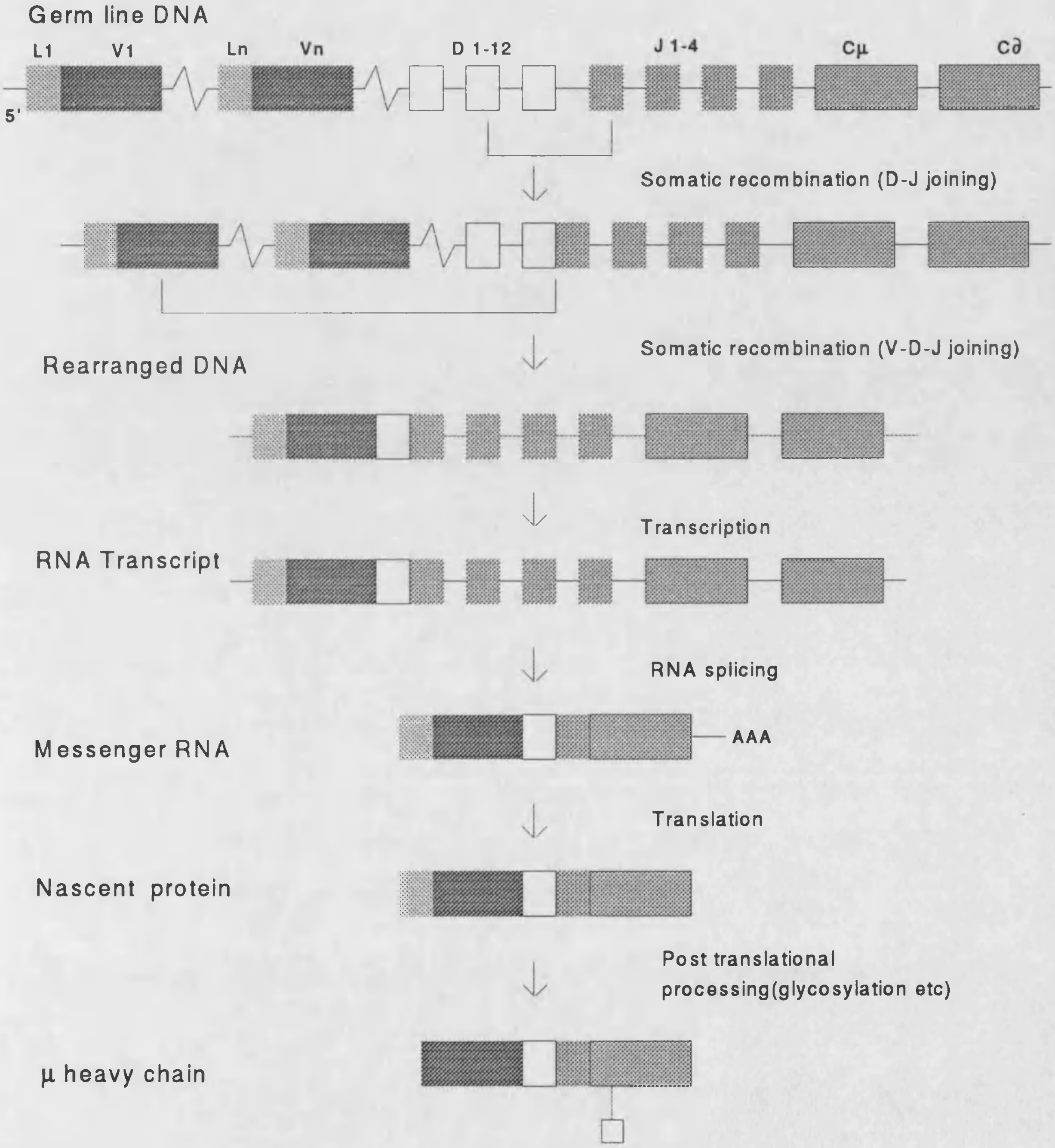


Figure 1.5: Events which lead to the generation of mature heavy chains. Two rounds of somatic recombination occurs, 1) D-J joining and 2) V-D-J joining. In the final step the leader (L) is cleaved off and the protein is glycosylated (□ indicates glycosylation site), to give the mature heavy chain.

Diversity occurs as a result of disorder in the recombination of immunoglobulin genes, and is obtained by a combination of **somatic recombination** of many germ-line genes and **somatic mutation**(Figure 1.4 and 1.5). A light chain is generated from three gene fragments **V-J-C**, and a heavy chain from four fragments **V-D-J-C**. This means (assuming approximately equal numbers of segments) that there exists a larger number of heavy chains than light chains. Figure 1.4 and 1.5 shows how heavy and light chains are generated. The murine immunoglobulin gene complex has been mapped and much of it has been sequenced. As a result of this work it is possible to estimate that the size of the antibody repertoire is of order $10^9 - 10^{11}$ different antibodies (Abbas *et al.*, 1991). Table 1.2 outlines the basis of this estimate. This is only an estimate of possible sequences and not of possible complementary shapes.

CDR's L3 and H3 are the most variable in length and sequence. This "extra" variability is obtained because these CDR gene sequences are situated right at the position where joining of either the J-segment (light chain) or J and D segments (heavy chain) occurs. The implications of this variability for modelling of antibody combining sites are addressed in Chapter 2.

1.5 Antigen recognition

The interaction of an antibody with its cognate antigen is a widely accepted paradigm of molecular recognition. To understand the antibody-antigen interaction in atomic detail requires knowledge of the three-dimensional structure of antibodies and of their antigen complexes.

The thermodynamic process of antigen binding is the result of changes in both

Diversity Factor	H chain	κ	λ
Germline Gene segments			
V	250-1000	250	2
J	4	4	3
D	12	0	0
Combinatorial joining			
V·D·(J)	10^4-4·10^4	10^3	6
H-L chain association			
H·κ	$1 - 4{\cdot}10^7$		
H·λ	$5 - 10{\cdot}10^4$		
Total potential repertoire	$10^9 - 10^{11}$		

Table 1.2: Simple calculation of the size of the immunoglobulin repertoire. Note that there are approximately ten times more possible heavy chains than light chains. Table reproduced after (Abbas *et al.*, 1991)

enthalpy and entropy of the system. The entropic changes arise from changes in entropy of water on exclusion from the binding site, and loss of motional entropy of both antibody and antigen on binding. The enthalpic changes involve complex exchange of H-bonds, charge-charge interactions and van der Waals interactions. The binding of antigen is believed to be a diffusion controlled process, characterised by second-order rate constants, with k_2 values in the range $0.6 - 1.0{\cdot}10^6 M^{-1}s^{-1}$. These rate constants are slow when compared to enzymatic reactions which have k_2 values in the order $10^7 - 10^8 M^{-1}s^{-1}$ (Northrup and Erickson, 1992).

1.5.1 The antigen

An **antigen** is defined as a substance which may be specifically bound to an antibody. Antigens which are capable of eliciting an immune response are called **immunogens**.

Small molecules (M_R < 5000 kDa) are generally unable to generate an immune response unless bound to a larger **carrier** molecule or unless they can react as superantigens (independent of MHC processing). Small antigens are called **haptens**. Where the antigen is macromolecular and larger than the ACS, the antibody only binds to a part of the macromolecule called a **determinant** or **epitope**.

There is some controversy about the origin of antigenicity. Early work by Atassi *et al* indicated that the antigenic profile of a molecule is defined by very few, well defined epitopes (Atassi, 1975; Atassi, 1978). A later review of a larger body of work by Benjamin *et al* showed that any region of the surface of a macromolecule can be a potential antigenic epitope (Benjamin *et al.*, 1984). However, the capacity of a given individual to respond to any particular epitope is determined by the regulatory processes of the immune system operating in that individual. Despite numerous debates (Tainer *et al.*, 1985) it is still not clear what influence the flexibility of the antigen has on the capacity to trigger an immune response.

1.5.2 Antibody types

Antibodies can be classified according to the topology of the antigens which they recognise (Wang *et al.*, 1991) (Figure 1.6). There are three groups which in this thesis will be called : 1) **cavity antibodies**, 2) **groove antibodies**, and 3) **planar antibodies**. This classification describes the overall topography of the ACS. The classification is based on 20 x-ray crystallographic structures of antibody $\mathbf{F}_{\mathrm{AB}}$ fragments some of which have an antigen bound (Wang *et al.*, 1991).

It is not clear whether any of these combining site types are preferred by particular

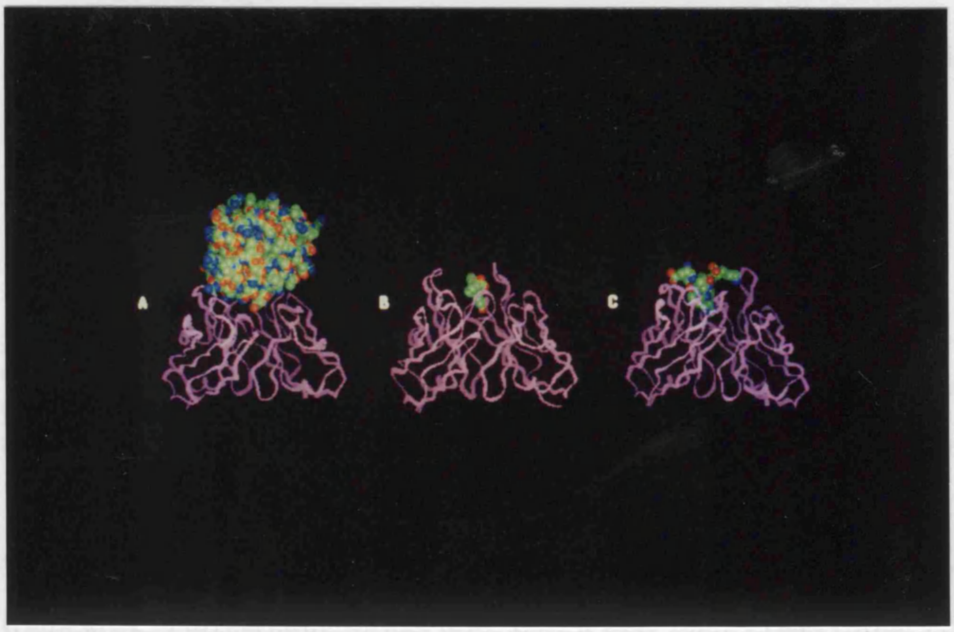


Figure 1.6: The three antibody binding site types, exemplified by: a) Planar: Hy-Hel-10 (Padlan *et al.*, 1989) b) Cavity: 4-4-20 (Herron *et al.*, 1989), c) Groove: B13I2 (Stanfield *et al.*, 1990). Classification as is outlined by Wang *et al* (Wang *et al.*, 1991) In this picture the F_V framework is shown as a magenta ribbon, and the antigen is shown in CPK representation.

types of antigens. There are however some indications (Rini *et al.*, 1992; Novotny, 1991; Herron *et al.*, 1989) that smaller antigens bind best when they are almost buried in the surface of of the combining site, usually in a hydrophobic hole. Larger protein antigens prefer less curved surfaces and appear to bind over a larger surface area (Amit *et al.*, 1986; Sheriff *et al.*, 1987), often with many charge-charge interactions.

1.5.3 CDR sidechains

The recognition of an antigen is largely mediated by the exposed sidechains in the CDR loops. Several studies (Padlan, 1990; Mian *et al.*, 1991; Kabat and Wu, 1971) of the amino acid distribution and the accessibility of sidechains in

Residue group	Residues	Specificity (H-bonding)	Binding (Hydrophobic effect)
Aliphatic	Ile **Leu** Val Ala		+
O and S functional	Cys **Ser** Met Thr	++	
Acidic	**Asp** Glu	+++	
Basic	Lys **Arg**	+++	+
Bifunctional	**Asn** **Tyr** Gln His	++	++
Aromatic	**Phe** Trp		+++
Structural	**Gly** Pro		

Table 1.3: The function of various sidechain groups in the ACS, (+) signs indicate the influence of a residue on a particular effect. The most abundant residues occurring in CDRs as determined from the database of Kabat *et al* (Kabat *et al.*, 1992), are outlined in bold.

the CDRs has shown that, overall, sidechains are more exposed than those in the framework. Furthermore, it has been shown that there exists (Mian *et al.*, 1991) a higher frequency of bifunctional residues such as Tyr, His, and Asn which are capable of engaging in hydrogen bonding and of contributing to the hydrophobic effect in the binding loops. The usage of bifunctional residues is yet another way of broadening specificity of the antibody. Table 1.3 outlines the function of amino acid types as compiled by Padlan and Mian (Padlan, 1990; Mian *et al.*, 1991).

There is also an unusually high frequency of exposed hydrophobic residues. It can therefore be speculated (Colman, 1988) that the hydrophobic sidechains account for **binding**, and the charged sidechains for **specificity** in the process of antigen recognition. Frequency tables of various residue types within the framework and CDR's can be found in Appendix A.

1.6 Homology modelling

When the first x-ray crystallographic structures of homologous proteins emerged it was discovered that proteins which have homologous sequences also have similar folds. This provided the basis for performing **Homology Modeling** (Browne *et al.*, 1969) or folding of a sequence over a known three dimensional structure. Since these first modeling experiments this method of generating three-dimensional structures has been used on a large variety of protein families (for some of the latest examples see (Bank *et al.*, 1990; Dalgleish *et al.*, 1992; Greer, 1990; James *et al.*, 1991; Mas *et al.*, 1992; Mosimann *et al.*, 1992; Weber, 1990)).

When modeling homologous families of proteins the first step is to align the known sequences of the family. This will usually result in an optimised alignment, containing gaps in the sequences (Needleman and Wunsch, 1970). Similarly, if there are more three-dimensional structures known within the family these have to be superimposed (McLachlan, 1979). From the alignment and the superimposition **Structurally Conserved Regions** (SCR's) and **Variable Regions** (VR's) can be assigned. SCR's are usually located in the protein core, whereas VR's are surface located. This distribution of SCR's and VR's also reflects the accuracy with which a new sequence can be modelled (Greer, 1991). The protein core can be predicted with high confidence and the conformation of VR's are predicted with lower confidence. The reason for this lower confidence when modelling VR's is that these are usually located in regions with less well defined secondary structure (loops) on the protein surface. This confidence problem has also been the main obstacle when attempting to model antibody F_V structures.

Many algorithms have been developed to automate and improve the accuracy of models obtained from homology modelling. The difference between these methods lies largely in the way they model VR's. The VR modelling methods fall into two groups : Database methods (Sutcliffe *et al.*, 1987b; Sutcliffe *et al.*, 1987a; Jones and Thirup, 1986; Martin *et al.*, 1989; Martin *et al.*, 1991a) and *ab initio* methods (Palmer and Sheraga, 1991; Bruccoleri and Karplus, 1987; Moult and James, 1986; Havel and Snow, 1991).

The application of these methods to antibody F_V structures is reviewed in Chapter 2. For other reviews of these methods see (Greer, 1991; Maggiora *et al.*, 1991).

1.7 Molecular mechanics calculations

The objective of any modelling study is to obtain insight into molecular structure and function. Almost any modelling procedure does contain a stage of objective evaluation of the model produced, and this is most frequently done using potential energy functions. In this thesis several of these Molecular Mechanics (MM) packages are made use of, such as VFF (Valence Force Field) (Lifson *et al.*, 1979), DISCOVER (b), CHARMM (Brooks *et al.*, 1983), and a Monte Carlo/Metropolis (Metropolis *et al.*, 1953) package written by the author.

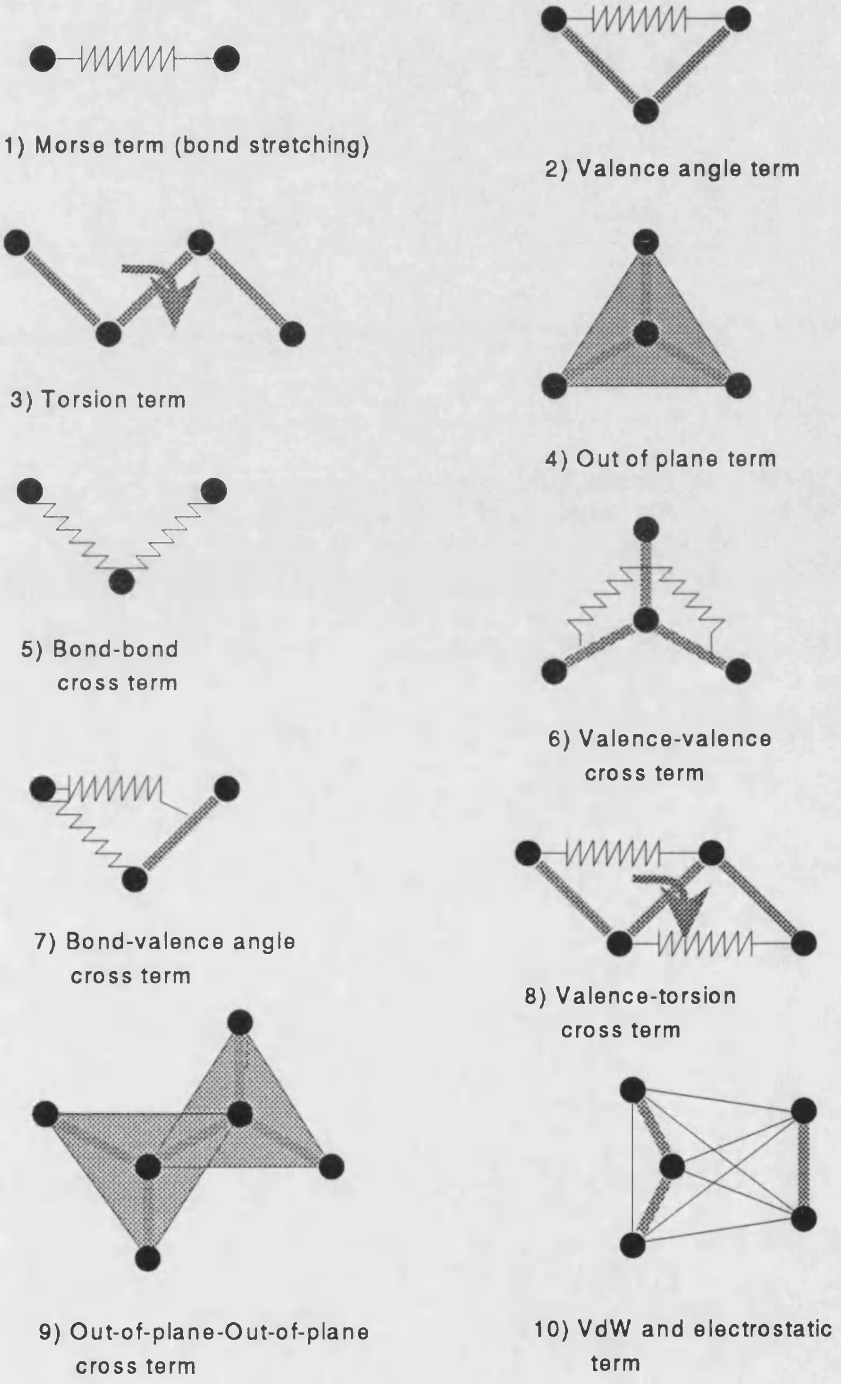
A prerequisite for any MM calculation is a potential function defining the energy of a molecular system as a function of atomic position. In principle an exact solution to this problem can be obtained by solving the quantum mechanical equations which describe the ground state energy of the electrons and nuclei at each possible nuclear position. The resulting energies form a continuous **Born-Oppenheimer surface** (McCammon and Harvey, 1987) as a function of nuclear position. The surface describes the energy of virtually any type of atomic motion in molecular systems (McCammon and Harvey, 1987). Unfortunately quantum mechanical descriptions of systems the size of proteins (thousands of atoms) is not yet possible. Therefore there is a need to derive simple empirical energy functions of the atomic positions.

The global energy functions used are functions which are a sum of several terms, each describing a single atomic or molecular force. The model used in VFF (Dauber-Osguthorpe *et al.*, 1991) is outlined in equation 1.1.

$$\begin{aligned}
V &= \sum[D_b(1-\exp^{-\alpha(b-b_0)})^2 - D_b] + \\
&\frac{1}{2}\sum H_\theta(\theta-\theta_0)^2 + \\
&\frac{1}{2}\sum H_\phi(1-s\cos(n\phi)) + \\
&\frac{1}{2}\sum H_\chi \chi^2 + \\
&\sum\sum F_{bb'}(b-b_0)(b'-b'_0) + \\
&\sum\sum F_{\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0) + \\
&\sum\sum F_{b\theta}(b-b_0)(\theta-\theta_0) + \\
&\sum F_{\phi\theta\theta'}\cos\phi(\theta-\theta_0)(\theta'-\theta'_0) + \\
&\sum\sum F_{\chi\chi'}\chi\chi' + \\
&\sum\varepsilon[(\frac{r^\star}{r})^{12} - 2(\frac{r^\star}{r})^6] + \sum\frac{q_i q_j}{r} \qquad (1.1)
\end{aligned}$$

Here, b is bond length, θ is valence angle, ϕ is torsion angle and χ out of plane angles. r is the distance between atoms, q partial atomic charges and ϵ is the energy of interaction at the most favorable interaction distance $r^\star$. H,F and D are force constants.

In this force field the first four terms describe the energy required to distort internal bonds, valence angles, torsion angles, and out of plane angles. The next five terms describe relations between the first four terms, and are called cross terms. The last term describe the relation between non-bonded atoms. The physical meaning of these terms is illustrated in Figure 1.7.



1) Morse term (bond stretching)

2) Valence angle term

3) Torsion term

4) Out of plane term

5) Bond-bond cross term

6) Valence-valence cross term

7) Bond-valence angle cross term

8) Valence-torsion cross term

9) Out-of-plane-Out-of-plane cross term

10) VdW and electrostatic term

Figure 1.7: Pictorial description of the force field outlined in Equation 1.1, the ten terms correspond to the ten terms in the equation. Reproduced after INSIGHT manual (TM Biosym Inc., San Diego, CA)

The constants in the system are then derived by fitting the function to experimental data, or quantum mechanical calculations on small molecules. Forcefields are frequently refined by adding terms which better describe specific phenomena observed in molecular structures, such as hydrogen bonds etc (Dauber-Osguthorpe *et al.*, 1991).

Three main molecular mechanics methods are used in this work: **molecular dynamics** (MD), **minimisation**, and **Monte Carlo simulation** (MC). The MC simulation method is discussed in further detail in Section 2.5.

1.7.1 Molecular dynamics

The aim of MD is to simulate the motions of molecules, using the basic Newtonian equations of motion (Newton, 1729 (1960)). In this description the atom i is assumed to be a singular point, with the mass m_i. If the position of the particle is called r_i, the velocity is given by the first derivative of position with respect to time (∂t):

$$v_i = \frac{\partial r_i}{\partial t} = \frac{p_i}{m_i} \tag{1.2}$$

Where p_i is the momentum of the particle. The net force exerted on the particle is given by:

$$F_i = \frac{\partial p_i}{\partial t} = -\frac{\partial V}{\partial r_i} \tag{1.3}$$

and James, 1986; Havel and Snow, 1991).

The application of these methods to antibody F_V structures is reviewed in Chapter 2. For other reviews of these methods see (Greer, 1991; Maggiora *et al.*, 1991).

1.7 Molecular mechanics calculations

The objective of any modelling study is to obtain insight into molecular structure and function. Almost any modelling procedure does contain a stage of objective evaluation of the model produced, and this is most frequently done using potential energy functions. In this thesis several of these Molecular Mechanics (MM) packages are made use of, such as VFF (Valence Force Field) (Lifson *et al.*, 1979), DISCOVER (b), CHARMM (Brooks *et al.*, 1983), and a Monte Carlo/Metropolis (Metropolis *et al.*, 1953) package written by the author.

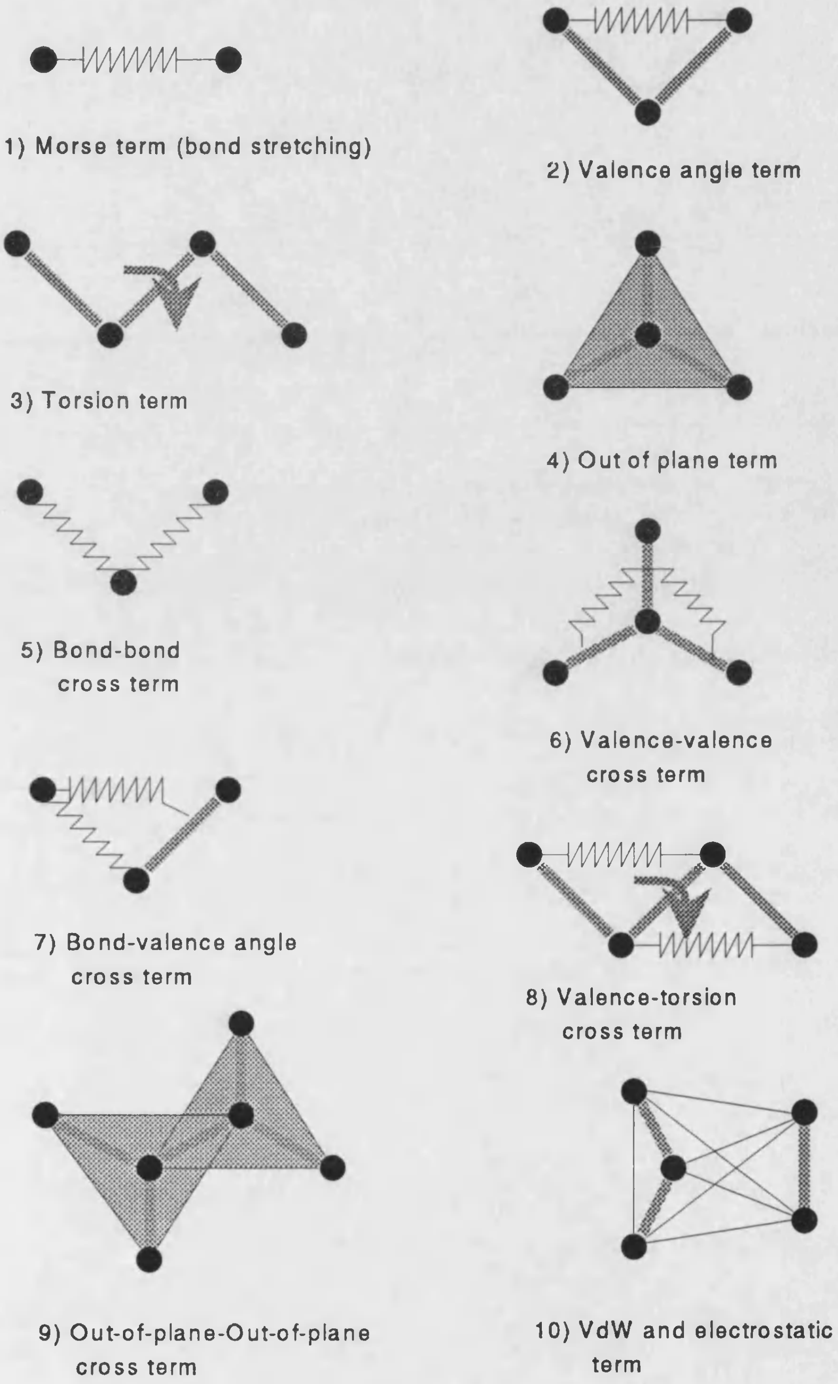
A prerequisite for any MM calculation is a potential function defining the energy of a molecular system as a function of atomic position. In principle an exact solution to this problem can be obtained by solving the quantum mechanical equations which describe the ground state energy of the electrons and nuclei at each possible nuclear position. The resulting energies form a continuous **Born-Oppenheimer surface** (McCammon and Harvey, 1987) as a function of nuclear position. The surface describes the energy of virtually any type of atomic motion in molecular systems (McCammon and Harvey, 1987). Unfortunately quantum mechanical descriptions of systems the size of proteins (thousands of atoms) is not yet possible. Therefore there is a need to derive simple empirical energy functions of the atomic positions.

The global energy functions used are functions which are a sum of several terms, each describing a single atomic or molecular force. The model used in VFF (Dauber-Osguthorpe *et al.*, 1991) is outlined in equation 1.1.

$$
\begin{aligned}
V = {} & \sum[D_b(1-\exp^{-\alpha(b-b_0)})^2 - D_b] + \\
& \frac{1}{2}\sum H_\theta(\theta-\theta_0)^2 + \\
& \frac{1}{2}\sum H_\phi(1-s\cos(n\phi)) + \\
& \frac{1}{2}\sum H_\chi\chi^2 + \\
& \sum\sum F_{bb'}(b-b_0)(b'-b'_0) + \\
& \sum\sum F_{\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0) + \\
& \sum\sum F_{b\theta}(b-b_0)(\theta-\theta_0) + \\
& \sum F_{\phi\theta\theta'}\cos\phi(\theta-\theta_0)(\theta'-\theta'_0) + \\
& \sum\sum F_{\chi\chi'}\chi\chi' + \\
& \sum\varepsilon[(\frac{r^\star}{r})^{12} - 2(\frac{r^\star}{r})^{6}] + \sum\frac{q_iq_j}{r}
\end{aligned}
\tag{1.1}
$$

Here, b is bond length, θ is valence angle, ϕ is torsion angle and χ out of plane angles. r is the distance between atoms, q partial atomic charges and ϵ is the energy of interaction at the most favorable interaction distance $r^\star$. H,F and D are force constants.

In this force field the first four terms describe the energy required to distort internal bonds, valence angles, torsion angles, and out of plane angles. The next five terms describe relations between the first four terms, and are called cross terms. The last term describe the relation between non-bonded atoms. The physical meaning of these terms is illustrated in Figure 1.7.



1) Morse term (bond stretching)

2) Valence angle term

3) Torsion term

4) Out of plane term

5) Bond-bond cross term

6) Valence-valence cross term

7) Bond-valence angle cross term

8) Valence-torsion cross term

9) Out-of-plane-Out-of-plane cross term

10) VdW and electrostatic term

Figure 1.7: Pictorial description of the force field outlined in Equation 1.1, the ten terms correspond to the ten terms in the equation. Reproduced after INSIGHT manual (TM Biosym Inc., San Diego, CA)

The constants in the system are then derived by fitting the function to experimental data, or quantum mechanical calculations on small molecules. Forcefields are frequently refined by adding terms which better describe specific phenomena observed in molecular structures, such as hydrogen bonds etc (Dauber-Osguthorpe *et al.*, 1991).

Three main molecular mechanics methods are used in this work: **molecular dynamics** (MD), **minimisation**, and **Monte Carlo simulation** (MC). The MC simulation method is discussed in further detail in Section 2.5.

1.7.1 Molecular dynamics

The aim of MD is to simulate the motions of molecules, using the basic Newtonian equations of motion (Newton, 1729 (1960)). In this description the atom i is assumed to be a singular point, with the mass m_i. If the position of the particle is called r_i, the velocity is given by the first derivative of position with respect to time (∂t):

$$v_i = \frac{\partial r_i}{\partial t} = \frac{p_i}{m_i} \tag{1.2}$$

Where p_i is the momentum of the particle. The net force exerted on the particle is given by:

$$F_i = \frac{\partial p_i}{\partial t} = -\frac{\partial V}{\partial r_i} \tag{1.3}$$

Where V is the energy calculated in the potential function 1.1. The force is thus the negative gradient of potential energy in point i with respect to position of point i. The final equation needed to describe the motion of the system is Newtons second law, describing the acceleration of particle i:

$$a_i = \frac{\partial^2 r_i}{\partial t^2} = \frac{F_i}{m_i} \tag{1.4}$$

In MD a system of atoms is set in motion by assigning a random set of velocities, usually drawn from a Boltzman distribution of velocities at a given temperature (energy). The new position of atom $i(X)$ after a short time interval Δt can be described by the Taylor series:

$$X(t+\Delta t) = X(t) + \frac{\partial X(t)}{\partial t}\Delta t + \frac{1}{2}\frac{\partial^2 X(t)}{\partial t^2}\Delta t^2 \ldots \tag{1.5}$$

Producing a numerical solution to this equation involves the calculation of velocity (first derivative) and acceleration (second derivative). It is however necessary to make approximations to the higher derivatives in the infinite series. The difference between various molecular mechanics algorithms basically lies in the way these higher derivatives are handled. A review of various MD algorithms is given in McCammon and Harvey (McCammon and Harvey, 1987).

1.7.2 Minimisation

The second molecular mechanics methodology which is used in this thesis is minimisation. The aim of minimisation is to find positions for atoms in a molecule such that the global potential energy function has a minimum. It is easy to find the minimum of a function with few (one to five) degrees of freedom, using analytical methods. Minimising structure coordinates for large molecular systems is a many body problem with $3N$ degrees of freedom, where N is the number of atoms in the system, and requires nonlinear optimisation. All methods involve the Taylor expansion of the potential energy function V as a function of coordinate position x:

$$V(x) = V(x_0) + (x - x_0)\frac{\partial V(x)}{\partial x}\Delta x + (x - x_0)^2\frac{1}{2}\frac{\partial^2 V(x)}{\partial x^2}\Delta x^2 \dots \tag{1.6}$$

Where x is the change of one degree of freedom

Minimisation methods are classified in order of the highest derivative involved in the method. The most frequently used methods are: Steepest decent (first order), Conjugate gradients (first order), Newton-Raphson (second order). All these methods are described in detail by Jacoby *et al* (Jacoby *et al.*, 1972). Since minimisation is a very difficult problem to solve, because the large systems get trapped in local minima, only a very small part of the phase space is searched in a minimisation. There are many minimisation methods which seek to remedy this but the description of these are not within the scope of this thesis (Jacoby *et al.*, 1972).

1.7.3 Monte Carlo methods

J. von Neumann and S.M. Ulam introduced, around 1945, the Monte Carlo method of solving problems which have a large solution space. They showed that a solution could be computed by performing a random walk through the solution space, and a practical approach was outlined by (Metropolis *et al.*, 1953). Instead of computing the analytical solution, a solution is generated by random sampling of the solution space. Metropolis developed the method further by introducing a probability density function and an objective evaluation function E, in a process of simulated annealing or simulation of a cooling process. The result becomes a biased random walk, having an initial state where all moves are allowed. By slowly lowering the probability for accepting an unfavorable move the system is moved towards a global minimum.

In terms of molecular structure determination the objective evaluation function is an energy function, and the probability function is derived from the Bolzman distribution. Assuming that a given molecular structure will adopt a conformation which represents a global minimum and a well "packed" (no space between the atoms) conformation, a simple energy function can be used for evaluation:

$$E = \varepsilon_o \sum_{i=1}^{n} ((\frac{r_o}{r})^6 - 2(\frac{r_o}{r})^{12}) + \kappa_o \cdot cos(3\omega) \tag{1.7}$$

Where the first term is a simple *Lennard-Jones* potential which evaluates the non-bonded contacts between the atoms in a given molecule and the second term is a simple torsional term which only applies to C-C bonds. The torsional term biases the function towards 60° rotamers. ε_o and κ_o are constants. The Metropolis function:

$$P = e^{\frac{-\Delta E}{T}} \tag{1.8}$$

is used to evaluate the energy function. Any move which results in a decrease in energy is accepted, and any move which results in a positive ΔE is only accepted with the probability P. This method can be used to search the large conformational space defined by a set of torsion angles and find or define the global minimum which exist for a molecule. It is necessary to emphasise that the Metropolis method of simulated annealing is not a minimisation, but merely a biased random walk. The value T is the simulation parameter which determines how fast the function should approach a minimum. In the case of thermic motion this is temperature, thus the denotation T. In the following chapters this will be termed the simulation temperature.

1.8 The aim of this thesis

The scope of this thesis is two-fold. First, to improve upon existing methods and algorithms that will enable the user to model antibody combining sites from amino acid sequence alone. Second, to use these algorithms in the *de novo* design of of an antibody combining site specific for a known antigen. The methods are based on the earlier work of the previous members of the group (Martin, 1990), which led to a combined modelling algorithm, CAMAL developed by Martin *et al* (1989;1991a).

The requirement for antibody modelling and design originates from the slow speed at which structure elucidation progresses, compared to the rate at which mutagenesis experiments can be performed. The present rate is such that only four to six new crystal structures of antibodies are published each year. The time it takes to solve the structure of an antibody can be in the range of one to three years since the work includes many stages of biochemical characterisation, purification and crystallisation etc. In order to get a reasonably fast turnover in the **protein engineering cycle** (Blundell and Sternberg, 1985; Rees and de la Paz, 1986) there is a requirement for fast access to structural data of mutant proteins. In some instances it might not be possible to crystallise the protein at all. Molecular modelling is one answer to this problem, although application of methods such as NMR (Rees *et al.*, 1989) and Laue crystallography (Hajdu *et al.*, 1987) show promise for the future.

The further development and testing of the combined algorithm is presented in Chapter 2. New methods for the construction of frameworks and sidechains have been developed and tested by modelling of three antibodies, which later had their structure solved by x-ray crystallography.

In Chapter 3 it is shown how the antibody modelling programs and databases can be used to make changes to F_V structures without changing the specificity of the antibody in a new method of antibody humanisation called "resurfacing".

Finally in Chapter 4 a method for the *ab-initio* design of antibodies (changing specificity) is presented and tested by modelling an anti-morphine antibody from the crystal structure structure of the anti-peptide antibody Gloop-2 (Jeffrey *et al.*, 1991).

Chapter 2

Modelling antibody combining sites

The modelling of antibody combining sites was first attempted by Padlan & Davies at a time when very few antibody structures were known (Padlan *et al.*, 1976). Nonetheless, Padlan and colleagues recognized that the key lay in the high structural homology that existed within the β-sheet framework regions of different antibody variable domains. The antigen combining site is formed by the juxtaposition of six inter-strand loops, or CDRs (Complementarity Determining Regions) (Kabat *et al.*, 1992), on this framework. If the framework could be modelled by homology then it might be possible to model the CDRs in the same way. Padlan and Davies reasoned that CDR length was the important determinant of backbone conformation though the number of antibody structures was insufficient to thoroughly test this maximum overlap procedure (**MOP** (Padlan *et al.*, 1976)).

In the MOP procedure a framework is chosen from one single structure on the basis of sequence similarity. Loops are then sampled from the Brookhaven (Bernstein *et al.*, 1977) database, which fit the required length, these loops are then scored according to sequence identity and the most similar loop is chosen as the final conformation.

The MOP idea was not picked up again until the early 1980's when a similar approach to modelling antibody combining sites based on a more extensive analysis of antibody structures (Darsley *et al.*, 1985; de la Paz *et al.*, 1986), was proposed.

These knowledge-based procedures are further exemplified for antibodies by the work of Chothia & Lesk who, in 1986, extended and modified the MOP procedure by introducing the concept of "key" residues (Chothia *et al.*, 1986) (See Figure 2.1). These residues allow the further subdivision of CDRs of the same length into "canonical" structures which differ in having residues at specified positions that, through packing, hydrogen bonding or the ability to assume unusual values of the torsion angles ϕ,ψ and ω, determine the precise CDR conformation. Similar knowledge-based methods have been proposed for predicting loop conformations in general (Thornton *et al.*, 1988; Tramontano *et al.*, 1989). These methods rely on the crystallographic database of protein structures. However, none of the above knowledge-based methods has been totally successful. In particular, the MOP or canonical structure approaches have succeeded in modelling at most five of the six CDRs. This stems from the fact that the third CDR of the heavy chain, H3, is more variable in sequence, length and structure than any of the other CDRs. This extra variability arises from V-D-J-C splicing (see Section 1.4).

To deal with the CDR H3 problem several groups have attempted to use *ab-initio* methods to model the combining site (Bruccoleri and Karplus, 1987). The requirement of such methods is that the total accessible conformational space to a particular CDR is sampled. Typical of purely geometric approaches is that of Gō

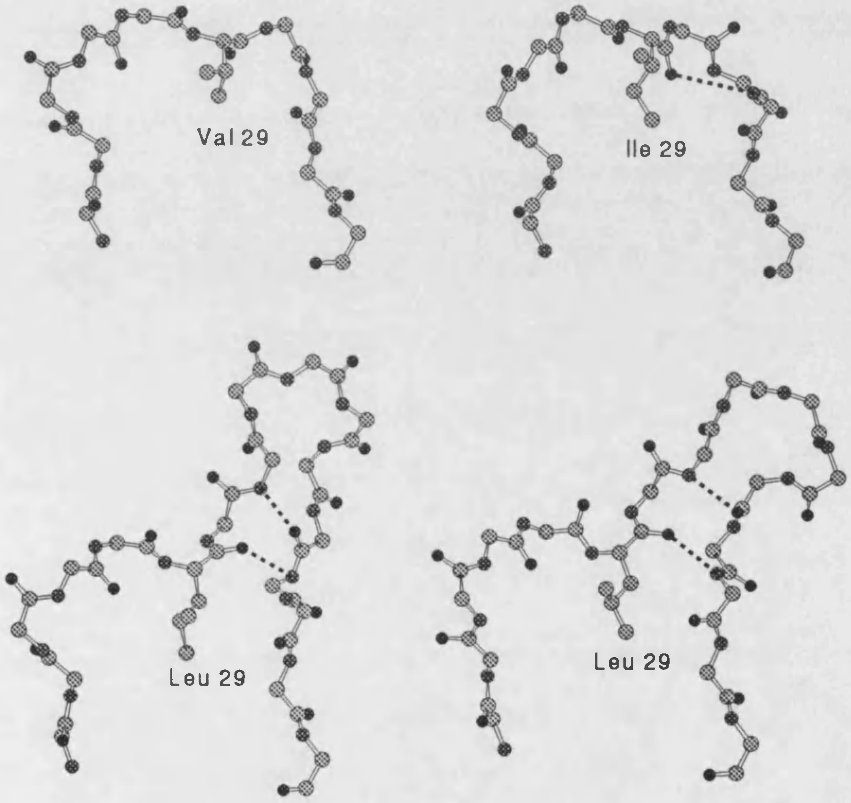


Figure 2.1: The canonical concept illustrated by the CDR L1 groups as defined by (Chothia *et al.*, 1989). The conformation of the loop is defined by the length of the loop and the existence of a small hydrophobic residue at position 29 in the light chain sequence. The small residue is packing to the framework of the F_V for short loops this leads to an "arch" like conformation. For longer loops the "arch" is retained, but with an additional "bulge" on the loop.

& Sheraga and more recently Palmer & Sheraga, where the problem is reduced to one in which the central region of the polypeptide backbone, having characteristic bond length and bond angles (rigid geometry), is constructed between the end points of the loop (CDR if an antibody loop) by a "chain closure" algorithm (Go and Sheraga, 1970; Palmer and Sheraga, 1991). In a modification of this algorithm, Bruccoleri & Karplus introduced an energy minimisation procedure which greatly expanded the domain of conformational space searched during the chain closure procedure (Bruccoleri and Karplus, 1987). This modification is incorporated into the conformational search program CONGEN (Bruccoleri and Karplus, 1987), which also allows the user to choose any set of standard bond length and bond angles such as the CHARMM (Brooks *et al.*, 1983) standard geometry parameter sets. Other approaches such as minimisation (Moult and James, 1986), or molecular dynamics (Fine *et al.*, 1986) either fail to saturate conformational space or are unable to deal with the problem of long CDRs. Whichever of the *ab initio* methods is employed, the consequence is one of defining the selection criteria in such a way as to allow the unambiguous identification of the *correct* structure (in this context *correct* is defined by reference to an appropriate X-ray structure) within the ensemble of candidates, for every CDR. To date this problem has not been solved.

In this thesis a more holistic approach has been applied when modelling CDRs which combines the advantages of knowledge-based and *ab initio* methods in a single algorithm known as A*b*M (Anti*b*ody Modelling), which includes CAMAL (Combined Algorithm for Modelling Antibody Loops) (Martin *et al.*, 1989; Martin *et al.*, 1991a; Gregory *et al.*, 1990).

2.1 A combined algorithm

The combined algorithm (CAMAL) developed by Martin *et al* (1989;1990;1991a) attempts to combine the advantages of both *ab initio* and knowledge based or database methods, and minimise the disadvantages at the same time. The conformational search program CONGEN searches all of the conformational space for small fragments of proteins (three to seven residues). The computational time is short for small peptides of three to five residues, but increases exponentially with the number of residues searched (N complete problem (Press *et al.*, 1990)). For database search methods involving loops this time is almost constant for any length of peptide since the same number of constraints is applied to short and longer loops. The major disadvantage of database methods is that they fail to saturate the conformational space available to long peptide fragments.

The whole procedure (A*b*M) is outlined in Figure 2.2. This flowchart also contains the modifications added during the course of the work presented in this thesis (Indicated by shaded boxes). The individual steps in the modelling procedure are described in the following sections.

2.2 Sequence analysis

The comparative analysis of protein sequences is the first step in the study of protein structure and function. When this is coupled to three-dimensional information for a given family of homologous proteins it becomes a powerful tool for determining residues which are important for a particular structural or functional role. The large number of sequences and structures available make antibodies an

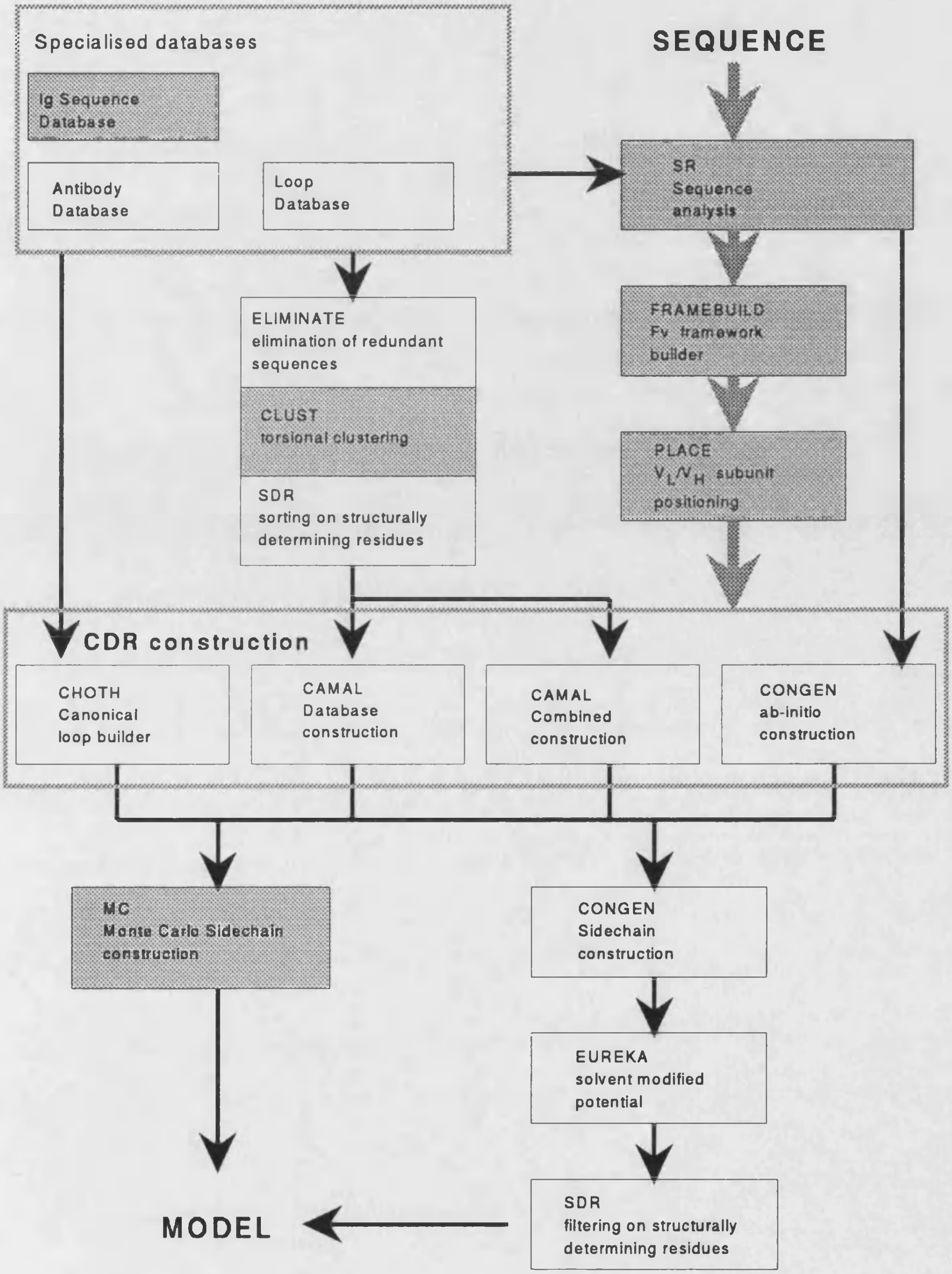

Specialised databases
Ig Sequence Database
Antibody Database
Loop Database
SEQUENCE
SR Sequence analysis
FRAMEBUILD Fv framework builder
PLACE V_L/V_H subunit positioning
ELIMINATE elimination of redundant sequences
CLUST torsional clustering
SDR sorting on structurally determining residues
CDR construction
CHOTH Canonical loop builder
CAMAL Database construction
CAMAL Combined construction
CONGEN ab-initio construction
MC Monte Carlo Sidechain construction
CONGEN Sidechain construction
EUREKA solvent modified potential
SDR filtering on structurally determining residues
MODEL

Figure 2.2: Flowchart of the antibody modelling algorithm A*b*M. The various stages of the modelling protocol are outlined in the text. The capitalised names refer to program names. Shaded boxes indicate algorithmic steps added during the course of this thesis

Sequence species	Chain	
	Heavy	Light
Caiman	3	-
Chicken	4	26
Canine	3	1
Duck	-	2
Frogs	15	-
Gold fish	8	-
Human	129	164
Mouse	490	369
Shark	3	-
Sheep	-	1

Table 2.1: Current number of sequence entries in the A*b*M sequence database. Alignments of human sequences, and some statistics appear in Appendix A. Assigned descriptors to date are: Species, Canonical classification, V_H/V_L Pairing, Pairing residues, Accessible surface residues, CDR-framework contacting residues. The database was compiled from: Swissprot Rel 17, GenBank Rel 67, NBRF Rel 28, PDS-Kyoto V.5, NEWAT, Brookhaven (may 92)

ideal family for this type of analysis.

A specific antibody sequence database has been set up using data from available DNA and protein databanks of aligned heavy and light chain antibody sequences (Figure 2.2). The sequence alignments are performed on the L and H chains separately, and independently for each of the species for which sequence information is available. The specification of the database is outlined in Table 2.1. The sequences have been aligned using the sequence alignment program, AMPS (Barton and Sternberg, 1987; Barton and Sternberg, 1990; Barton, 1990). Alignments were then inspected using the sequence handling program SR written by S.M.J. Searle (1992). Within any group of germline related somatically mutated sequences only one was retained in order to obtain a database of unique sequences for use in statistical analysis. Also, all incomplete variable region sequences were eliminated, such that the database only contained sequences covering the complete V_H or V_L region.

The database entries conform to NBRF format (Bleasby, 1990) which is the current standard for protein sequences, and supported by most sequence databanks. This format enables the assignment of any **descriptor** to a sequence, allowing the sequence database to become a "knowledge database". In this database a descriptor is a set of numbers or a string of descriptive text. An example of a sequence entry is given in Appendix A.1. These descriptors can be used for sorting the data in the database after any required combination of properties, such as the combination of canonical loops present within a particular chain. The legend to Table 2.1 contains a list of currently assigned descriptors.

This database was tabulated before the sequence database of Kabat *et al* (1992) became available on computers. This database, although it contains more sequences, does not have all the property descriptors available in the A*b*M database.

The construction of a three dimensional model for a given sequence is preceded by consultation of the sequence database in order to determine any variation of CDR length from the statistical consensus (see Figure 2.3).

For example if a 7 residue loop is to be modelled for L2, then this can be done with high confidence since 95 percent of all the CDR L2's are of this length, and conformational space can be saturated adequately or a canonical loop can be selected. In contrast if an H3 loop of length 14 residues is to be built, confidence will be lower. The distribution of loop length in the sequence database reflects the distribution in the structural database, and the average loop length for CDR H3 loops in the sequence database is 9-12. The conseqence is that conformational space will not be saturated adequately by a database search alone.

The sequence comparison is a step in the direction of validating a given model,

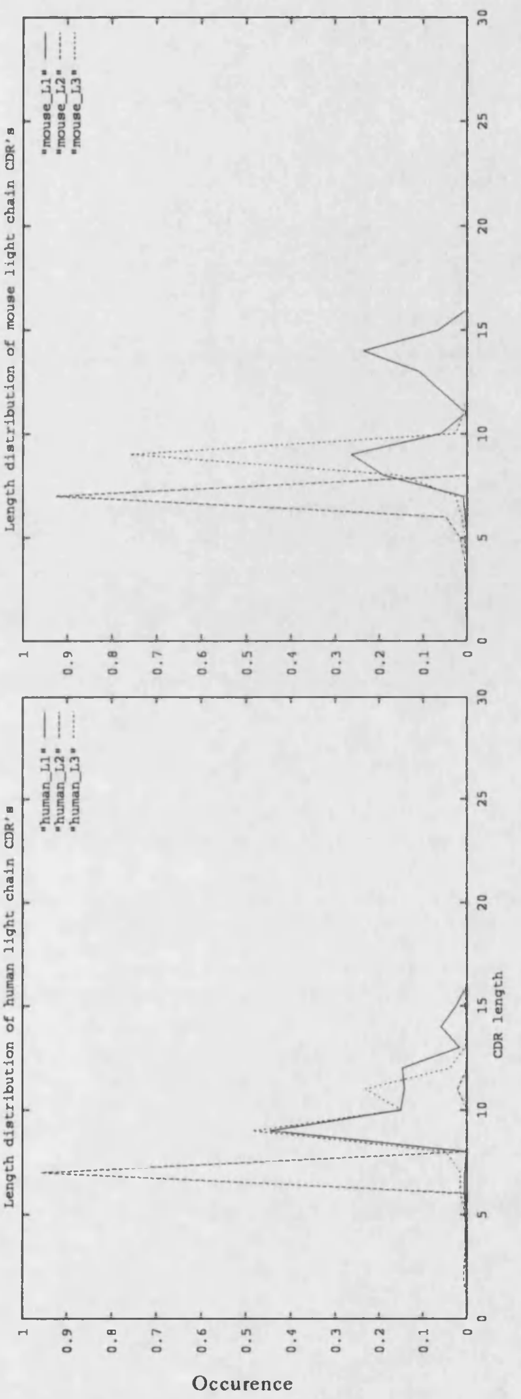


Figure 2.3: CDR length distribution for sequences in the Kabat sequence database (Kabat *et al.*, 1992). Number of sequences used are: human light chains 239, human heavy chains: 155, mouse light chains: 585, mouse heavy chains: 836. Distributions for human and mouse chains are shown

and to pinpoint any weaknesses in the modelling. The length distribution of each of the six CDR's has been tabulated and is the major descriptor for abnormality when comparing a sequence. These distributions are found in Figure 2.3.

2.3 The framework region

Antibody framework regions consist of conserved sequences that form a β-barrel structure (see Figure 1.2).

In the original method developed by Martin *et al* (1989) the framework of the antibody was generated using a simple interactive homology modelling protocol. In this protocol the light and heavy chain structures were selected on the basis of sequence similarity, where similarity was defined as the number of identical residues in an optimal sequence alignment between the crystal structure sequences and the sequence to be built. If the light and heavy chains came from different parent crystal structures the light and heavy chains were paired by superimposing the heavy chain selected onto the heavy chain of the structure from which the light chain was derived. Subsequently the redundant light and heavy chains were removed. The amino acids were then corrected to match the sidechains of the required sequence using the sidechain replacement algorithm of Jones and Thirup (1986), which is implemented in the molecular graphics program FRODO. No further refinement of the framework was performed before the CDR's were constructed. This method has several disadvantages: it does not take variations in the V_H/V_L interface residues into account, and it relies to a large extent on interactive, intuitive model building which generates results that cannot be consistently reproduced.

Brookhaven entry	name	resolution (Å)	chain types	reference
2hfl(*)	HyHel-5	2.54	κ/γII	(Sheriff *et al.*, 1987)
3hfm(*)	HyHel-10	3.0	κ/γI	(Padlan *et al.*, 1989)
1bjl/2bjl	LOC	2.8	κ/κ	(Schiffer *et al.*, 1989)
2fbj(*)	j539	1.95	κ/γIII	(Mainhart *et al.*, 1984)
3fab/7fab(*)	NEW	2.0	λI/γII	(Saul *et al.*, 1978)
4fab(*)	4-4-20	2.7	κ/γII	(Herron *et al.*, 1989)
5fab/6fab(*)	36-71	1.9	κ/γI	(Rose *et al.*, 1990)
1mcp/2mcp(*)	McPC603	3.0	κ/γIII	(Segal *et al.*, 1974)
3mcg	MCG	2.0	λI/λI	(Ely *et al.*, 1989)
1mcw	WEIR/MCG	3.5	λI/λI	(Ely *et al.*, 1985)
2rhe	RHE	1.6	λI/λI	(Furey-Junior *et al.*, 1983)
1rei	REI	2.0	κ/κ	(Palm and Hilschmann, 1975)
2fb4/2ig2(*)	KOL	1.9	λI/γIII	(Marquart *et al.*, 1980)
1f19(*)	R19.9	2.8	κ/γIIb	(Lascombe *et al.*, 1989)
1fdl(*)	D1.3	2.5	κ/γII	(Amit *et al.*, 1986)
1mam	YS*T9.1	2.5	/γIIb	(Rose *et al.*, 1992)
8fab	HIL	1.8	λ/γI	(Saul and Poljak, 1992)
1baf	ANO2	2.9		(Brunger *et al.*, 1991)
1hil/1hin/1him	17/9	2.0	κ/γIIa	(Rini *et al.*, 1992)
(*)	Gloop-2	2.8		(Jeffrey *et al.*, 1991)
1igf/2igf	B13I2	2.8	κ/γI	(Stanfield *et al.*, 1990)
1dfb(*)	3D6	2.7	κ/γI	(He *et al.*, 1992)

Table 2.2: List of antibodies used in the antibody modelling program A*b*M. Structures which do not have a brookhaven entry are not yet deposited. Antibodies used for β-barrel analysis are marked with a (*).

In this study the frameworks are built from a database of known antibody structures (see Table 2.2), using sequence homology for selection of the light (L) and heavy (H) chain V-domains, and are then paired by least squares fitting on the most conserved strands of the antibody. These β-barrel strands differ from the strands constituting the domain interface as defined by Chothia *et al* (1985), as they are selected on the basis of secondary structure and sequence conservation and not excluded surface area.

The most conserved strands were determined by analysing the barrels of known antibody crystal structures. Twelve antibodies (in Table 2.2 twenty two structures are listed; at the time of the analysis only twelve were available) were fitted using a multiple structure fitting program (Pedersen, 1992). Eleven structures were fitted onto one of the set selected at random and mean coordinates were calculated. Twelve structures were then fitted onto these mean coordinates and new mean coordinates determined. This procedure was iterated until the mean coordinate set converged (5–10 cycles). The variance for the mean coordinates at each barrel point (N,Cα,C) was then calculated. The conjugated axis of the β-barrel is here calculated from the fitting of the mean β-barrel to the surface of a hyperboloid:

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} - \frac{z^2}{C^2} = 1 \tag{2.1}$$

The parameters for A, B and C are taken from (Novotny *et al.*, 1984; Novotny *et al.*, 1983). This fit is shown in Figure 2.4.

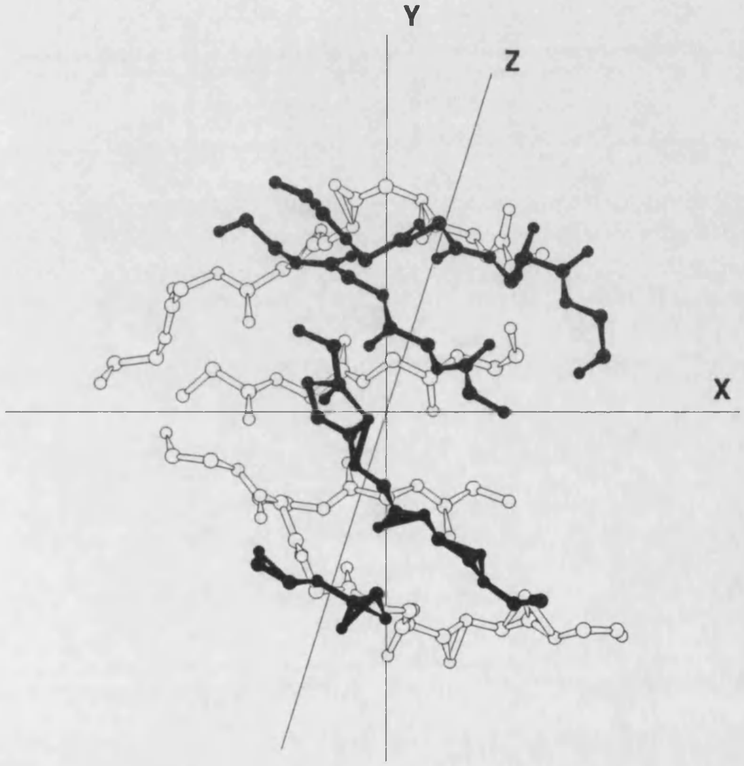


Figure 2.4: Plot of the average β-barrel strands derived form the multiple fitting procedure. The conjugate axis is here equivalent to the x-axis, shown on the figure. The RMS deviation of the fit to a hyperboloid is 2.01 Å. The light chain strands are shown in white, and the heavy chain strands have black bonds and atoms.

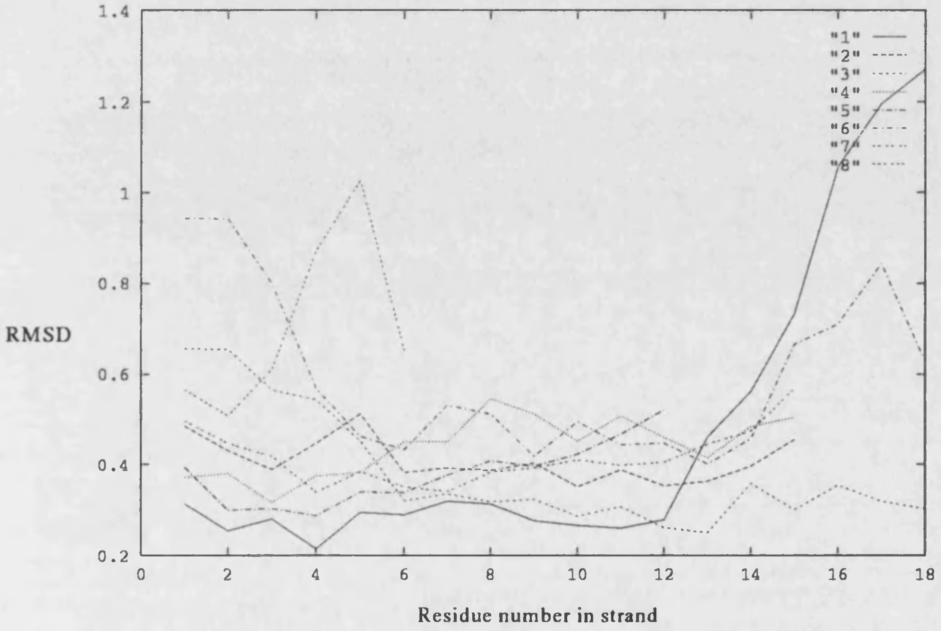


Figure 2.5: Plot of RMS deviation from the mean of the eight β–sheet strands comprising the framework. The RMS was calculated from structures R19.9, 4-4-20, NEW, FBJ, KOL, HyHEL-5, HyHEL-10, D1.3, Gloop-2 and McPC603. N,Cα,C atoms are included in the plot. The residues used are listed in Appendix B.3.4). The most disordered residues are all the residues of strand HFR4, the last residue of LFR1, and the first and last residue of HFR2.

In Figure 2.5 the variance is plotted against the sequence position. Strand 8 and all but two residues of strand 7 in both light and heavy chains were eliminated as they showed deviations greater than 3σ (standard deviation units) from the mean coordinates. These two strands comprise the takeoff points of CDR H3, and suggests that any knowledge-based prediction of CDR H3 would have to take into account sequence and length variation in the CDR itself, and the position of the participating strands. The remaining mean coordinates were used as a scaffold onto which the L and H chains were fitted. Strands 7 and 8 in the final framework were obtained from the database heavy chain structure used in the construction. It is also apparent from Figure 2.5 that strands 1 and 5 have a high variability. However, those variations were not considered to be important since the variability is at the end of the strands and in the $\mathrm{F_V/F_C}$ interface, and thus not likely to influence CDR position and conformation.

The distribution of residues in the framework strands for the human and the murine sequences is shown in Tables 2.3 & 2.4.

Position	Human	Mouse
41	W 99	W 98
42	Y 88 F 8	Y 74 F12
43	Q 93 L 4	Q 74 L 22
44	Q 98	Q 88 E 5
45	K 58 H 13 L 10	K 81 R 13
46	P 89 A 5	P 80 S 13
51	K 52 R 27	K 73 Q 12
52	L 72 V 10	L 70 R 9
53	L 75 I 8 V 11	L 81 W 15
54	L 92	I 91 V 4
55	V 86 F 6	V 83
91	E 42 F 37 V 9	A 21 I 13 L 26 F 9 V 13
90	A 90 G 8	A 67 E 28
91	D 38 T 22 V 28	D 7 I 11 T 36 V 30
92	Y 99	Y 99
93	Y 90 F 9	Y 67 F 30
94	C 99	C 99
106	F 93 Y 5	F 91
107	G 94	G 92
108	G 44 Q 34 T 10	G 56 A 21 S 12
109	G 95	G 95
110	T 95	T 92

Table 2.3: This table contains the distribution of residues in the sequence database, of human and mouse light chain sequences, that make up the β-barrel V_L/V_H interface. The numbering used is the same as described at the end of the program documentation for the *framebuild* program in Appendix B.3.4. Distributions are in percent occurrence at this position of the alignment, and only occurrences higher than 5 % are included.

Position	Human	Mouse
155	W 96	W 98
156	V 70 I 23	V 86 I 10
157	R 90	R 53 K 44
158	Q 93	Q 90 K 5
164	G 83 A 6 S 6	G 58 R 20 K 8 S 7
165	L 95	L 98
166	E 88	E 96
167	W 98	W 92
168	V 46 I 22 M 17 L 13	I 61 V 18 L11 M 9
169	G 58 A 22 S 15	G 68 A 29
215	Y 98	Y 96
216	V 90 F 9	Y 80 F 18
217	C 97	C 98
218	A 82 T 10	A 80
219	R 66 K 12 P 6	R 83
237	W 91	W 95
238	G 94	G 96

Table 2.4: This table contains the distribution of residues in the sequence database, of human and mouse heavy chain sequences, that make up the β-barrel V_L/V_H interface. The numbering used is the same as described at the end of the program documentation for the *framebuild* program in Appendix B.3.4. Distributions are in percent occurrence at this position of the alignment, and only occurrences higher than 5 % are included.

The residues at equivalent framework positions in human and murine sequences are virtually identical, indicating that the V_H/V_L pairing is extremely well conserved in different species. It is surprising that the sequences of strands 7 and 8 in the β-barrel are some of the most conserved in the sequence database and the most variable in terms of structure (see Figure 2.5).

When the framework strands have been positioned the sidechains are replaced using a 'maximum overlap' method where sidechain templates were fitted on backbone atoms with the sidechain torsion angles being adjusted to match those of equivalent torsions in the parent sidechain. Various other methods, as implemented in available molecular modelling packages, were tested but found inferior to the maximum overlap method (see Appendix B.3.4 for results of comparison).

2.4 CDR main-chain construction

The procedure for predicting the structure of combining sites combines a database search with a conformational search procedure. The architecture of the program suite to perform this task is outlined in Figure 2.2.

Using CAMAL, conformations for short loops (l < 5 residues) are determined using either a database search or CONGEN. Both succeed in saturating conformational space. For medium length peptides ($5 < l < 8$) the conformation is determined by saturating the conformational space with conformations sampled in a database generated from the complete Brookhaven Crystallographic Database (Bernstein *et al.*, 1977). For long loops (l > 8 residues) the conformational space is saturated using both database search, and CONGEN conformation generation in combination. Since the takeoff positions of the CDR's are conserved in the antibody structure (see later), the base of each loop has less conformational freedom than the central part. It is therefore assumed, that the conformational space of the loop base can be saturated adequately by a structural database. The conformational ensemble of the central sections of the longer CDR's is then expanded by generating conformations *ab initio* from each of the database loops using CONGEN.

The database search utilises distance constraints for each of the six CDR loops determined from known antibody structures. These constraints are determined by calculating Cα–Cα distances within known loops and using a search range of $\bar{x} \pm 3.5\sigma$ (the mean ± 3.5 standard deviation units). A specialised database containing all the proteins in the Brookhaven Protein Databank (Bernstein *et al.*, 1977) is then searched for fragments which satisfy the constraints for a loop of the required length. The selected loop fragments are then filtered using three different screens (ELIMINATE, CLUST and SDR sorting in Figure 2.2).

ELIMINATE In the database search method by Martin (1990) the redundancy check was performed solely on the basis of structure name and position of loop in structure (ELIMINATE), and not on the basis of the actual loop conformation. This was found to be inaccurate and was modified. Removing redundancies using structure names and positions in structures alone, as performed originally in ELIMINATE, will fail to identify an ensemble of unique conformations as the structural database gets larger. There are many homologous/identical structures in the database, which have different entry ID-codes (Brookhaven name). The torsional clustering will usually remove approximately 1/3 of the database loops found in a search. Without the torsional clustering the final SDR screen of the database loops would fail, since in many cases it would rank 50-100 identical loops from different structures as the best and thus fail to saturate conformational space.

CLUST CLUST is a torsional cluster algorithm which uses a standard euclidian distance clustering (Lazlo, 1975):

$$D_{ij} = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{in} - x_{jn})^2 \tag{2.2}$$

Where D_{ij} is the square of the euclidian distance between the two conformations i and j in the n dimensional phase space. In this case n is the number of backbone torsions in a given loop. The clustering is performed as a nearest neighbor clustering. A search distance (d) is determined as the mean distance between the two closest neighbors in the complete set of loops. The classification is done iteratively and for each step the search distance to neighbors is increased by the distance d. The clustering is terminated when no neighbors are found within $\pm 3.5SD$ units of the mean of all the euclidian distances. The clustering is able to eliminate any similar or closely related loops, with respect to backbone conformation.

SDR Finally, the database loop are sorted using a Structurally Determining Residue (SDR) protocol. In SDR (Sutcliffe *et al.*, 1987a) each residue in each database fragment is assigned as being structurally determining if it causes the next Cα to be moved relative to the position of that Cα in any of the other database loops and the structurally determining residues are scored against the sequence being modelled using the Dayhoff mutation matrix (Sutcliffe *et al.*, 1987a; Dayhoff *et al.*, 1983). Only the best 200 loops are used in the further construction, in order to reduce the computational task.

In the combined algorithm (CAMAL in Figure 2.2)the middle section of the loop is deleted and reconstructed using the conformational search program CONGEN (Bruccoleri and Karplus, 1987) (CONGEN in Figure 2.2). For loops of six or seven residues, the structural database appears to saturate the conformational space available to the backbone adequately and only sidechains are built by conformational search (see below for a further analysis and description of the sidechain reconstruction). Loops shorter than six residues are built by conformational search alone since this is computationally feasible and because the number of loops selected from the database becomes unacceptably large as loop length decreases.

When modelling a complete combining site, loops of 6 or more residues are modelled individually with the other loops absent. If the loops are built consecutively, small errors can accumulate leading to a poor result (Martin, 1990). However, recent work by S.M.J. Searle suggests that, where canonical loops are identified, their presence as backbone structures during the modelling of non-canonical loops gives greater accuracy of the final model (Searle, 1992).

2.5 Sidechain reconstruction

A number of different methods of sidechain reconstruction have been evaluated. The methods currently available fall into two main groups:

- **Knowledge based** - using statistical information of χ angle distributions in different types of secondary structure, in the crystallographic database.

- *ab-initio* **based** - using different types of conformation generation methods, and evaluating generated conformations using an objective function.

Database methods There are several studies in the literature which indicate that the conformational preference of amino acid sidechains depends upon which secondary structure they are in (Mcgregor *et al.*, 1987; Summers *et al.*, 1987; Sutcliffe *et al.*, 1987a). Unfortunately, there is only limited documentation of preferences in loop structures (Sutcliffe *et al.*, 1987a). The information which is available is for all types of loops or turns collectively, thus giving a low confidence when trying to assign the sidechain conformation to particular types of loop such as antibody CDR's. The loop or turn structure is not a random coil, but falls into many different, and some still unclassified, sub-groups. Ponder and Richards have shown that the occurrence of sidechain conformations in proteins is limited to a set of rotamers for each of the amino acid sidechains, and have constructed a library of these conformations (Ponder and Richards, 1987b; Ponder and Richards, 1987a). Unfortunately these rules only apply to internal (core) residues of the protein. For exposed, surface residues (most CDR residues) this rotamer library can not be used.

The main disadvantage of database methods is that they do not take local environment conditions into consideration, except for the geometric contribution of the backbone conformation (secondary structure).

Again, the methods are limited to the knowledge present in the database though when the loop type being modeled is a canonical CDR, these methods usually have a higher confidence (Sutcliffe *et al.*, 1987a; Chothia *et al.*, 1989) than *ab initio* methods. Thus, it seems obvious that one should use a combination of knowledge based and *ab initio* methods in order to obtain the best from both.

Ab initio methods The conformational search program CONGEN has an interesting treatment of the sidechain problem. The program has implemented a set of different side chain reconstruction algorithms, all using the CHARMM (Brooks *et al.*, 1983) potential for the evaluation of conformations. CONGEN uses a torsional grid search for the generation of conformations, and extensive tree pruning during the recursive generation, in order to avoid combinatorial explosions (when few sidechains are reconstructed). The different generation options available are outlined in Table 2.5. The major disadvantage of CONGEN is that reconstruction of sidechains of more than five to six residues results in combinatorial explosion. This problem could be overcome by using a coarser (30-60 °) grid. Unfortunately the algorithm is then not able to saturate the conformational space and other methods have to be considered. In Table 2.5 a test is shown of these CONGEN methods on the antibody 3D6 and the cpu time spent on the calculation.

An alternative approach is to search sidechain conformations using Monte Carlo simulated annealing. When the the evaluation function outlined in equation 1.7 is applied, the system usually gets trapped in an energetic minima well before the global minimum is encountered, at a high temperature and without the solution space having been searched sufficiently. This problem can be solved by truncating the *Lennard-Jones* potential in equation 1.7, thereby allowing atoms to pass through each other. In reality this function would converge towards infinity when the distance r between the atoms goes towards zero.

The torsional potential is pre-calculated and only updated every 10 steps and, since the average movement over 10 random steps is no more than $10\cdot\sqrt{10}$ (when using 10° grid) the precision of the energy calculation is maintained. The torsional potential term has only little influence when trying to determine internal side chain conformations, but becomes significant for surface sidechains. The above method of generating sidechain conformations has been successfully used to determine sidechain conformations for core residues (Lee and Levitt, 1991; Lee and Subbiah, 1991).

Evaluation of side chain conformations generated by simulated annealing is done solely on the basis of energy for internal (core) residues, since good van der Waal's interactions are considered to be equivalent to a good packing of the residues. The situation becomes more complicated when trying to predict the conformation of surface residues.

The lowest van der Waal's interaction is obtained by a combination of side chain conformations which minimise the overlap of atoms. There is however nothing in the simple potential (Equation 1.7) which takes the surface environment into account. The sidechains can adopt many well packed conformations on the surface, all equally favorable. The implication of this is that the method described by Lee (Lee and Levitt, 1991; Lee and Subbiah, 1991) cannot be applied directly when predicting surface sidechain conformations.

Adapted Monte Carlo method Using the fact that hydrophobic, bulky residues will be shielded by the hydrophilic sidechains, and will be buried in the surface, it is possible to generate simple functions which will evaluate these macroscopic observations. These functions can either be implemented in the objective evaluation function of the Monte Carlo simulation, or as is done here, added as a post processing step. Including an accessibility/hydrophobicity term in the evaluation function would slow down the calculation considerably, hence the term has been added as a screening function.

In the functions used here the accessibilities and the hydrophobicities have been scaled appropriately. All residual accessibilities are relative to the accessibility of of a given amino acid in isolation in the conformation in which it is found in the protein structure. The accessibilities are therefore in the range $[0;1]$. Residue hydrophobicities are taken from Cornette *et al* (1987), but have been scaled in the range $[-1;1]$. The simplest type of function can be either of two:

$$f_a = -\sum \frac{A_{rel}}{H_{rel}} \quad f_a \in \,]-\infty;\infty[, H_{rel} \neq 0 \tag{2.3}$$

or

$$f_a = -\sum A_{rel} \cdot H_{rel} \quad f_a \in [-1;1] \tag{2.4}$$

In these equations A_{rel} denotes the relative accessibility of a given amino acid sidechain. H_{rel} denotes the relative hydrophobicity a given amino acid. The main difference between the two functions above is the ranges in which they are defined. In Equation 2.3 the score for a favorable conformation is exponential, whereas in Equation 2.4 the score is linear for the relative exposed area of a given group. f_a in Equation 2.3 is not defined for H_{rel} or A_{rel} equal to zero. f_a in Equation 2.4 is a continuous function in the range $[-1;1]$. The surface area is calculated using the tessellated icosahedron approach (Chau and Dean, 1987), which is not very precise (0.1 percent), but is able to evaluate a large number of conformations in a short time.

Similar semianalytical expressions have been suggested by Still *et al* (1990). These have been included in energy calculations and have been shown to be able to generate conformations of sidechains which are similar in conformation to crystal structure conformations. The traditional (Still *et al.*, 1990) treatment of solvation free energy (G_{sol}), is a function consisting of three terms:

$$G_{sol} = G_{cav} + G_{vdW} + G_{pol} \tag{2.5}$$

G_{cav} is a solvent cavity term, G_{vdW} is a solute van der Waals term, and G_{pol} is a solute solvent electrostatic term. For saturated hydrocarbons in water G_{sol} is linearly related to the solvent-accessible surface area A_s.

Vila and Sheraga use an even simpler expression for the free energy of hydration:

$$G_i = \sum_{k=1}^{N} \rho_k A_k \tag{2.6}$$

Here, A_k is the solvent accessible surface area of atom k and ρ_k is the atomic solvation parameter for atom k. The solvation parameters used were determined by NMR (Vila *et al.*, 1991). This simple term was included directly in a forcefield to describe solvation (Vila *et al.*, 1991).

When generating sidechains using the MC approach it is possible to integrate over a large phase space with many degrees of freedom, and get a complete sampling of the phase space. The generation and evaluation of sidechains using

	RMS deviation						
Method	L1	L2	L3	H1	H2	H3	time (*min*)
CONGEN methods							
-all							
-independent							
-combination							
-first	2.36	1.16	1.72	1.28	1.92	2.40	0.023
-itera + seq	2.40	1.01	0.88	1.23	1.82	1.98	221.0
-itera + order	2.31	0.83	0.92	1.06	1.39	1.62	220.0
Monte Carlo							
+HpH function	1.74	0.98	1.20	1.16	1.16	1.91	16000.0
+E$_{min}$	1.56	1.10	0.93	1.15	1.16	1.76	16000.0
Random	2.82	1.76	2.46	1.76	2.30	2.39	

Table 2.5: Evaluation of possible sidechain reconstruction schemes, reconstructing 49 sidechains of the CDR's of antibody 3D6. The first three CONGEN methods have not been tested since they are unsuitable. *All* tries to generate all the possible conformations, using nested loops, thus for 49 residues this would be in the order of 3^{49} conformations, and the *cpu* time needed in the same order of magnitude. *Independent* generates all the sidechains independently of each other, only taking backbone into consideration, the CHARMM energy evaluation function can not be used for the evaluation in this case since many large repulsive van der Waals clashes are generated. *Combination* generates a small number of energetically favorable conformations for each sidechain, and then evaluates all the possible combinations of these, unfortunately if just the two lowest energy conformations are chosen for each of the sidechains in this case 2^{49} conformations would have to be evaluated, this renders only the two final methods possible for this type of problem. *First* uses the same algorithm as all, but just retains the first acceptable energy conformation, thus selecting a more or less random low energy conformation, which is detected in the RMS values. The last two methods are variations on the *Iterative* method of CONGEN. In the *Iterative* method the sidechains which are to be constructed are twisted around their χ angles in a specified order. For each sidechain the lowest energy conformation is retained and the next sidechain is searched, this procedure is repeated until the total energy of the system converges. In the first of the two methods the sidechains are generated sequentially and in the second they are generated as a function of Cβ distance from center of gravity of the F$_V$ fragment. The philosophy behind the last method is to generate the sidechains first which have least conformational freedom, thus higher confidence in the construction. These new sidechains will then add more conformational constraints when constructing the more exposed sidechains. The last method (Monte Carlo) which performs a complete search of the conformational space is described in the text. The final set of RMS values is for a conformation of the 49 residues which is generated using a pseudo random number generator to generate the sidechain torsions. The sidechains in the Monte Carlo simulated annealing represent the average conformation of the 1000 lowest energy conformations. E_{min} refers to lowest energy conformation, and *HpH* refers to best conformation with respect to hydrophobicity/accessibility score.

this approach has been implemented in the program MC (Monte Carlo). The method of simulated annealing is described further in the documentation to the MC program in Appendix B.

The CDR sidechains of antibody 3D6 were reconstructed using the MC method and were compared to the results obtained with CONGEN using the *iterative* method (Table 2.5). The Monte Carlo/Metropolis method has a better performance than CONGEN which is evident from the RMS values, and Figure 2.6. The major performance difference is seen in the hydrophobic sidechains where CONGEN consistently fails to find the right conformation. Using the MC algorithm the conformation is selected which gives the best shielding of hydrophobic sidechains. Since the Monte Carlo reconstruction is not a minimisation the final conformations have also been minimised and the results are also shown in Table 2.5.

2.6 Selection of CDR conformation

All the loop conformations for which sidechains have been constructed, using CONGEN, are evaluated using a solvent modified potential, which excludes the attractive van der Waals and electrostatic terms of the non-bonded energy function contained within an appropriate potential energy function. Both the GROMOS (Åqvist *et al.*, 1985) and EUREKA (Lifson *et al.*, 1979; OML, 1992) have been shown to give identical results. All the generated conformation are then passed through the cluster algorithm again and the lowest five *different* energy conformations are selected and filtered using an SDR algorithm (FILTER), based on backbone torsion angles observed in the original database loops. Since the database search is not used for the shortest loops (5 residues or fewer) the FILTER

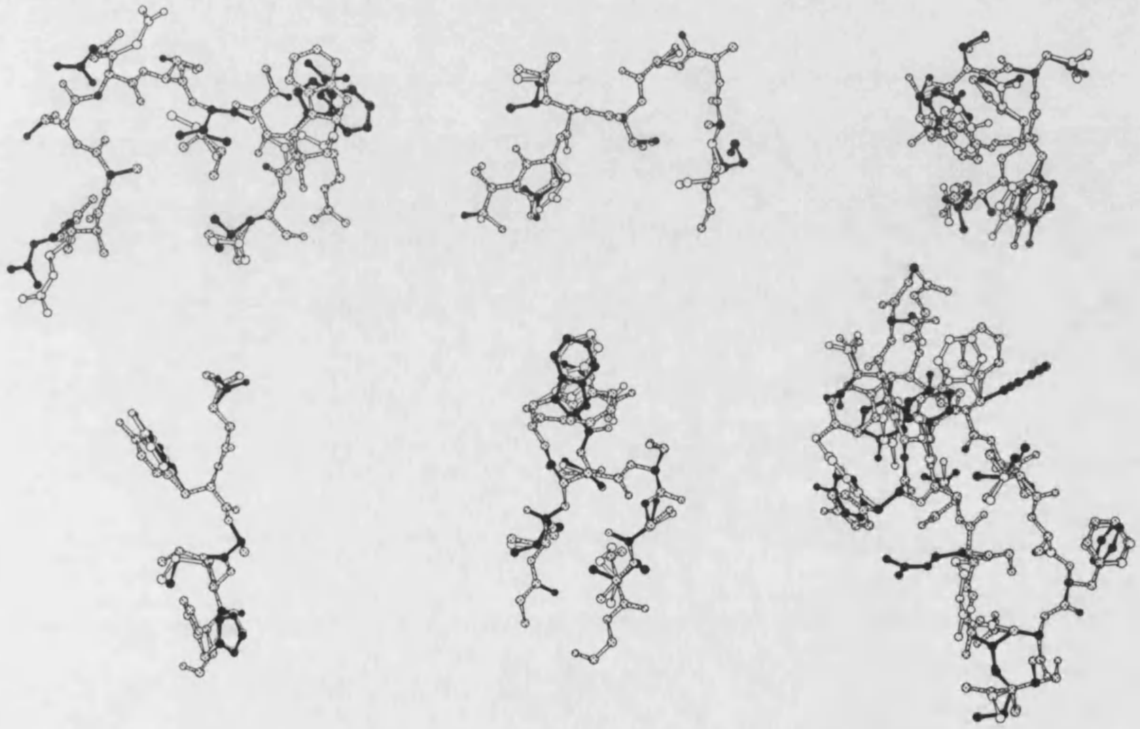


Figure 2.6: Sidechain reconstruction of the six CDR's of antibody 3D6(He *et al.*, 1992). Top: L1,L2,L3. Bottom: H1,H2,H3. White: crystal structure. Grey: sidechains reconstructed with CONGEN (iterative). Black: sidechains reconstructed using MC. The Trp in L1 and H2 are predicted correctly using MC (Black), CONGEN fails to determine this conformation.

algorithm cannot be used. Energy is thus the only available selection criterion and the short loops are built last, in the presence of the longer loops.

2.7 Modelling of three antibodies

The A*b*M algorithm has been blind tested on four F_V structures which have had their structures determined independently (Pedersen *et al.*, 1991).

In the following section the analysis of three model structures is presented. The fourth structure (Gloop-2) was modelled earlier by Martin *et al* (1989), and is included here for comparison and completeness. The three new models are:

- D1.3 (Amit *et al.*, 1986), an anti-lysozyme antibody.
- 36–71 (Rose *et al.*, 1990), an anti-phenylarsonate antibody.
- 3D6 (He *et al.*, 1992), an anti-protein (GP41 of HIV) antibody.

For all of these three antibodies the crystal structure coordinates were obtained only after the model coordinates had been deposited with the authors.

All three models were subjected to both restrained and unrestrained energy minimisation using the DISCOVER (TM Biosym Technology) potential with 300 cycles of steepest descents, followed by conjugate gradient minimisation until convergence to within 0.01 Kcal/mol between steps occurred.

The resolution and R-factors of the x-ray structures are given in Table 2.2 together with the parent frameworks selected in building the models. The structures and models were compared by global fits of the loops. The β-barrel strands 1–6, as described above, were least squares fitted and the RMS deviation was then calculated over the loops. The backbone (N,Cα,C) RMS values for fitting model and crystal structure frameworks were between 0.4 and 0.9Å, illustrating the conservation of the core β-barrel. Using all eight strands RMS deviations between 0.6 and 1.2Å were observed.

Global fits (Table 2.6) give a more realistic measure of the accuracy of the model than a local least-squares fit over the loops since they account for the overall positioning of the loops in the context of the F_V structure. Local fits, which give lower RMS deviations, are also shown in Table 2.6. Differences between local and global RMS deviations arise from differences in V_H/V_L domain packing and differences in loop 'take off' angles and positions. The antibody Gloop-2 is included in some of the comparisons, since it was the first antibody to be modelled solely using the CAMAL method (Martin *et al.*, 1989; Martin, 1990).

Table 2.7 shows the canonical loops selected for modelling 3D6. Backbone structures of the modelled CDRs, superimposed on the x-ray structures after global fitting are shown in Figure 2.7. General features and points of interest for each of the six CDRs are discussed below.

			RMS local (Å)			RMS global (Å)		
Antibody	CDR	sequence	$C\alpha$	$N,C\alpha,C$	All	$C\alpha$	$N,C\alpha,C$	All
Gloop-2	L1	RAS[Q(EIS)G]YLS	0.73	0.71	2.05	0.86	0.87	2.09
D1.3		RAS[G(NIH)N]YLA	2.29	1.93	4.34	2.72	2.43	4.59
36–71		RAS[Q(DIN)N]FLN	2.71	2.43	4.80	3.51	3.31	5.19
3D6		RAS[Q(SIG)N]NLH	0.51	0.54	2.48	0.81	0.78	2.88
Gloop-2	L2	AASTLDS	0.25	0.23	0.80	0.66	0.68	1.10
D1.3		Y[T(TTL)A]D	0.67	0.73	1.80	0.99	1.02	2.01
36–71		F[T(SRS)Q]S	0.64	0.66	2.34	0.73	0.72	2.43
3D6		KASSLES	0.41	0.42	1.37	0.83	0.86	1.73
Gloop-2	L3	LQ[Y(LSY)P]LT	0.58	0.52	1.73	0.75	0.74	2.00
D1.3		QH[F(WST)P]RT	1,41	1.35	2.89	1.76	1.79	3.46
36–71		QQ[G(NAL)P]RT	1.09	1.00	2.26	1.48	1.36	2.37
3D6		Q[Q(YNS)Y]S	1.48	1.88	3.84	2.31	1.97	3.96
Gloop-2	H1	[T(FGI)T]	0.60	0.70	2.00	1.03	1.01	2.04
D1.3		[G(YGV)N]	0.44	0.62	2.33	0.85	0.90	3.24
36–71		[S(NGI)N]	0.90	0.83	2.22	1.04	0.97	2.51
3D6		DYAMH	0.67	0.77	1.52	0.81	0.72	1.59
Gloop-2	H2	EI[F(PGN)S]KTY	0.63	0.64	1.63	1.20	0.94	2.23
D1.3		MI[W(GDG)N]TD	0.42	0.42	1.55	0.87	0.85	1.88
36–71		YNN[P(GNG)Y]IA	0.84	0.78	2.01	1.47	1.41	1.79
3D6		ISWDSSSIG	0.45	0.52	2.35	0.95	0.89	2.85
Gloop-2	H3	[R(EIR)Y]	0.66	0.89	3.44	0.87	1.07	3.68
D1.3		ER[D(YRL)D]Y	0.38	0.53	1.68	1.25	0.81	1.96
36–71		SEYY[G(GSY)K]FDY	1.95	1.75	4.40	2.65	2.53	4.60
3D6		GRDYY[D(SGG)YF]TVAFDI	3.66	3.42	5.93	4.01	3.95	6.30

Table 2.6: Sequence and conformational search construction scheme for each of the 24 CDRs, [ ]=construction area, ( )= Chain closure, all sidechains are constructed. RMS (Root Mean Square) difference between model and crystal structure loop coordinates. The RMS values are a global fit calculated by least-squares fitting the conserved core of the two structures upon each other and calculating the RMS over the loops. The total RMS of the frameworks ($N,C\alpha,C$) is 0.81, 0.60, 0.86 and 0.56 respectively. Gloop-2 is modelled solely by the CAMAL method.

Loop	Canonical	Sequence
L1	HyHEL-10	R A S Q S I G N N L H
	(3D6)	R A S Q S I S R W L A
L2	REI	E A S N D L A
	(3D6)	K A S S L E S
H1	McPC603	D F Y M E
	(3D6)	D Y A M H
H2	KOL	I I W D D G S D Q
	(3D6)	I S W D S S S I G

Table 2.7: Canonical loops selected for the model of 3D6.

2.7.1 Analysis of the CDR regions

During the comparison of CDR conformations in the V-region models and the x-ray Fab structures it was observed that at certain positions in a CDR, the peptide backbone may adopt either of two conformations by undergoing a "peptide flip" (1,4 shift). This phenomenon is also seen in type 2 β-turns (Paul *et al.*, 1990). Dynamics simulations of β-turns show that the transformation energy between $\phi1 = -60$, $\psi1 = -30$, $\phi2 = -90$, $\psi2 = 0$ and $\phi1 = -60$, $\psi1 = 120$, $\phi2 = 90$, $\psi2 = 0$ has a maximum value of 5 kcal (Paul *et al.*, 1990). This is low enough to populate both both conformations at physiological temperature (310 °K) The peptide flip is observed within several canonical classes (as described by (Chothia *et al.*, 1989)) and the hydrogen bonding pattern used to determine the conformation of a canonical class does not disallow the peptide flip. Thus, while selection of a canonical class may describe the overall conformational status of the loop, local deviations of this type will not be defined. Any modelling procedure should therefore take these, or any other multiple conformations, into consideration where the transformation energies are sufficiently low to permit population of the different conformational forms. Table 2.8 shows an example of the "peptide-flip" phenomenon from two antibody structures in the crystallographic database

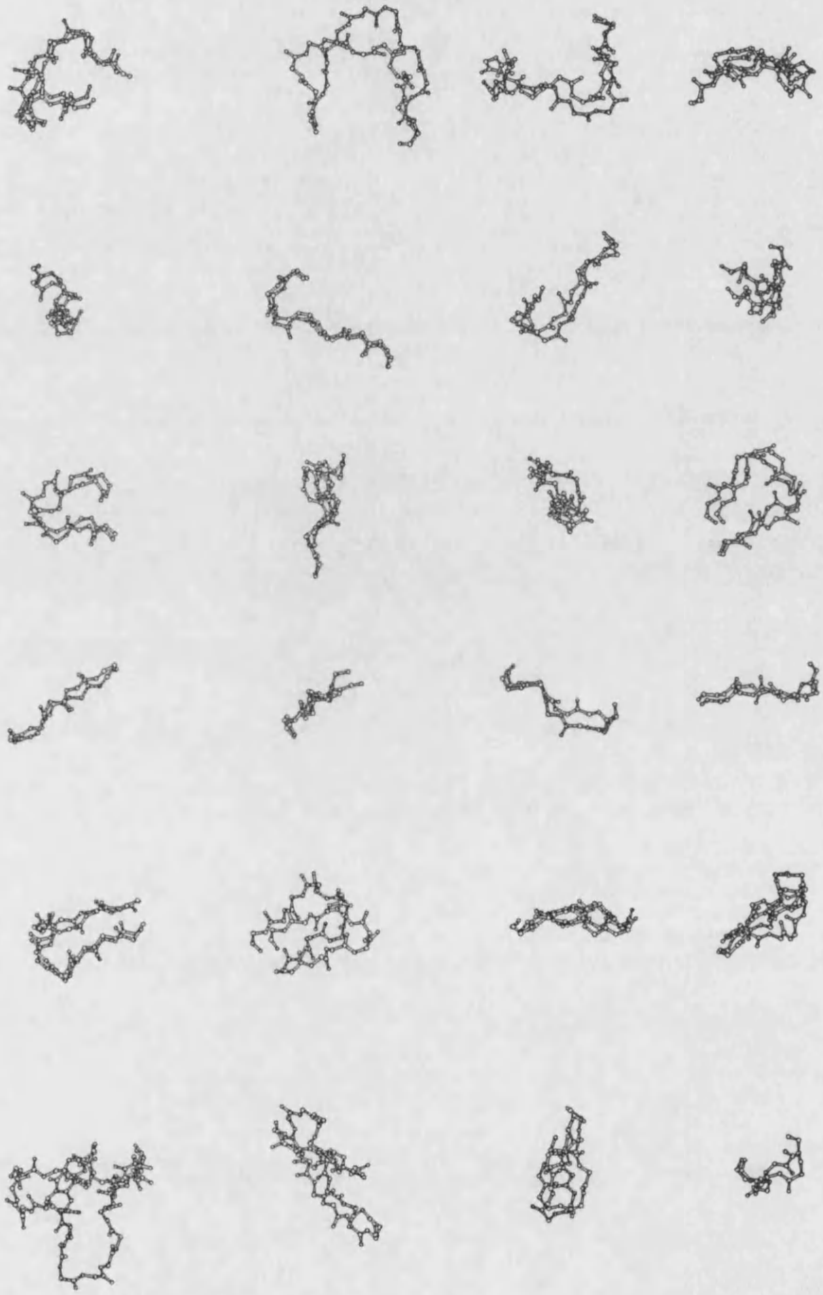


Figure 2.7: Plot of loop backbones for all the models and x-ray structures. The loops are positioned after global framework fit. This does not represent the best local least squares fit, but shows how the loops are positioned globally onto the framework. White: crystal structure. Grey: Structures modelled with A*b*M. Major deviations are only seen in H3 of 3D6 - this loop is also the longest in the set. The loops are from top to bottom L1,L2,L3,H1,H2 and H3. The structures are from left to right 3D6, 36-71, D1.3, and Gloop-2.

```
               10        20         30        40          50          60
       |--------|---------|--- ***************** ---------|----- *****
glb2:  DIQMTQSPSSLSASLGERVSLTC RASQEISG------YLS WLQQKPDGTIKRLIY AASTL
d1_3:  DIVLTQSPASLSASVGETVTITC RASGNIHN------YLA WYQQKQGKSPQLLVY YTTTL
3671:  DIQMTQIPSSLSASLGDRVSISC RASQDINN------FLN WYQQKPDGTIKLLIY FTSRS
3d_6:  DIQMTQSPSTLSASVGDRVTITC RASQSISR------WLA WYQQKPGKVPKLLIY KASSL

2hfl:  DIVLTQSPAIMSASPGEKVTMTC SASSSVN-------YMY WYQQKSGTSPKRWIY DTSKL
3hfm:  DIVLTQSPATLSVTPGNSVSLSC RASQSIGN------NLH WYQQKSHESPRLLIK YASQS
2fbj:  EIVLTQSPAITAASLGQKVTITC SASSSVSS-------LH WYQQKSGTSPKPWIY EISKL
2fb4:  QSVLTQPPSASG-TPGQRVTISC SGTSSNIG----SSTVN WYQQLPGMAPKLLIY RDAMR
1fb4:  ESVLTQPPSASG-TPGQRVTISC TGTSSNIG----SITVN WYQQLPGMAPKLLIY RDAMR
2mcp:  DIVMTQSPSSLSVSAGERVTMSC KSSQSLLNSGNQKNFLA WYQQKPGQPPKLLIY GASTR
3fab:  -SVLTQPPSVSG-APGQRVTISC TGSSSNIG---AGNHVK WYQQLPGTAPKLLIF HNNA-
1rei:  DIQMTQSPSSLSASVGDRVTITC QASQDII------KYLN WYQQTPGKAPKLLIY EASNL
2rhe:  ESVLTQPPSASG-TPGQRVTISC TGSATDIG----SNSVI WYQQVPGKAPKLLIY YNDLL
4fab:  DVVMTQTPLSLPVSLGDQASISC RSSQSLVHS-QGNTYLR WYLQKPGQSPKVLIY KVSNR
1f19:  DIQMTQTTSSLSASLGDRVTISC RASQDISN------YLN WYQQKPDGTVKLLVY YTSRL
1mcw:  -SALTQPASVSG-SPGQSITVSC AGHTSDVA---DSNSIS WFQQHPDKAPKLLIY AVTFR
3mcg:  -SALTQPPSASG-SLGQSVTISC TGTSSDVG---GYNYVS WYQQHAGKAPKVIIY EVNKR
b13i:  DVLMTQTPLSLPVSLGDQASISC RSNQTILLS-DGDTYLE WYLQKPGQSPKLLIY KVSNR
2bjl:  -SVLTQPPSASG-TPGQRVTISC SGSSSNIG---ETNSVS WYQHLPGTAPKLLIY EDNSR
       |--------|---------|--- ***************** ---------|----- *****
                                     CDR L1                       CDR

              70        80        90       100         110       120
       ** -------|---------|---------|---- *********** ----|---------|
glb2:  DS GVPKRFSGRRSGSDYSLTISSLESEDFADYYC LQYLS--YPLT FGAGTKLELKRA
d1_3:  AD GVPSRFSGSGSGTQYSLKINSLQPEDFGSYYC QHFWS--TPRT FGGGTKLEIKR-
3671:  QS GVPSRFSGSGSGTDYSLTISNLEQEDIATYFC QQGNA--LPRT FGGGTKLEIKRA
3d_6:  ES GVPSRFSGSGSGTEFTLTISSLQPDDFATYYC QQYNS----YS FGPGTKVDIKRT

2hfl:  AS GVPVRFSGSGSGTSYSLTISSMETEDAAEYYC QQWGR--NP-T FGGGTKLEIKRA
3hfm:  IS GIPSRFSGSGSGTDFTLSINSVETEDFGMYFC QQSNS--WPYT FGGGTKLEIKRA
2fbj:  AS GVPARFSGSGSGTSYSLTINTMEAEDAAIYYC QQWTY--PLIT FGAGTKLELKRA
2fb4:  PS GVPDRFSGSKSGASASLAIGGLQSEDETDYYC AAWDVSLNAYV FGTGTKVTVLGQ
1fb4:  PS GVPTRFSGSKSGTSASLAISGLEAEDESDYYC ASWNSSDNSYV FGTGTKVTVLGQ
2mcp:  ES GVPDRFTGSGSGTDFTLTISSVQAEDLAVYYC QNDHS--YPLT FGAGTKLEIKRA
3fab:  -- ----RFSVSKSGSSATLAITGLQAEDEADYYC QSYDR--SLRV FGGGTKLTVLRQ
1rei:  QA GVPSRFSGSGSGTDYTFTISSLQPEDIATYYC QQYQS--LPYT FGQGTKLQIT--
2rhe:  PS GVSDRFSASKSGTSASLAISGLESEDEADYYC AAWNDSLDEPG FGGGTKLTVLGQ
4fab:  FS GVPDRFSGSGSGTDFTLKISRVEAEDLGVYFC SQSTH--VPWT FGGGTKLEIKRA
1f19:  HS GVPSRFSGSGSGTDYSLTISNLEHEDIATYFC QQGST--TPRT FGGGTKLEIKRR
1mcw:  PS GIPLRFSGSKSGNTASLTISGLLPDDEADYFC MSYLS-DASFV FGSGTKVTVLRQ
3mcg:  PS GVPDRFSGSKSGNTASLTVSGLQAEDEADYYC SSYEGSD-NFV FGTGTKVTVLGQ
b13i:  FS GVPDRFSGSGSGTDFTLKISRVEAEDLGVYYC FQGSH--VPPT FGGGTKLEIKRA
2bjl:  AS GVSDRFSASKSGTSASLAISGLQPEDETDYYC AAWDDSLDVAV FGTGTKVTVLGQ
       ** -------|---------|---------|---- *********** ----|---------|
       L2                                    CDR L3
```

Figure 2.8: Sequence alignment for antibody crystal structures (light chains only) used in the A*b*M algorithm. The CDR's are indicated with stars. The four sequences separated at the top are the antibodies which have been modelled during the development of A*b*M. Gloop-2 was modelled by (Martin *et al.*, 1989)

```
               10        20        30            40       50          60
       |--------|---------|---------| ******* --|---------|- *********
glb2:  QVQLQQSGTELARPGASVRLSCKASGYTFT TFGIT-- WVKQRTGQGLEWIG EIFPGNS--
d1_3:  QVQLQESGPGLVAPSQSLSITCTVSGFSLT GYGVN-- WVRQPPGKGLEWLG MIWGDG---
3671:  EVQLQQSGVELVRAGSSVKMSCKASGYTFT SNGIN-- WVKQRPGQGLEWIG YNNPGNG--
3d_6:  EVQLVESGGGLVQPGRSLRLSCAASGFTFN DYAMH-- WVRQAPGKGLEWVS GISWDSS--

2hfl:  -VQLQQSGAELMKPGASVKISCKASGYTFS DYWIE-- WVKQRPGHGLEWIG EILPGSG--
3hfm:  DVQLQESGPSLVKPSQTLSLTCSVTGDSIT SDYWS-- WIRKFPGNRLEYMG YVSYSG---
2fbj:  EVKLLESGGGLVQPGGSLKLSCAASGFDFS KYWMS-- WVRQAPGKGLEWIG EIHPDSG--
2fb4:  EVQLVQSGGGVVQPGRSLRLSCSSSGFIFS SYAMY-- WVRQAPGKGLEWVA IIWDDGS--
1fb4:  EVQLVQSGGGVVQPGRSLRLSCSSSGFIFS SYAMY-- WVRQAPGKGLEWVA IIWDDGS--
2mcp:  EVKLVESGGGLVQPGGSLRLSCATSGFTFS DFYME-- WVRQPPGKRLEWIA ASRNKGNKY
3fab:  -VQLEQSGPGLVRPSQTLSLTCTVSGTSFD DYYST-- WVRQPPGRGLEWIG YVFYHG---
4fab:  EVKLDETGGGLVQPGRPMKLSCVASGFTFS DYWMN-- WVRQSPEKGLEWVA QIRNKPYNY
1f19:  QVQLKESGAELVAASSSVKMSCKASGYTFT SYGVN-- WVKQRPGQGLEWIG YINPGKG--
b13i:  EVQLVESGGDLVKPGGSLKLSCAASGFTFS RCAMS-- WVRQTPEKRLEWVA GISSGGS--
       |--------|---------|---------| ******* --|---------|- *********
                                       CDR H1                  CDR H2

                70        80        90       100          110         120
       *** ------|---------|---------|---------|-- ***************** |-
glb2:  KTY YAERFKGKATLTADKSSTTAYMGLSSLTSEDSAVYFCAR EIR-------------Y WG
d1_3:  NTD YNSALKSRLSISKDNSKSQVFLKMNSLHTDDTARYYCAR ERDYRL---------DY WG
3671:  YIA YNEKFKGKTTLTVDKSSSTAYMQLRSLTSEDSAVYFCAR SEYYGGSYKF-----DY WG
3d_6:  SIG YADSVKGRFTISRDNAKNSLYLQMNSLRAEDMALYYCVK GRDYYDSGGYFTVAFDI WG

2hfl:  STN YHERFKGKATFTADTSSSTAYMQLNSLTSEDSGVYYCLH GNYDF----------DG WG
3hfm:  STY YNPSLKSRISITRDTSKNQYYLDLNSVTTEDTATYYCAN WDG------------DY WG
2fbj:  TIN YTPSLKDKFIISRDNAKNSLYLQMSKVRSEDTALYYCAR LHYYGYN--------AY WG
2fb4:  DQH YADSVKGRFTISRNDSKNTLFLQMDSLRPEDTGVYFCAR DGGHGFCSSASCFGPDY WG
1fb4:  DQH YADSVKGRFTISRNDSKNTLFLQMDSLRPEDTGVYFCAR DGGHGFCSSASCFGPDY WG
2mcp:  TTE YSASVKGRFIVSRDTSQSILYLQMNALRAEDTAIYYCAR NYYGSTWYF------DV WG
3fab:  TSD TDTPLRSRVTMLVNTSKNQFSLRLSSVTAADTAVYYCAR NLIAGCI--------DV WG
4fab:  ETY YSDSVKGRFTISRDDSKSSVYLQMNNLRVEDMGIYYCTG SYYGM----------DY WG
1f19:  YLS YNEKFKGKTTLTVDRSSSTAYMQLRSLTSEDSAVYFCAR SFYGGSDLAVYYF--DS WG
b13i:  YTF YPDTVKGRFIISRNNARNTLSLQMSSLRSEDTAIYYCTR YSSDPFYF-------DY WG
       *** ------|---------|---------|---------|-- ***************** |-
                                                         CDR H3
```

Figure 2.9: Sequence alignment for antibody crystal structures (heavy chains only) used in the A*b*M algorithm. The CDR's are indicated with stars. The four sequences separated at the top are the antibodies which have been modelled during the development of A*b*M. Gloop-2 was modelled by A.Martin (Martin, 1990)

Residue Number		24	25	26	27	28*	29*
REI	Sequence	Q	A	S	Q	S	I
	ϕ/ψ	-/138	-103/157	-96/7	-158/142	-40/108	-112/9
HyHEL-10	Sequence	R	A	S	Q	S	I
	ϕ/ψ	-/108	-85/135	-88/64	172/160	-64/-38	9/63
Residue Number		30*	31*	32	33	32	
REI	Sequence	I	K	Y	L	N	
	ϕ/ψ	79/-77	-146/21	-104/89	-143/133	-144/-	
HyHEL-10	Sequence	G	N	N	L	H	
	ϕ/ψ	-63/107	85/-15	-105/72	-129/118	-126/-	

Table 2.8: Backbone ϕ andψ angles of residues in CDR-L1 from HyHEL-10 and REI classified in the same canonical group by (Chothia *et al.*, 1989). The residues exhibiting a peptide flip are indicated by a ’*’.

of antibody structures. It should be noted that a single crystal structure will not show multiple conformations since the crystallisation will ‘freeze out’ one of the conformations. During the modelling procedure the two populations of conformers are easily extracted from a set of *ab initio* generated loops, by using a torsional clustering algorithm (see documentation in Appendix B.3).

2.7.2 CDR-L1

In D1.3, all five low energy conformations selected by the EUREKA step (Figure 2.2) were very similar with RMS deviations differing by less than 0.25Å (backbone) and 0.35Å (all atoms). The FILTER algorithm was unable to distinguish between the conformations and the lowest energy structure was selected.

Although CDR-L1 of 3D6 was originally built using the canonical loop from HyHEL-10, the mid-section was rebuilt by conformational search, for the following reason. HyHEL-10 and REI CDR-L1 loops are placed in the same canonical ensemble (Chothia *et al.*, 1989) although they contain a 1–4 shift (peptide flip) relative to one another between the fifth and eighth residues of the loop (residues

28–31) (see Table 2.8).

36–71 shows the same 1–4 shift between the model and crystal structure CDRs. Both crystal structure and model were compared with other loops of the same canonical class as defined by (Chothia *et al.*, 1989). It was found that the hydrogen bonding pattern which determines the conformation was conserved. Thus, the canonical loop method does not discriminate between conformations of this type.

2.7.3 CDR-L2

CDR-L2 of D1.3 has two adjacent threonines (sequence positions 49 and 50) which in the x-ray structure are packed against the Tyr at the fourth position of CDR-H3, thus minimising the exposed hydrophobic sidechains. In the unminimised model the Thr sidechains are exposed to the solvent, but after energy minimisation the correct packing is observed. This CDR is correctly modelled in 3D6 and 36-71.

2.7.4 CDR-L3

In D1.3 and 36–71 the Pro at the seventh position in the loop is correctly predicted in the *cis* conformation. It has previously been suggested that the conformation of CDR-L3 is dictated by the presence of a Pro in position 8 or 9 (Chothia *et al.*, 1989) within the loop. 3D6 does not have a Pro in either position. Only 7 out of 290 CDR-L3 sequences (Kabat *et al.*, 1992) lack a Pro at both positions and in all of the published x-ray structures this Pro is present. This is an example of a situation where either a new canonical class may need to be defined or where the canonical rule breaks down altogether, and an alternative method must be employed.

The 3D6 L3 loop is 7 residues in length and was built using database loops alone where conformational space is saturated by means of fragments selected from the crystallographic database (Global RMS: 2.01 Å, N,Cα,C), and by using CAMAL (Construction: Q[Q(YNS)Y]S, Global RMS: 1.97Å, N,Cα,C). The similarity of the structures generated by the two procedures illustrates the utility of the database search and suggests that for shorter loops it is capable of saturating the available conformational space.

2.7.5 CDR-H1

The Kabat and Wu definition of CDR-H1 places this loop as an extension of the β-sheet. The extended nature of this stretch of peptide limits its conformational flexibility and CDR-H1 is generally modelled accurately (Martin *et al.*, 1989; Chothia *et al.*, 1989).

In D1.3, the Phe or Tyr sidechain at the second position in the loop is poorly placed and packs against Leu at the penultimate position in HFR1 (see Figure 2.9). 36–71 has a well-placed Asn at this position, rather than the more common bulky hydrophobic sidechain.

2.7.6 CDR-H2

CDR-H2 of 36-71 is similar in sequence to R19.9 (Strong *et al.*, 1991), (36–71: YNNPGNGYIA; R19.9: YINPGKGYLS). While the structurally determining residues specified by (Chothia *et al.*, 1989) are conserved, the backbone conformations are different: R19.9 has a bulge at the -PGN- Gly, compared with 36–71, giving the loop a 'kink' in the middle. The model of 36–71 shows a 1–4 shift, though the sidechains are still well placed.

2.7.7 CDR-H3

Problems and analysis CDR-H3 is the most variable of the six CDR's with all lengths up to 21 residues being represented in the database of (Kabat *et al.*, 1992). This extreme variability results from V–D–J splicing (Schilling *et al.*, 1980) and has always been a problem when attempting to model antibodies. Such loops may be divided into short (up to 7 residues), medium (up to 14 residues) and long (15 or more residues). Using the CAMAL procedure, we are now confident that short and medium CDR-H3's can be modelled as accurately as other CDR's of similar lengths. Although long CDR-H3's are more difficult and cannot, at present, be built to the same accuracy, the chain trace is still essentially correct.

It is unlikely that the longer loops consist of 'pure' loops (i.e. all random coil or turn). In crystal structures of antibodies with medium to long CDR-H3 loops (McPC603 (Rudikoff *et al.*, 1981): 11 amino acids (aa); KOL (Marquart *et al.*, 1980): 17 aa; R19.9 (Lascombe *et al.*, 1989): 15 aa) the loops consist of a disordered β-sheet extension from the β-barrel core and a 5–8 residue random coil/turn connecting these two strands.

To determine the nature of medium to long loops ($>$ 8 residues) which satisfy the CDR-H3 constraints, a complete search of the Protein Databank for loops of length 8–20 residues, was performed using the inter-Cα distance constraints

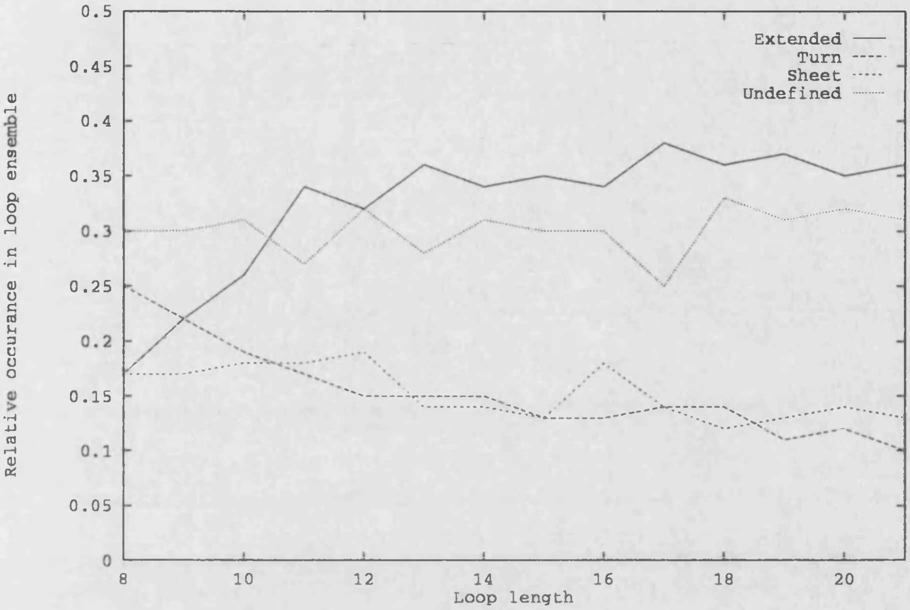


Figure 2.10: Relative distribution of secondary structure in CDR H3 loop ensemble obtained using constraints calculated from known F_V structures. The secondary structure is calculated using the DSSP (Kabsch and Sander, 1983) program. Calculations were not done for loops shorter than eight residues, due to loss of information caused by chain termini (No assignment possible).

determined from known antibody crystal structures for CDR-H3. The resulting loops were then analysed using the DSSP (Kabsch and Sander, 1983) program, which is able to assign secondary structure to polypeptide structures. The amount of secondary structure for each length of loop was calculated (Figure 2.10), and it was observed that for loops longer than 12 residues the amount of secondary structure within each of the classes described in DSSP was constant. The number of loops selected is also constant (approx 150 loops) for loops longer than 12 residues. A closer inspection of each of the length ensembles shows indeed that the loop are the same between the groups.

This analysis shows (Figure 2.10) that, like the long CDR-H3 crystal structures, the selected fragments consist of β-strands connected by 5–8 residue loops. We find that for loops above 12–13 residues in length, the same loops are selected, but with extensions to the β-strands. This is termed the "sliding-ladder" effect. In addition, the maximum size of a random coil or turn fragment in any of the structures contained in the Protein Databank tends not to exceed 8 residues, as determined by DSSP. This implies that the conformational space of longer loops is not saturated by the database and, although it is unlikely that long loops in antibodies will differ significantly from long loops in other structures, confidence in the prediction must be correspondingly lower.

By how much is the usefulness of the CAMAL algorithm reduced by this observation ?

The frequency of occurrence of different CDR-H3 lengths in antibody sequences described by Kabat *et al.* (Kabat *et al.*, 1992) was analysed. The distribution plot in Figure 2.3 shows that more than 85% of H3 loops have lengths between 4 and 14 residues which can be modelled accurately by the CAMAL algorithm.

Modelling results CDR-H3 of D1.3 is of average length (8 residues), though no loops of this length are seen in the available antibody structures. The crystal structure coordinate set showed an RMS of 1.9Å compared with the model.

The 36–71 loop is 12 residues long. The conformation is correctly predicted as a short loop connecting an extension of the β-sheet.

The 3D6 H3 loop is 17 residues long. While KOL (Marquart *et al.*, 1980) has the same length it has only one residue in common with 3D6 and only one conservative mutation. There is thus no reason to believe that the conformations would be similar. The final predicted conformation of 3D6 is an extended β-sheet, as in the crystal structure. The difference between the predicted and the crystal structure of 3D6-H3 is due to a twist of 5-7° in the extended β-sheet loop (see Figures 2.11 & 2.12). Such a twist has also been observed for complexed and uncomplexed antibodies by Rini *et al* (1992). This suggests that long CDR-H3 loops may be flexible and actively involved in antigen binding. Thus, attempting to assign a single conformation to such loops may be meaningless

2.7.8 The complete variable region - Summary of results

Prediction of the strand positions and V_L/V_H orientation in the framework β-barrel was exact for all the three antibodies. The backbone (N,Cα,C) RMS deviations from the crystal structures were between 0.56 and 0.86 Å, despite the fact that in all cases the V_L and V_H regions of a particular model were derived from different antibody structures. This suggests that this method will do well in procedures such as humanisation (Gorman *et al.*, 1991), where correct framework positioning is important. The backbones of all six CDRs in all three antibodies are essentially correctly predicted, as shown in Figure 2.7. There are two important points to make about these predictions. First, the position of each CDR on its framework barrel is correct. Thus, CDR-framework interactions can be confidently monitored. The only deviation from the x-ray structure is CDR-H3 of antibody 3D6 which has been discussed above. Second, the all atom RMS deviation between models and x-ray structures is dominated by sidechain positions. In most instances this deviation is due to a small number of incorrectly positioned, exposed sidechains (for example in D1.3 the only sidechains which are incorrectly predicted are Tyr 9 of L1, Trp 4 of L3, Tyr 2 of H1 and Tyr 4 of H3). Since each CDR is constructed in the absence of other CDRs, the forcefield may choose a rotamer which is 120° away from that found in the crystal structure.

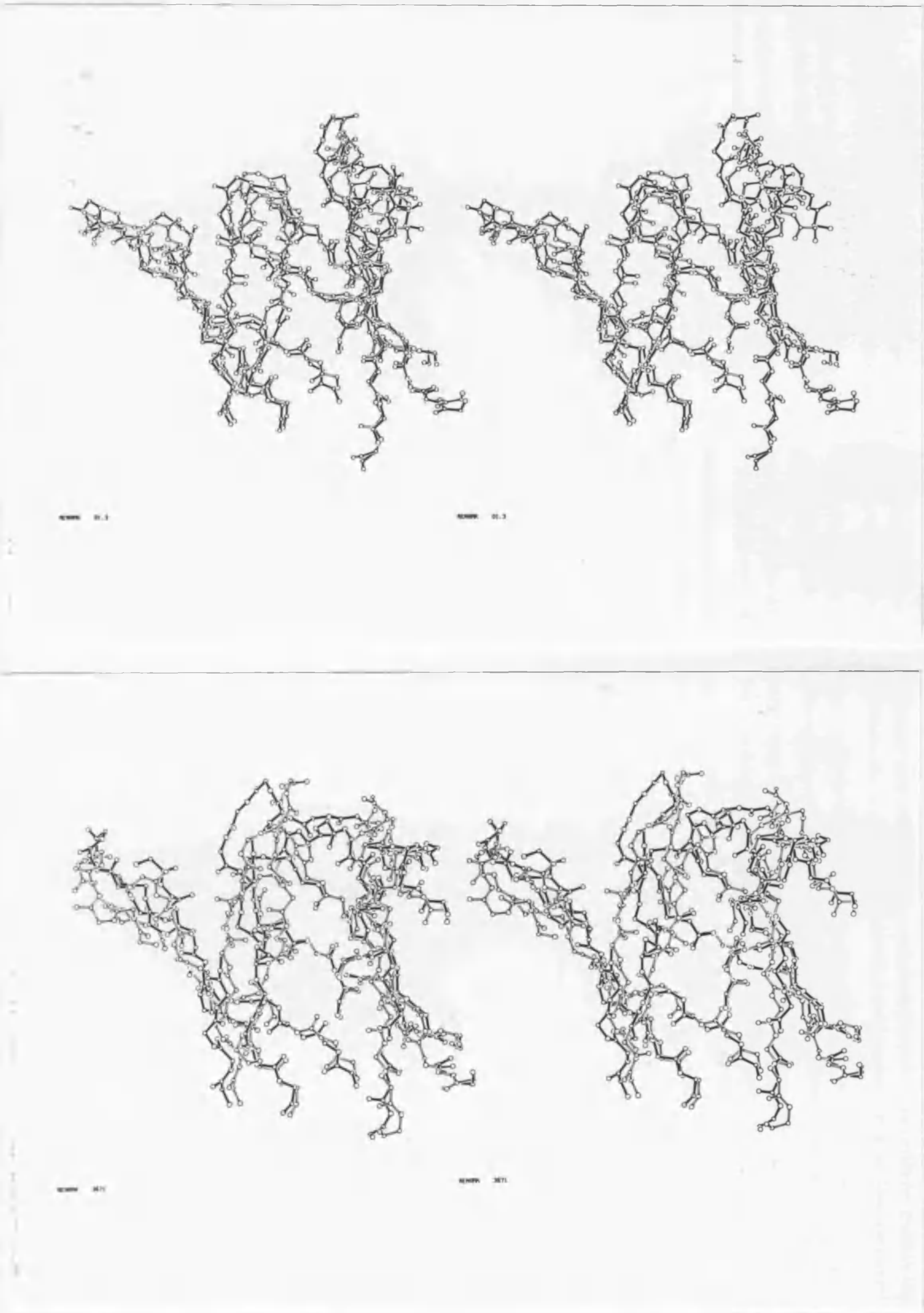


Figure 2.11: Stereo (N,C-α,C,O) representation of crystal structures and models of **D1.3 and 3671** variable domain and β-barrel strands . Crystal structure are shown with open bonds, model with solid bonds. Top: D1.3, Bottom: 36-71

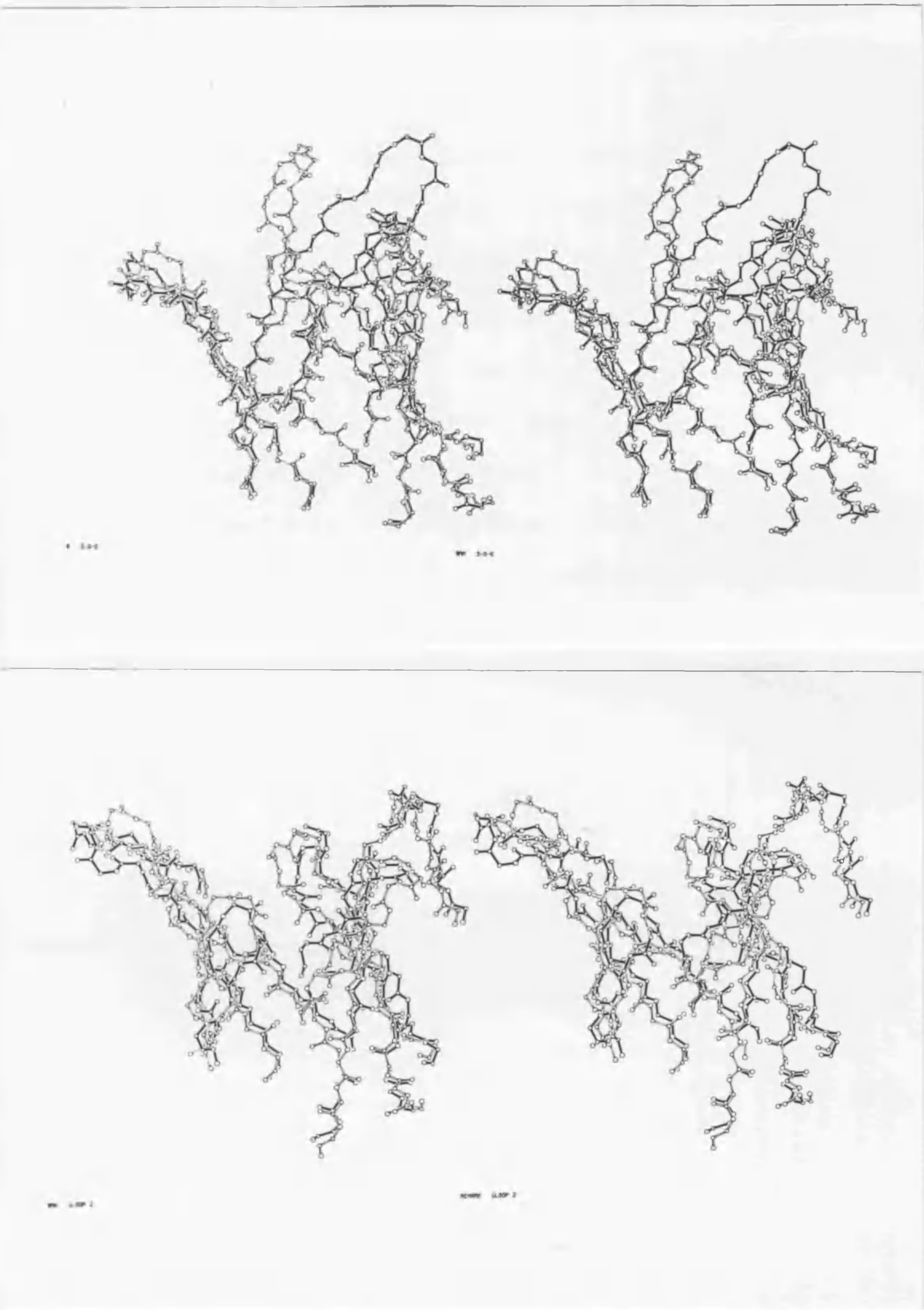


Figure 2.12: Stereo (N,C-α,C,O) representation of crystal structures and models of **3D6 and Gloop-2** variable domain and β-barrel strands . Crystal structures are shown with open bonds, model with solid bonds. The difference between the 3D6-H3 in the model and the crystal structure is due to a 5-7° twist in the extended β sheet conformation of this loop. Top: 3D6, Bottom: Gloop-2

.

This effect has also been observed by (Lee and Levitt, 1991).

A present limitation of the A*b*M algorithm is the assumption that all CDR's can be predicted independently of any of the other loops. Modelling all the loops independently works well when the antibody being modelled (e.g. Gloop-2) has short CDR loops. When modelling antibodies with longer CDR loops (e.g. 36-71) the effects of other CDR's should ideally be taken into account. In the antibody 36-71 a number of clashes are observed as a result of modelling the loops independently. CDR L1 is modelled such that it overlaps with L2. Since canonical loops can usually be defined for at least five of the six CDRs these could placed in the combining site before the remaining loops are modelled. Using this protocol conformational space will be appropriately limited during the *ab-initio* search and progression of error through the model will be minimised.

When predicting sidechain conformations using the Monte Carlo method it is necessary to have a model where the backbone has been predicted with high confidence. If the backbone is not well defined the position of all the sidechains can be wrong as they are generated all at once. This problem is avoided when generating sidechains in CONGEN/CAMAL since the CDR's are modelled independently. Thus, the choice of sidechain construction method will be dictated by the confidence level for the backbone construction.

2.8 Antibody modelling: further developments

Recently (October 1992), a commercially available version (OML, 1992) of the A*b*M program has become available. In order to get a wider view of how the algorithms in the program suite perform, all the antibody crystal structures available in the crystallographic database (Bernstein *et al.*, 1977) were modelled (Results shown in Table 2.9 and in the complete Table of RMSD values in Appendix A.3). The data presented here is the result of a joint effort between S.M.J.Searle and the author.

The main problem when modelling complete combining sites using A*b*M is the determination of the takeoff angles for the loops. As shown in the Tables in Appendix A.2 there is up to 90° variation between the takeoff angles of CDR H3 in different crystal structures. The overlap and framework selection algorithm has therefore been modified in A*b*M. Three structural classes of CDR H3 loops have been defined, using the table for CDR H3 takeoff angles in Appendix A.2. The structures 2hfl and 1f19 are different with respect to takeoff angles to the remaining structures. The difference appears to be due to a structural residue position at the C-terminal end of the H3 loop. Most structures have a conserved Tyr or Val at the C-terminal position of the loop. However, 2hfl and 1f19 have a Gly and Ser respectively at this position, resulting in a kink in the loop, and a resulting change in takeoff angle. The result of this observation and the fact that there appears to be two populations of loop takeoff angles, depending on the CDR length, lead to the definition of three classes of H3 loops (see also Figure 2.13:

- Loops shorter than seven residues.
- Loops equal to or longer than seven residues.
- Loops which do not have a structural Val or Tyr at the penultimate position of the loop sequence.

The framework of the heavy chain is therefore selected, in A*b*M, on the basis

Structure	CDR	CDR Length	Global RMSD (N,Cα,C,O)
glb2	L1	11	1.161
	L2	7	0.647
	L3	9	1.031
	H1	5	1.785
	H2	10	1.609
	H3	4	1.273
	Total		1.251
2hfl	L1	10	1.150
	L2	7	0.712
	L3	8	2.524
	H1	5	1.261
	H2	10	2.155
	H3	7	2.310
	Total		1.685
2mcp	L1	17	0.784
	L2	7	0.538
	L3	9	0.739
	H1	5	1.004
	H2	12	2.014
	H3	11	2.306
	Total		1.231
4fab	L1	16	2.470
	L2	7	0.792
	L3	9	1.255
	H1	5	0.721
	H2	12	2.028
	H3	7	2.132
	Total		1.566
3hfm	L1	11	0.775
	L2	7	1.021
	L3	9	0.394
	H1	5	2.012
	H2	9	0.942
	H3	5	1.683
	Total		1.138
lmam	L1	11	1.302
	L2	7	1.362
	L3	9	1.289
	H1	5	1.845
	H2	12	2.976
	H3	8	2.524
	Total		1.883
b13i	L1	16	2.667
	L2	7	0.763
	L3	9	0.877
	H1	5	1.310
	H2	10	1.202
	H3	10	2.970
	Total		1.632
d1.3	L1	11	0.799
	L2	7	0.928
	L3	9	1.138
	H1	5	0.846
	H2	9	1.413
	H3	8	2.188
	Total		1.219

Table 2.9: Modelling for some of the crystal structures with CDR H3 length less than twelve residues. The complete data set for all sixteen antibodies in the crystallographic database can be found in Appendix A.3.

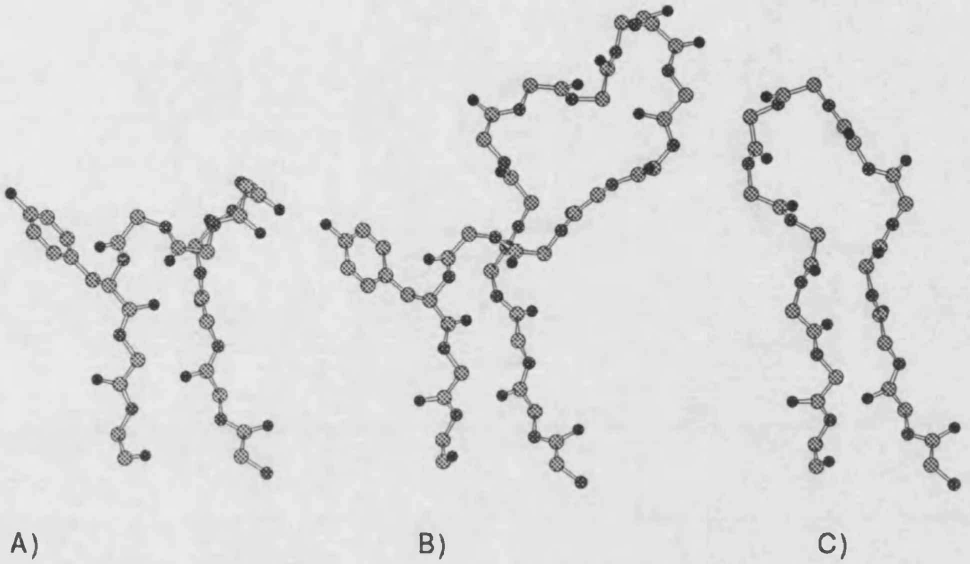


Figure 2.13: Structural classes of CDR H3 as defined in the text. Here they are illustrated by the three structures A) 3hfm B) 8fab C) 2hfl. The structural Tyr at at the penultimate position is shown, the structure 2hfl (C) has a Gly at this position.

of the most homologous CDR H3 with respect to the above classes, and *not* on the basis of the complete heavy chain sequence. This dramatically improves the quality of the final conformation of longer (> 10 residues) CDR H3s (data using the original framework selection method are not shown).

Chapter 3

A new method of humanisation: resurfacing

3.1 Antibody fragments and their properties

Recently a large interest has been shown in the reshaping (**humanisation**) of non-human antibodies (Winter and Milstein, 1991; Lewis and Crowe, 1991) in order to make these non-immunogenic in man. These reshaped antibodies are then used as therapeutic drugs in the treatment of diseases (Reichman *et al.*, 1988). The main theme in the development of antibodies as drugs has been the reduction of the size of the antibody to obtain a minimal recognition unit (MRU). Smaller compounds are more easily transported across cell membranes and tissue barriers. Figure 3.1 shows how these fragments are derived from the native antibody.

Chimeric antibodies are antibodies with variable domains from a rodent (usually mice) antibody, and constant domains from a human antibody. In these chimeric antibodies the presence of the murine variable region framework often leads to

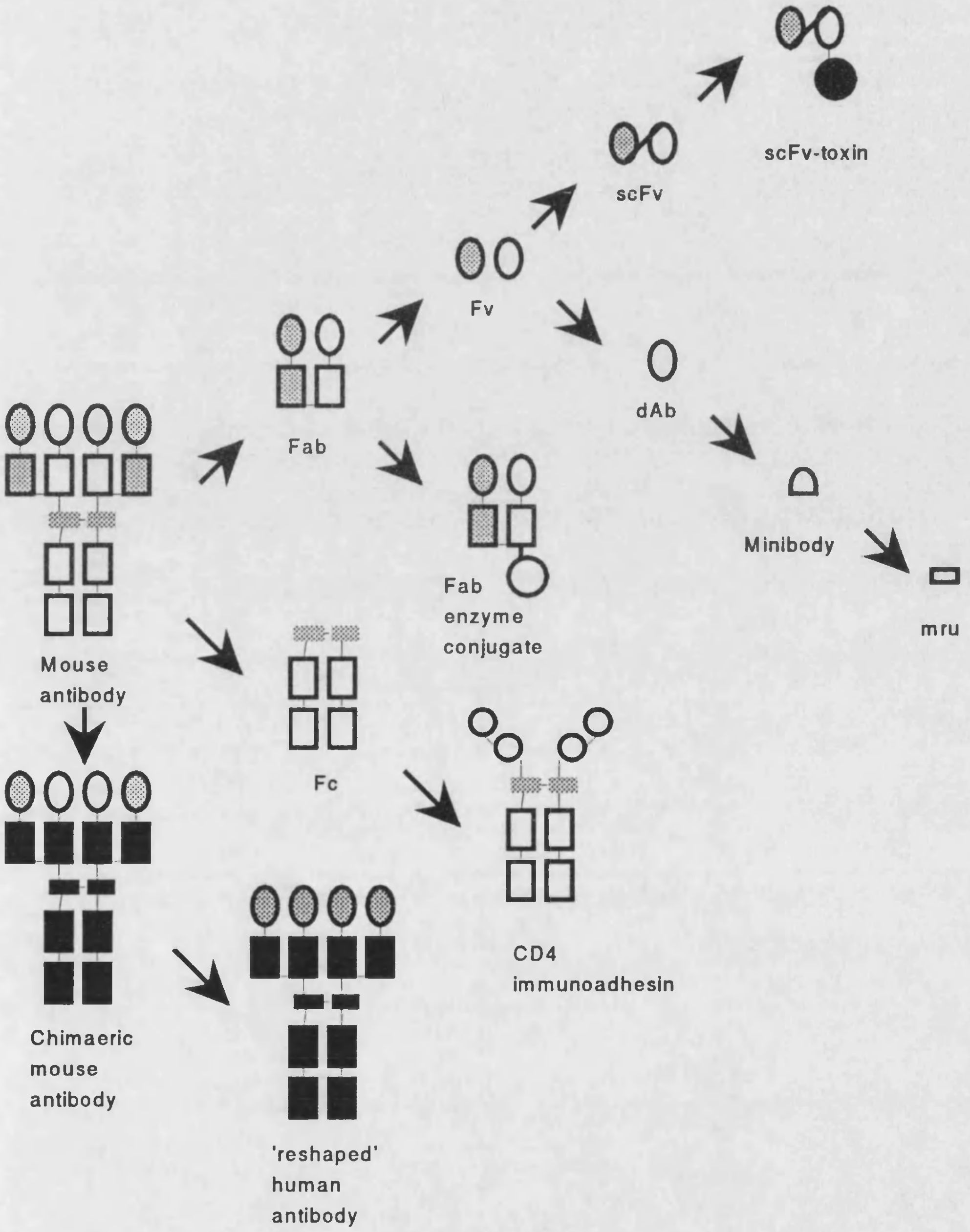


Figure 3.1: Various antibody fragments and engineered antibodies which have been reported in the literature. Each box or circle represents a protein domain. The various fragments are described in the text. (Reproduced after (Winter and Milstein, 1991))

immunogenicity (Lobuglio and Saleh, 1992). This can be overcome by grafting only the CDR loops from the original mouse antibody onto the human antibody (see later) (Hale *et al.*, 1991; Verhoeyen *et al.*, 1991; Verhoeyen *et al.*, 1988; Kyle *et al.*, 1991; Crowe *et al.*, 1992).

F_C domains from mice have been linked to receptor specific molecules such as CD4. This conjugate binds to protein **gp120** of the human immunodeficency virus **HIV** on the surface of infected cells and kills the infected cells by antibody dependent cell-mediated cytolysis (Byrn *et al.*, 1990).

F_{AB} fragments have been used in many ways both as therapeutic agents and as diagnostics. The smaller fragments are more attractive to use *in vivo* since they have a higher capability to penetrate tissue boundaries, and are cleared faster from the blood stream. F_{AB} fragments and other small antibody fragments conjugated to cell toxins are frequently called "magic bullets", since these can be used to specifically target disease areas in the living organism, such as cancer tumors (Reichman *et al.*, 1988). F_{AB} fragments are also used for clearing toxic drugs, such as digoxin (Wenger *et al.*, 1985), from the blood stream. *In vitro*, F_{AB} fragments are conjugated to enzymes and used in ELISA (enzyme linked immuno sorbent assay) (Engvall and Pesce, 1978). In these assays the F_{AB}-enzyme conjugate is bound to immobilised antigen and the amount of antigen is then determined by performing an enzyme specific reaction (usually colourimetric) with the F_{AB} bound enzyme. The extent of this reaction is related to the amount of antigen initially bound (Engvall and Pesce, 1978).

The smallest fragment of an antibody which still contains the complete binding domain is the F_V fragment. F_V fragments have been used in the same way as F_{AB} fragments as conjugates. Single domain antibodies (**dAb**) and even single

CDR's have been shown to bind to antigens to which the original antibody was raised (Ward *et al.*, 1989; Taub *et al.*, 1989). In order to make dAb's useful it may be necessary to make large alterations to the surface of the domain in order to gain solubility, since these fragments have the V_L/V_H interface exposed to the aqueous surroundings. Tramontano has engineered a 60 amino acid sub-fragment (**minibody**) of the heavy chain to produce a more soluble recognition unit (Tramontano, 1992). More promising are single chain F_V fragments (**scF$_V$**). In scF$_V$'s the C-terminus of the light chain is linked to the N-terminus of the heavy chain, usually through a hydrophilic poly-Ser-Gly linker (Bird *et al.*, 1988b). Single chain F_V fragments have been bound to cell toxins such as ricin (See (Pimm, 1988; Bagshawe, 1987) for review). These antibody-fragment conjugates are in development in many Biothecnology companies for cancer treatment.

Immunomimetics are compounds which are derived from a known antibody structure and which resembles the action of of the antibody. The scope of immunomimetics is to avoid the inherent disadvantage of proteins. Proteins are degraded fast and rapidly cleared from the blood stream. This is not desirable in a clinical situation where it is necessary to have longer retention times although this may be a useful property for imaging of tissue targets where background signals from bound antibody needs to be as low as possible. The first immunomimetics were peptides derived directly from CDR sequences in antibodies, and were cyclised MRU's (Taub *et al.*, 1989; Bruck *et al.*, 1986; Kang *et al.*, 1988; Williams *et al.*, 1991; Novotny *et al.*, 1986; Williams *et al.*, 1989b; Williams *et al.*, 1989a). Since then more advanced cyclical peptide compounds have been derived from antibody structures. Sargovi *et al* synthesised a β-turn peptide, which was derived from the model of an antibody combining site of an anti-retro virus type 3 cellular receptor (Reo3R). They used accessibility as the selection criteria, and rationalised that the most exposed CDR was the most likely interaction site of the antibody with its antigen. These smaller more rigid peptides, frequently contain modified amino-acids (D- amino acids, etc), which make them less prone to proteolytic degradation.

Several studies are concerned with the change of single residues in order to increase specificity or affinity of a given antibody or antibody fragment (Roberts *et al.*, 1987; Winter and Milstein, 1991). Metal binding sites, and catalytic triads of proteases have been engineered into antibodies (Tainer *et al.*, 1985; Gregory *et al.*, 1990). **Catalytic antibodies** have also been produced which are antibodies induced by immunisation with transition state analogues. Numerous examples have been reported where diverse reactions have been catalysed (see (Baum, 1991; Benkovic *et al.*, 1991; Gibbs *et al.*, 1991; Ikeda *et al.*, 1991; Jackson *et al.*, 1991; Khalaf *et al.*, 1992; Lerner *et al.*, 1991; Martin *et al.*, 1991c; Martin *et al.*, 1991b; Sastry *et al.*, 1991; Shokat and Schultz, 1991; Suckling *et al.*, 1992). However, it is not within the scope of this thesis to further describe these.

3.2 Humanisation of variable regions

The large interest in reshaping murine antibodies has been spawned by the large therapeutic potential of humanised antibodies (Hale *et al.*, 1991; Verhoeyen *et al.*, 1991; Verhoeyen *et al.*, 1988; Kyle *et al.*, 1991; Crowe *et al.*, 1992).

The first attempt to humanise a murine antibody was performed by Reichmann *et al* (1988). The CDRs from a rat antibody directed against human lymphocytes were grafted onto the framework of the human heavy chain of NEW (Saul *et al.*, 1978) and the light chain of REI (Palm and Hilschmann, 1975). These antibodies were capable of activating complement and thus mediating cell lysis (Reichman *et al.*, 1988). In order to regain the activity of the original antibody additional single residue changes had to be made in the human framework in order to restore the correct environment for the murine CDRs.

This process of "back mutations" has been necessary in virtually all the reported cases of reshaping (Hale *et al.*, 1991; Verhoeyen *et al.*, 1991; Verhoeyen *et al.*, 1988; Kyle *et al.*, 1991; Crowe *et al.*, 1992; Kettleborough *et al.*, 1991; Reichman *et al.*, 1988). Kettleborough identified important CDR interacting residues in the framework of the F_V region, using a molecular model of the murine antibody. In order to test the importance of these residues nine versions of the humanised antibody were produced. In this case the best construct only retained 60 % the avidity of the original antibody (Kettleborough *et al.*, 1991).

This loss of binding justifies the need for finding new ways of reshaping murine antibodies which avoid extensive changes to the framework region adjacent to the CDRs.

In this chapter it is shown how an F_V surface can be changed, retaining the specificity of the combining site, proving that major changes can be made to the F_V structure without changing its binding functionality.

3.3 Variable region surfaces

Several attempts have been made to rationalise, and explain the differences between human and murine antibody F_V domains (Arnold *et al.*, 1991; Strohal *et al.*, 1989; Zachau, 1990).

Recently Schroeder and colleagues have performed a thorough analysis of V_H mammalian germline sequences (Schroeder *et al.*, 1989) in order to determine regions of conservation. They identified a set of conserved regions in the sequences which are located on a solvent exposed face of the V_H chain. A similar analysis was performed by Kroemer *et al* for light chain V_L region sequences (Kroemer *et al.*, 1991). These phylogenetic studies pinpoint the divergent evolution of the human and murine immunoglobulin sequences, but do not clearly identify the different conserved regions in the two families.

An attempt to locate the conserved, exposed regions in human and murine antibodies has been presented by Padlan. He calculated the accessibility of the crystal structure of one human and one murine antibody (Padlan, 1991). Using an accessibility criteria the exposure of surface positions was determined. Although this study did not present a general algorithm there appeared to be differences in the presentation of surface residues of murine and human germline antibodies.

In order to more exhaustively characterise the surface of different V-regions a statistical analysis of antibody surface residues was carried out which has lead to a novel method for the reshaping of murine antibodies. The method is termed **Resurfacing**.

3.3.1 F_V surface analysis

In order to determine the amino acid positions which are usually accessible on the surface of the F_V domain, the accessibility was calculated for twelve F_{AB} x-ray crystallographic structures obtained from the Brookhaven database (Bernstein *et al.*, 1977). The relative accessibility was calculated using the program MC

(Appendix B.1), which implements a modified version of the DSSP (Kabsch and Sander, 1983) accessibility calculation routine in which explicit atomic radii are employed. Here the relative surface accessibility is defined as the accessibility of a given residue in the protein divided by the accessibility of the same residue in the same conformation but in a free blocked amino acid. A residue was defined as being surface accessible when the relative accessibility was greater than 30 %. Surface accessible positions of framework amino acids constitute 40 % of the F_V surface area. The remaining surface accessible residues are in the CDRs and in the interdomain C-terminal region. The Figures 3.3 and 3.4 show a sequence alignment of the twelve crystal structures, the average relative accessibility, and the 30 % accessibility cut-off.

The surface accessible framework positions were mapped onto a database of unique human and mouse F_V sequences . The frequency of particular residues in each of these positions is shown in Table 3.1 & 3.2. Only residue frequencies higher than 5 % are listed.

The justification for using a 30 % cut-off was tested by calculating the solvent accessibility of all the residues in hen egg lysozyme (Figure 3.2). The epitopes for four antibodies, HyHEL5 (Sheriff *et al.*, 1987), HyHEL10 (Padlan *et al.*, 1989), D1.3 (Amit *et al.*, 1986), and Gloop2 (Rees *et al.*, 1989), all of which were determined by either x-ray crystallography or NMR, are indicated. The 30 % cut-off position is also shown. The data show that the epitope residues in all four antibodies are included in the 30 % surface residue set and that residues below 30 % are largely inaccessible to the antibodies.

There are three major points to be made from the frequency data i Table 3.1 and 3.2.

Position	Human	Mouse
1	D 54 E 33	D 76 Q 9 E 7
3	V 39 Q 25 S 22	V 62 Q 22
5	T 66 L 33	T 87
9	S 31 P 23 G 15 A 15	S 37 A 28 L 17
15	P 59 V 24 L 15	L 46 P 33 V 9
18	R 57 S 18 T 12	R 38 K 22 S 14 Q 12 T 10
26	S 71 T 12	S 93
27	Q 57 S 24	Q 52 S 16 K 12 E 10
28	S 64 D 8 G 7	S 59 D 19 G 9
46	P 92	P 82 S 9
47	G 87	G 72 D 18
51	K 49 R 28	K 71 Q 13 R 8
62	S 58 T 23	S 71 P 7 D 7
63	G 91	G 96
66	D 41 S 27 A 9	D 37 S 27 A 26
73	S 97	S 91
76	D 44 S 18 T 18 E 14	D 65 S 16
86	P 42 A 29 S 17	A 49 P 12 S 9 T 8
87	E 73 D 12	E 91
108	G 44 Q 37 T 7	G 57 A 20 S 12
111	K 78 R 11	K 92
115	K 55 L 40	K 84
116	R 63 G 32	R 85 G 10
117	Q 48 A 21 T 18	A 68 Q 20

Table 3.1: Table of residue frequencies in surface positions of sequence alignment of Kabat (Kabat *et al.*, 1992) database of light chain sequences. Only residue types which occur with a higher frequency than 5 % are listed. Sequence numbering is the same as in Figure 3.10

Position	Human	Mouse
118	E 48 Q 45	E 59 Q 29 D 10
120	Q 81 T 6	Q 68 K 25
122	V 54 Q 15 L 14	Q 57 V 27
126	G 53 A 23 P 19	G 35 P 30 A 28
127	G 54 E 24 A 12	E 45 G 43
128	L 65 V 27 F 7	L 95
131	P 93	P 90
132	G 71 S 18 T 7	G 81 S 17
133	G 38 E 16 Q 14 R 12 A 10	A 34 G 29 Q 16 S 9
136	R 49 K 25 S 18 T 7	K 64 S 17 R 14
143	G 95	G 98
145	T 47 S 32 N 8	T 62 S 19 N 7
159	A 54 P 21	R 36 S 15 T 12 P 12 F 11 A 8
160	P 84 S 10	P 89 H 6
161	G 93	G 72 E 23
173	S 27 K 15 G 13 D 11	G 36 K 14 S 13 N 11 D 10 Y 7
174	G 48 D 14 S 13	G 31 N 23 S 19 A 9
183	D 25 P 24 A 16 Q 9 T 7	E 31 P 21 D 17 Q 11 A 11
184	S 68 K 10	K 42 S 37
186	K 57 Q 19 R 7	K 82 Q 6
187	G 65 S 22	G 61 S 18 D 10
195	T 32 D 24 N 19 K 8	T 35 K 29 N 26
196	S 89	S 75 A 16
197	K 63 T 7 I 7	S 46 K 34 Q 10
208	R 44 T 22 K 15	T 55 R 26 K 8
209	A 48 P 19 S 16 T 9	S 66 A 15 T 11
210	E 49 A 18 D 12 S 8	E 87 D 7
212	T 85	T 53 S 43
222	G 17 D 11 P 10 Y 9 V 8 N 8	D 67 A 18

Table 3.2: Table of residue frequencies in surface positions of sequence alignment of Kabat (Kabat *et al.*, 1992) database of heavy chain sequences. Only residue types which occur with a higher frequency than 5 % are listed. Sequence numbering is the same as in Figure 3.10

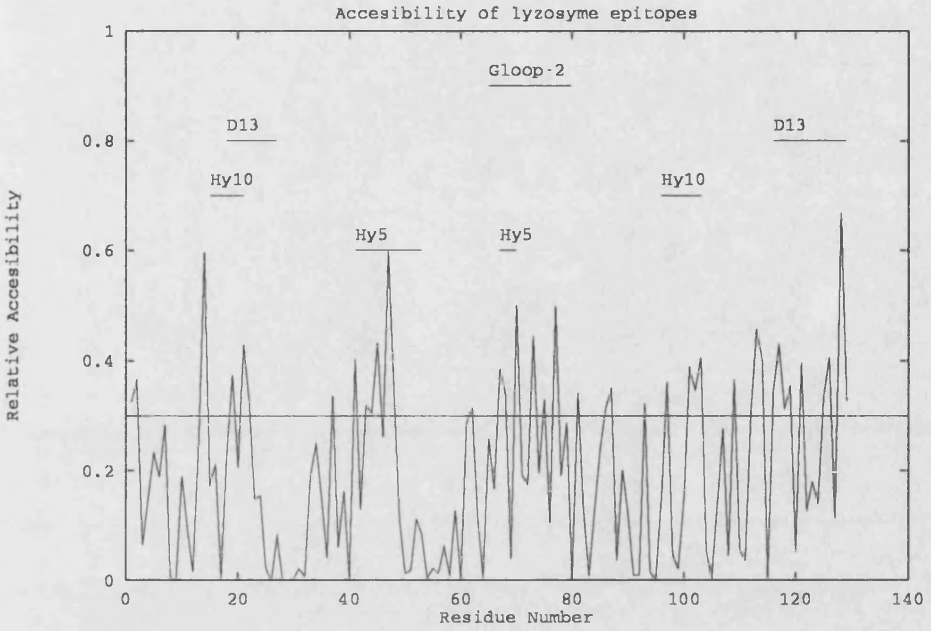


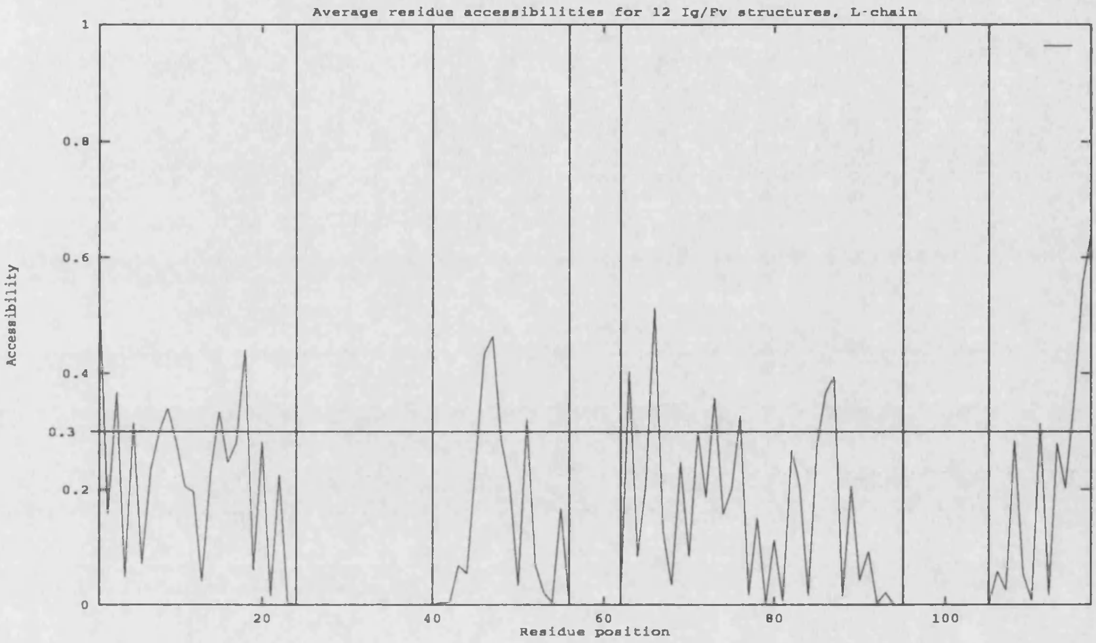
Figure 3.2: Accessibility plot of hen egg lysozyme (Moult *et al.*, 1976). The accessibility was calculated as described in the text. The epitopes of the four antibodies HyHEL5, HyHEL10, D1.3 and Gloop2 are shown with bars. The 30 % cut-off is also shown.

1. The residue frequencies at particular positions of the sequence are largely conserved.

2. At the amino acid positions identified by the above analysis, none of the entire combinations of surface residues in the human sequences are found in the murine sequences and *vice versa.*

3. Only at two of the surface positions are different distributions of amino acids found. At position 5 of the light chain Leu is found in 33 % of the sequences while only Thr is found in the mouse sequences. At position 159 of the heavy chain Arg is found in 36 % of the mouse sequences, but in none of the human sequences where it is an Ala or a Ser residue.

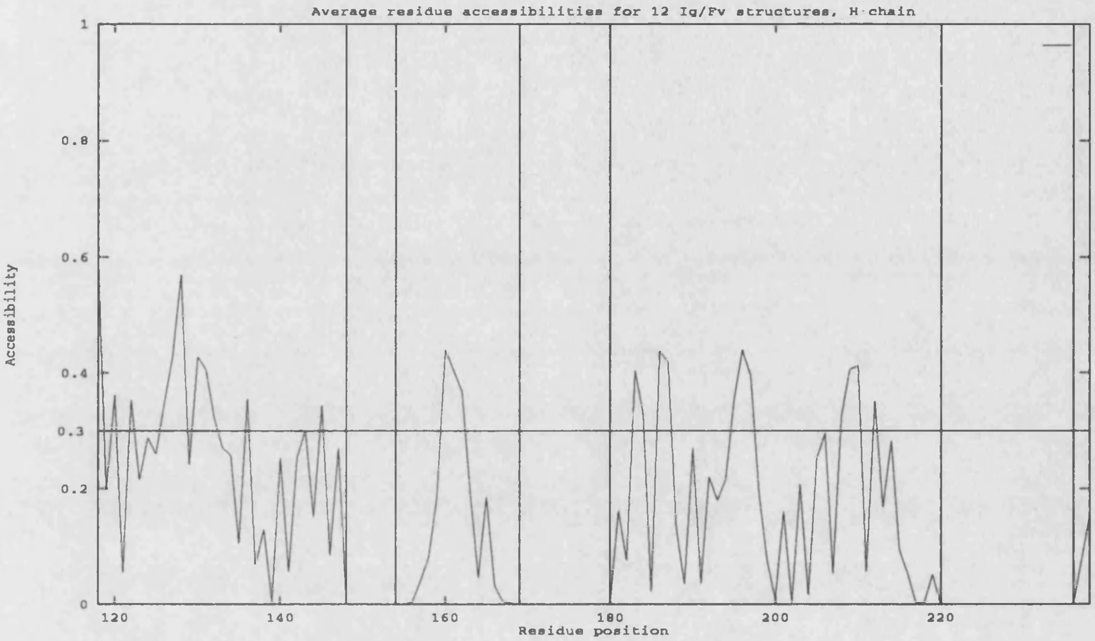
In order to determine whether the mouse sequences are more distantly related to human F_V sequences than to other mouse F_V sequences, the identity was calculated between all the sequences in a pool of both human and mouse sequence patches



```
               LFR 1           CDR L1           LFR 2       CDR L2           LFR 3              CDR L3    LFR 4

     | | |   |     |  |     *****************     ||   |    *******|  |      |  |         ||      ************     |  |||
glb2 DIQMTQSPSSLSASLGERVSLTCRASQEISG------YLSWLQQKPDGTIKRLIYAASTLDSGVPKRFSGRRSGSDYSLTISSLESEDFADYYCLQYLS--YPLTFGAGTKLELKRA
1fdl DIQMTQSPASLSASVGETVTITCRASGNIHN------YLAWYQQKQGKSPQLLVYYTTTLADGVPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWS--TPRTFGGGTKLEIKRR
2hfl DIVLTQSPAIMSASPGEKVTMTCSASSSVN-------YMYWYQQKSGTSPKRWIYDTSKLASGVPVRFSGSGSGTSYSLTISSMETEDAAEYYCQQWGR--NP-TFGGGTKLEIKRA
3hfm DIVLTQSPATLSVTPGNSVSLSCRASQSIGN------NLHWYQQKSHESPRLLIKYASQSISGIPSRFSGSGSGTDFTLSINSVETEDFGMYFCQQSNS--WPYTFGGGTKLEIKRA
2fbj EIVLTQSPAITAASLGQKVTITCSASSSVSS-------LHWYQQKSGTSPKPWIYEISKLASGVPARFSGSGSGTSYSLTINTMEAEDAAIYYCQQWTY--PLITFGAGTKLELKRA
2fb4 ESVLTQPPSASG-TPGQRVTISCTGTSSNIG---SITVNWYQQLPGMAPKLLIYRDAMRPSGVPTRFSGSKSGTSASLAISGLEAEDESDYYCASWNSSDNSYVFGTGTKVTVLGQ
2mcp DIVMTQSPSSLSVSAGERVTMSCKSSQSLLNSGNQKNFLAWYQQKPGQPPKLLIYGASTRESGVPDRFTGSGSGTDFTLTISSVQAEDLAVYYCQNDHS--YPLTFGAGTKLEIKRA
3fab -SVLTQPPSVSG-APGQRVTISCTGSSSNIG---AGNHVKWYQQLPGTAPKLLIFHNNA-------RFSVSKSGSSATLAITGLQAEDEADYYCQSYDR--SLRVFGGGTKLTVLRQ
4fab DVVMTQTPLSLPVSLGDQASISCRSSQSLVHS-QGNTYLRWYLQKPGQSPKVLIYKVSNRFSGVPNRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTH--VPWTFGGGTKLEIKRA
1f19 DIQMTQTTSSLSASLGDRVTISCRASQDISN------YLNWYQQKPDGTVKLLVYYTSRLHSGVPSRFSGSGSGTDYSLTISNLEHEDIATYFCQQGST--TPRTFGGGTKLEIKRR
5fab DIQMTQIPSSLSASLGDRVSISCRASQDINN------FLNWYQQKPDGTIKLLIYFTSRSQSGVPSRFSGSGSGTDYSLTISNLEQEDIATYFCQQGNA--LPRTFGGGTKLEIKRA
1dfb DIQMTQSPSTLSASVGDRVTITCRASQSISR------WLAWYQQKPGKVPKLLIYKASSLESGVPSRFSGSGSGTEFTLTISSLQPDDFATYYCQQYNS---YSFGPGTKVDIKRT
```

Figure 3.3: Average relative accessibility of F_V **light** chain structures plotted along sequence alignment. The numbering used here is described in Figure 3.10. Sequences are referenced by their Brookhaven entry code (For references see Table 2.2)



```
                 HFR 1             CDR H1      HFR 2     CDR H2                        HFR 3                    CDR H3

     | | |   ||| |||   |      | |  *******     |||      ************* || ||       |||          ||| |       *****************
glb2 QVQLQQSGTELARPGASVRLSCKASGYTFTTFGIT--WVKQRTGQGLEWIGEIFPGNS--KTYYAERFKGKATLTADKSSTTAYMQLSSLTSEDSAVYFCAREIR-------------YWG
1fdl QVQLKESGPGLVAPSQSLSITCTVSGFSLTGYGVN--WVRQPPGKGLEWLGMIWGDG---NTDYNSALKSRLSISKDNSKSQVFLKMNSLHTDDTARYYCARERDYRL---------DYWG
2hfl -VQLQQSGAELMKPGASVKISCKASGYTFSDYWIE--WVKQRPGHGLEWIGEILPGSG--STNYHERFKGKATFTADTSSSTAYMQLNSLTSEDSGVYYCLHGNYDF----------DGWG
3hfm DVQLQESGPSLVKPSQTLSLTCSVTGDSITSDYWS--WIRKFPGNRLEYMGYVSYSG---STYYNPSLKSRISITRDTSKNQYYLDLNSVTTEDTATYYCANWDG------------DYWG
2fbj EVKLLESGGGLVQPGGSLKLSCAASGFDFSKYWMS--WVRQAPGKGLEWIGEIHPDSG--TINYTPSLKDKFIISRDNAKNSLYLQMSKVRSEDTALYYCARLHYYGYN--------AYWG
2fb4 EVQLVQSGGGVVQPGRSLRLSCSSSGFIFSSYAMY--WVRQAPGKGLEWVAIIWDDGS--DQHYADSVKGRFTISRNDSKNTLFLQMDSLRPEDTGVYFCARDGGHGFCSSASCFGPDYWG
2mcp EVKLVESGGGLVQPGGSLRLSCATSGFTFSDFYME--WVRQPPGKRLEWIAASRNKGNKYTTEYSASVKGRFIVSRDTSQSILYLQMNALRAEDTAIYYCARNYYGSTWYF------DVWG
3fab -VQLEQSGPGLVRPSQTLSLTCTVSGTSFDDYYST--WVRQPPGRGLEWIGYVFYHG---TSDTDTPLRSRVTMLVNTSKNQFSLRLSSVTAADTAVYYCARNLIAGCI--------DVWG
4fab EVKLDETGGGLVQPGRPMKLSCVASGFTFSDYWMN--WVRQSPEKGLEWVAQIRNKPYNYETYYSDSVKGRFTISRDDSKSSVYLQMNNLRVEDMGIYYCTGSYYGM----------DYWG
1f19 QVQLKESGAELVAASSSVKMSCKASGYTFTSYGVN--WVKQRPGQGLEWIGYINPGKG--YLSYNEKFKGKTTLTVDRSSSTAYMQLRSLTSEDSAVYFCARSFYGGSDLAVYYF--DSWG
5fab EVQLQQSGVELVRAGSSVKMSCKASGYTFTSNGIN--WVKQRPGQGLEWIGYNNPGNG--YIAYNEKFKGKTTLTVDKSSSTAYMQLRSLTSEDSAVYFCARSEYYGGSYKF-----DYWG
1dfb EVQLVESGGGLVQPGRSLRLSCAASGFTFNDYAMH--WVRQAPGKGLEWVSGISWDSS--SIGYADSVKGRFTISRDNAKNSLYLQMNSLRAEDMALYYCVKGRDYYDSGGYFTVAFDIWG
```

Figure 3.4: Average relative accessibility of F_V **heavy** chain structures plotted along sequence alignment. The numbering used here is described in Figure 3.10. Sequences are referenced by their Brookhaven entry code (For references see Table 2.2)

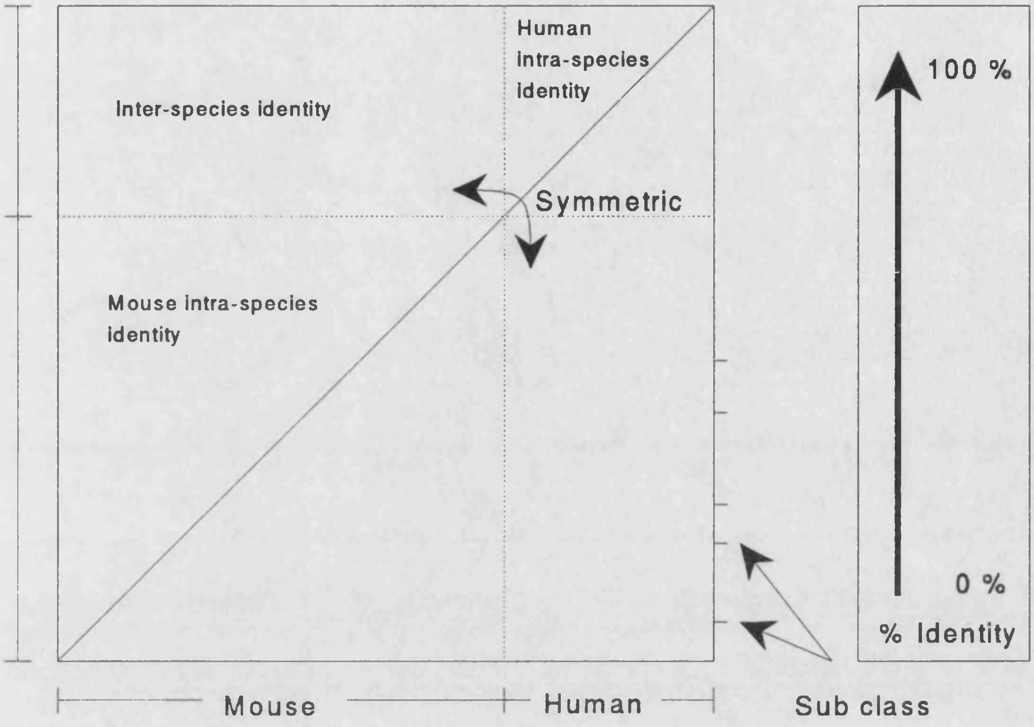


Figure 3.5: Key showing how the density maps in Figures 3.6-3.8 are assembled. The density plot is symmetric about the diagonal axis. F_V sequences are sorted according to species and sub-group along the x and y axis. This gives a clear separation of inter- and intra-species identities. The relative identity is shown by a grey scale. Each point in the plot shows the identity between two sequences in the F_V sequence database. Plots for different residue sets are shown.

made up of the surface accessible residues. The sequences are plotted against each other and are represented as density maps in Figures 3.6-3.8. Figure 3.5 shows how to interpret the maps.

The intensity of the colour indicates the homology between two sequences. The sequences within any of the groups are sorted according to sub-group classification as defined by (Kabat *et al.*, 1992), so that sequence families appear consecutively. The same plots are generated for the whole F_V framework sequences and for a set of human heavy chain germ-line sequences for comparison.

The identity plots shown in Figures 3.6-3.8 clearly demonstrate that the chain

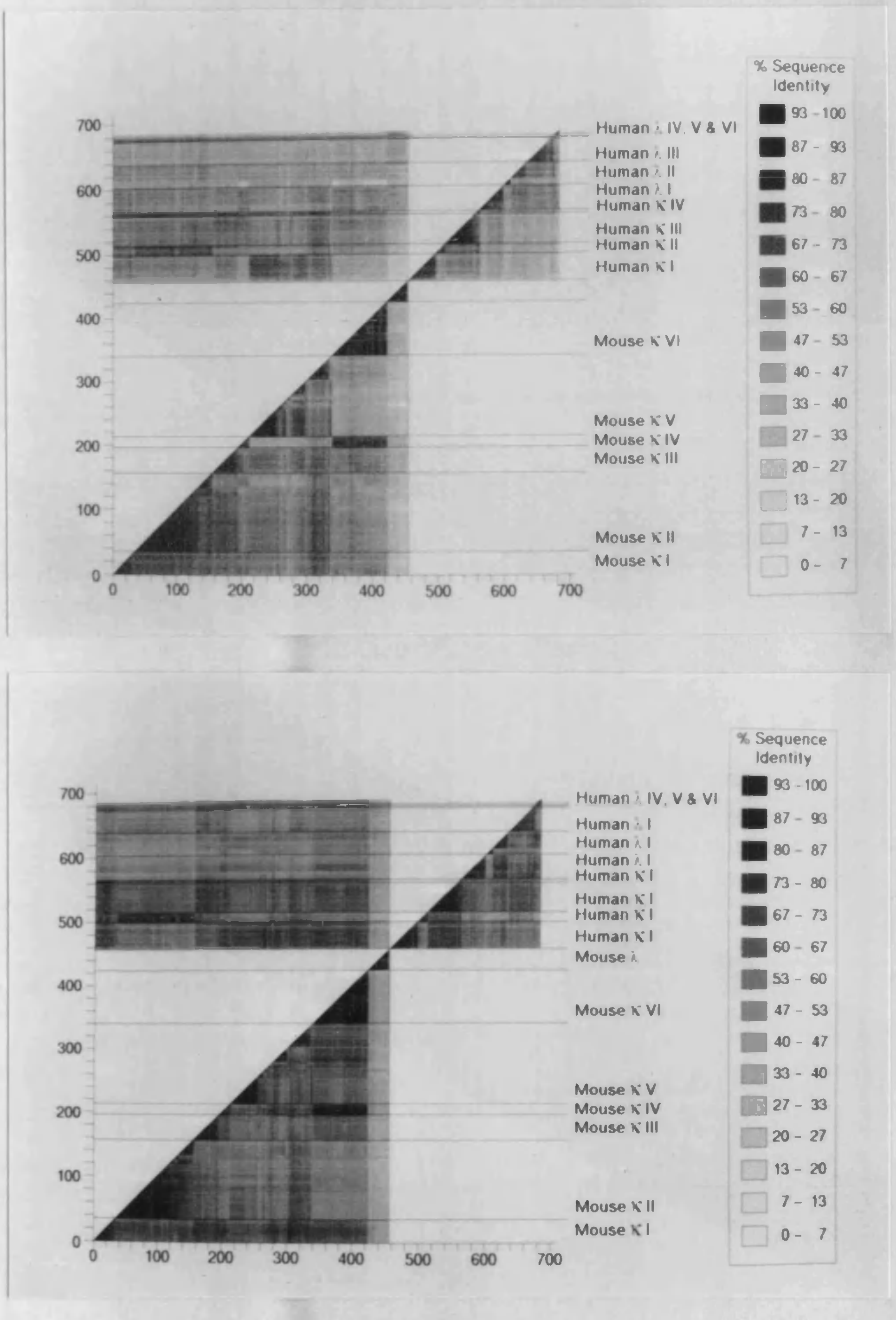


Figure 3.6: Homology plots of all the surface motifs extracted from sequence database, compared to whole sequence. The two plots shown are *a*) Light chain sequences surface residues only, and *b*) whole framework. For plots *a, b, c* and *d* the sub-group classification of (Kabat *et al.*, 1992) is used, for plots *e* and *f* the classification of (Tomlinson *et al.*, 1992) is used. Where whole framework sequence is shown, the CDR's and surface residues have been excluded from the identity calculation.

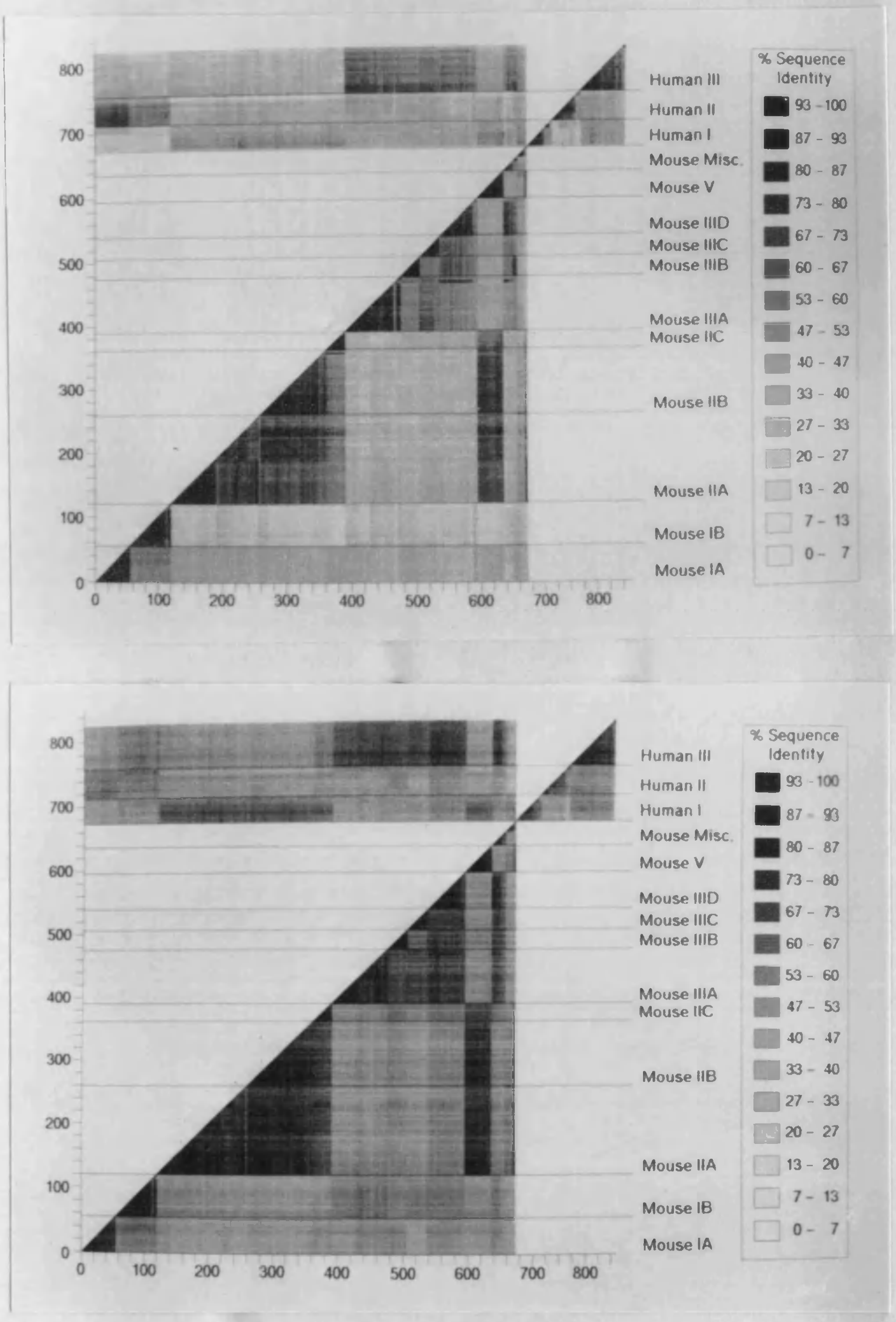


Figure 3.7: Homology plots of all the surface motifs extracted from sequence database, compared to whole sequence. The two plots shown are *c*) Heavy chain sequences surface residues only, and *d*) whole framework. For plots a, b, c and d the sub-group classification of (Kabat *et al.*, 1992) is used, for plots e and f the classification of (Tomlinson *et al.*, 1992) is used. Where whole framework sequence is shown, the CDR's and surface residues have been excluded from the identity calculation.

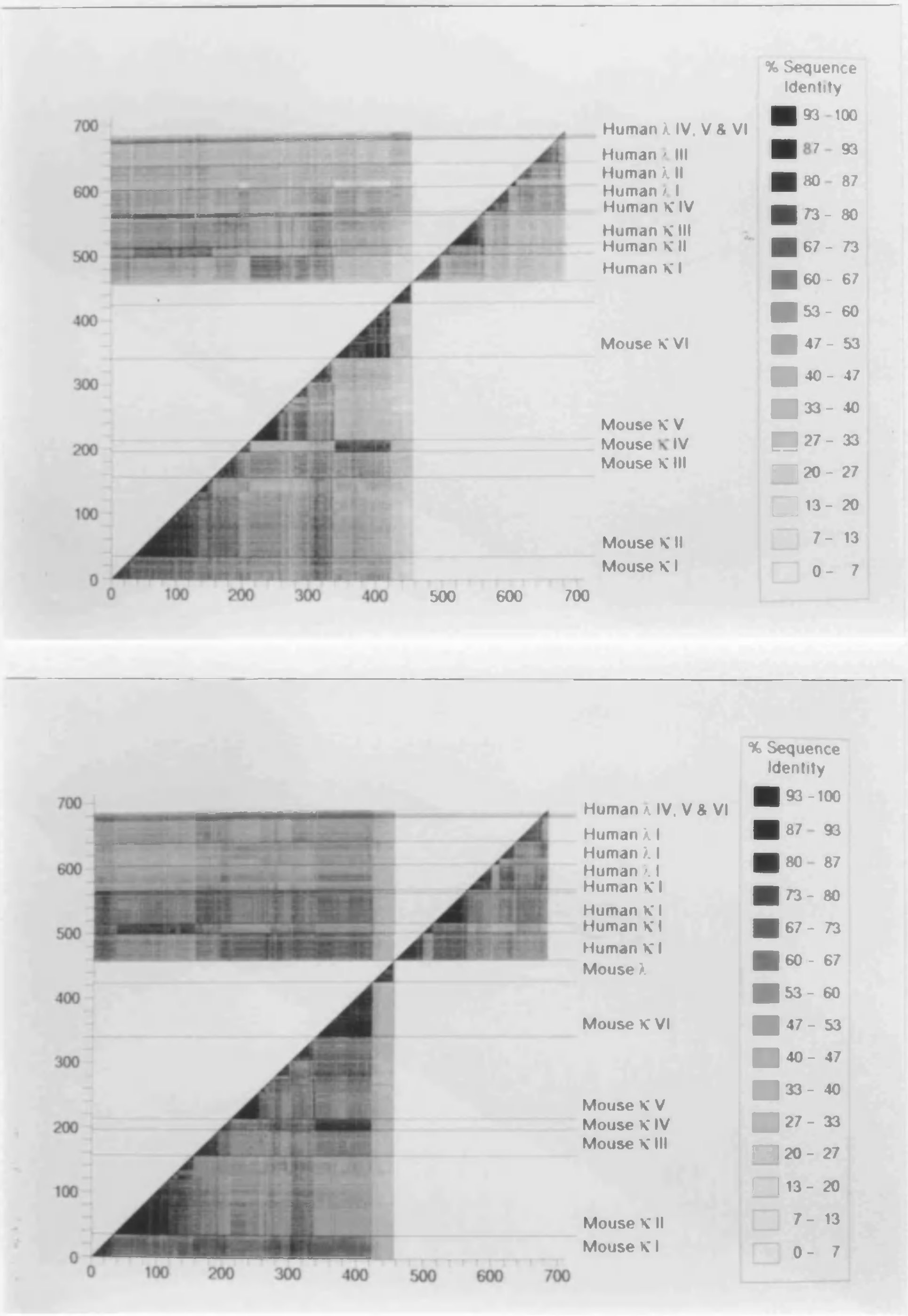


Figure 3.6: Homology plots of all the surface motifs extracted from sequence database, compared to whole sequence. The two plots shown are *a*) Light chain sequences surface residues only, and *b*) whole framework. For plots *a*, *b*, *c* and *d* the sub-group classification of (Kabat *et al.*, 1992) is used, for plots *e* and *f* the classification of (Tomlinson *et al.*, 1992) is used. Where whole framework sequence is shown, the CDR's and surface residues have been excluded from the identity calculation.

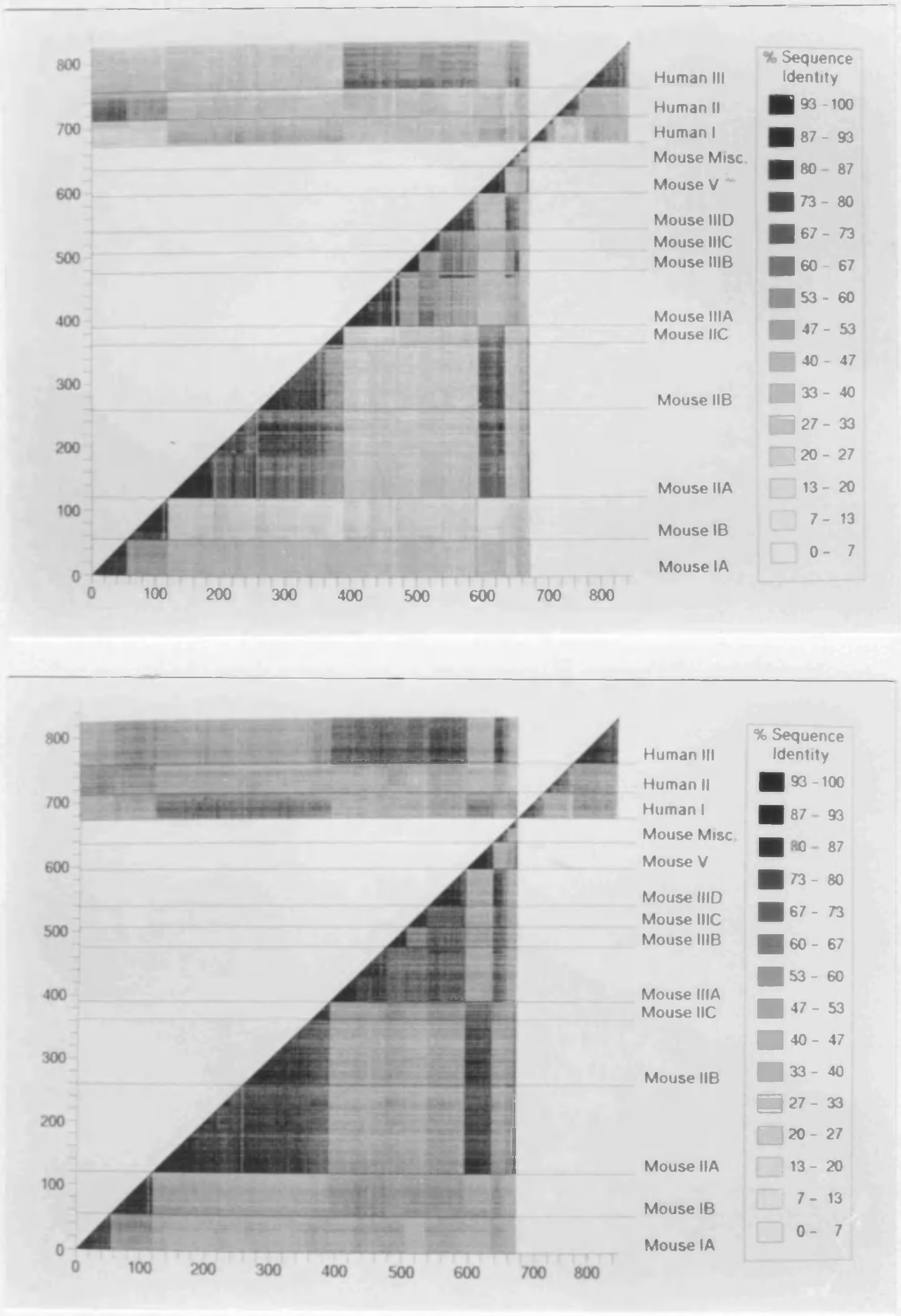


Figure 3.7: Homology plots of all the surface motifs extracted from sequence database, compared to whole sequence. The two plots shown are *c*) Heavy chain sequences surface residues only, and *d*) whole framework. For plots *a*, *b*, *c* and *d* the sub-group classification of (Kabat *et al.*, 1992) is used, for plots *e* and *f* the classification of (Tomlinson *et al.*, 1992) is used. Where whole framework sequence is shown, the CDR's and surface residues have been excluded from the identity calculation.

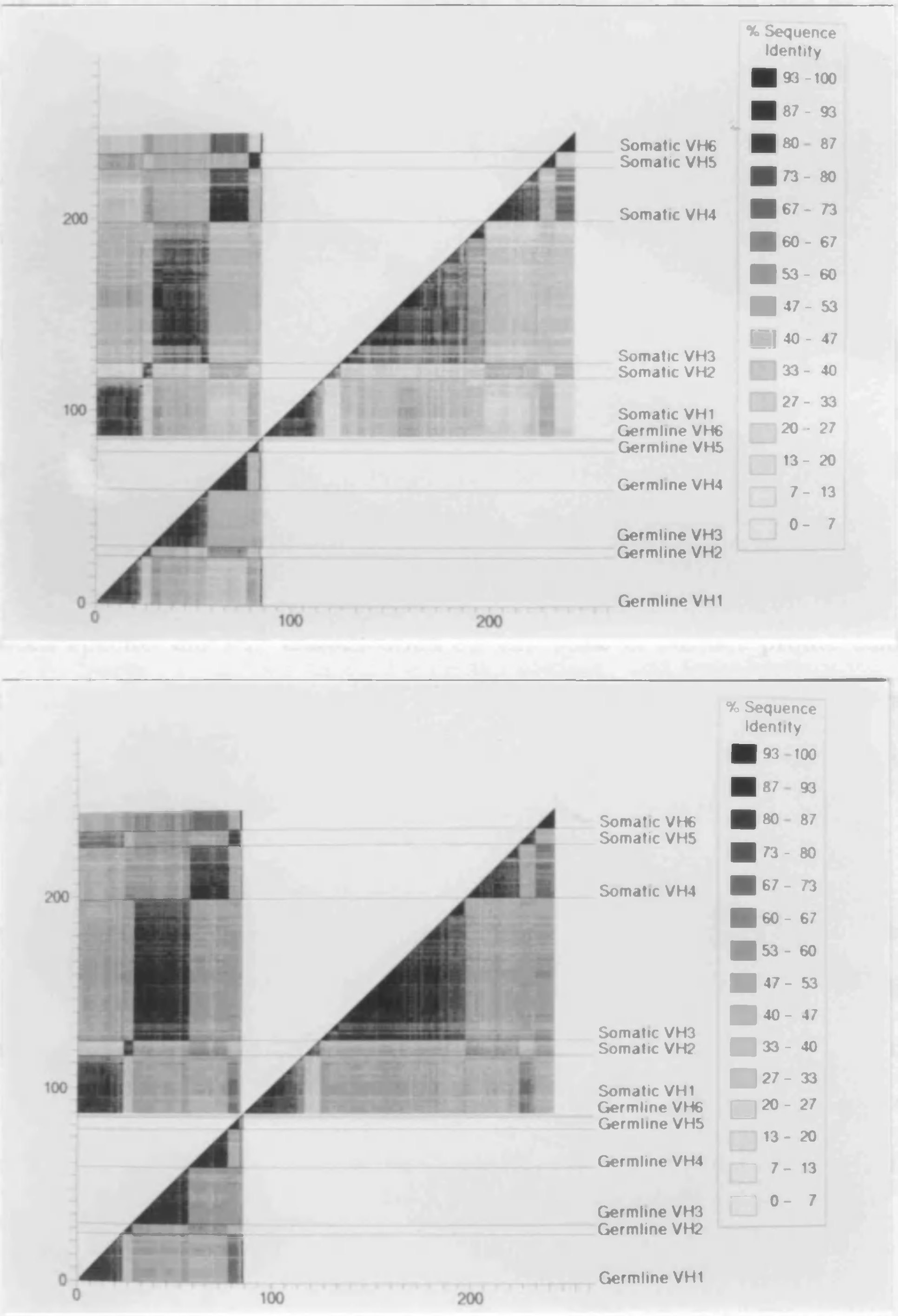


Figure 3.8: Homology plots of all the surface motifs extracted from sequence database, compared to whole sequence. The two plots shown are *e*) Heavy chain sequences surface residues only for germline sequences, and *f*) whole framework for germline sequences. For plots *a*, *b*, *c* and *d* the sub-group classification of (Kabat *et al.*, 1992) is used, for plots *e* and *f* the classification of (Tomlinson *et al.*, 1992) is used. Where whole framework sequence is shown, the CDR's and surface residues have been excluded from the identity calculation.

classification traditionally used to segregate families can be replaced by considering the surface residues alone.

Figure 3.6 *a* and *b* reveal that sub-group $\kappa 4$ and $\kappa 6$ murine sequences are very similar, and may belong to the same family.

In (Figure 3.7 *c* and *d*) it can be seen that that heavy chain sub-groups $VH2$ and $VH5$ are so similar as to warrant classification in the same group. Again, the human $VH4$ and $VH6$ families have considerable identity and may justifiably be clustered together.

Between species the VH classification on the basis of surface profile confirms that the mouse $VH1$ and human $VH2$ families are closely related as are the mouse $VH2$ and human $VH1$ families. For the $VH3$ family the classification is consistent between the two species.

In order to determine whether the surface patterns are conserved in the germline the same homology plots were produced for a set of human germline (Tomlinson *et al.*, 1992) and somatic mutant sequences (Figure 3.8 *e* and *f*). The somatic mutants do not show any significant difference in the chain classification when compared to the germline, suggesting that the surface residue positions are conserved during maturation of the immune response.

Table 3.3 shows the variability of all the residue positions on the surface compared to the variability of framework residues. There is no significant difference in the conservation of surface residues in the framework, compared to core framework residues, although the surface residues in the V_H sequences appear to be slightly more variable than any of the other surface residues.

	Framework		Surface	
	L chains	H chains	L chains	H chains
Number of sequences	841	627	539	836
Number of residues/sequence	68	70	20	26
Variability	18.70±13.70	14.68±8.82	17.56±8.90	24.36±16.44

Table 3.3: Average variability of residues in surface positions compared to framework. Framework is here defined as all residues in an F_V sequence which are not on the surface and not in any of the CDR's. The variability is calculated according to Kabat (Kabat *et al.*, 1992)

3.3.2 Resurfacing of variable domains

The analysis in the above section suggests an interesting approach to the reshaping of F_V fragments. Solely by changing the surface, and thus leaving the CDR interacting residues of the antibody framework untouched, it should be possible to make a hybrid F_V fragment which has the core and CDR's of a murine antibody and the surface of a human antibody (Figure 3.9).

The number of mutations required to make a human sequence from a particular sub-group sequence, using the suggested resurfacing protocol, are listed in table 3.4. For mouse $\kappa 2$ light chains an average of only four mutations are required to make a human surface, whereas 7 to 8 mutations is required to transform a λ chain surface into a human surface, showing that the selection of the initial murine sequence is a critical determinant of how many residues have to be changed in order to obtain a human F_V surface.

To test the resurfacing hypothesis three humanisation experiments were set up in collaboration with ImmunoGen Inc (Cambridge, Mass. USA): 1) Traditional CDR grafting (Verhoeyen *et al.*, 1991) onto a human F_V framework of known structure; 2) Surface humanisation using the most similar human heavy and

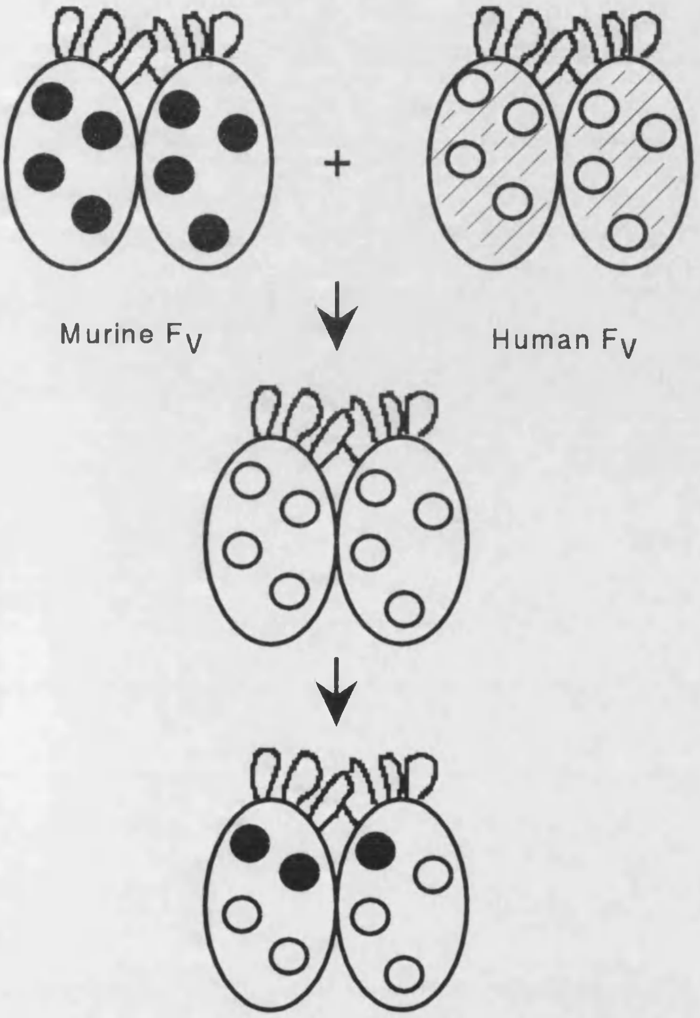


Figure 3.9: Diagram showing the resurfacing approach to humanisation described in the text. It can be divided into two stages. In the first, the mouse framework (white) is retained and only the surface residues changed from mouse (dark grey circles) to the closest human pattern (light grey circles). This should remove the antigenicity of the mouse antibody. In the second stage surface residues within 5 Åof the CDRs are replaced with the mouse equivalents in an attempt to retain antigen binding, and CDR conformation.

Chain type	Number of mutations	Standard Deviation	Number of sequences
$M1A$	5.5	1.12	55
$M1B$	8.0	1.41	65
$M2A$	10.5	2.29	139
$M2B$	10.0	2.00	103
$M2C$	9.0	1.41	28
$M3A$	7.0	1.41	85
$M3B$	7.0	1.41	31
$M3C$	7.0	0.82	33
$M3D$	8.5	2.29	56
$M5A$	10.0	1.41	40
$\kappa 1$	5.5	1.12	27
$\kappa 2$	4.0	1.41	112
$\kappa 3$	5.0	1.41	36
$\kappa 4$	7.0	1.41	18
$\kappa 5$	5.0	2.00	115
$\kappa 6$	7.0	1.41	60
λ	7.0	1.41	21

Table 3.4: Table showing the number of mutations required to change a V_L or V_H chain surface to that of the closest human counter part. Sub-groups of heavy chains are as defined by (Kabat *et al.*, 1992)

light chains; 3) Surface humanisation using human heavy and light chains with most similar surface residues.

The antibody used for the test was a murine anti-NCAM (CD-56) antibody (anti-N901 here called N901 (Griffin *et al.*, 1983)) of class IgG1, κ. N901 binds to natural killer cells (NK) and is highly specific to large granular lymphocytes (LGL). The antibody is being developed as an antibody-toxin conjugated for the treatment of various malignant tumors.

The alignment in Figure 3.10 shows the original N-901 antibody and the sequences used in each of the three approaches outlined here.

Humanisation by CDR grafting In traditional humanisation the CDR's of the rodent antibody are grafted onto a framework of known structure and CDR-framework interactions are monitored by a homology modelling procedure. Models of N901 and the initial humanised construct were made. The model of the humanised antibody was compared to that of the original rodent antibody, and possible CDR interacting framework residues were changed back to the murine sequence (marked with '*' in alignment) in order to retain the three-dimensional shape of the CDR's. For N901 antibody KOL was used, this resulted in a low identity score of only 59 % and 46 % in the heavy and light chains respectively. These low identities are likely responsible for the poor success rate of antibodies humanised in this way (Kettleborough *et al.*, 1991).

Most similar chain humanisation A total of 164 human heavy and 129 human light chains were sampled from the sequence database. Each of the rodent chains, L and H, were then matched and the most identical human heavy and light

```
Light Chain Sequences

                                   10        20         30        40         50         60         70
                           ---------+---------+--- ------+---------+ ---------+----- ----+-- -------+
  1 N901L                 :DVLMTQTPLSLPVSLGDQASISC RSSQIIIHSDGNTY-LE WFLQKPGQSPKLLIY KVSNRFS GVPDRFSG

  2 KOL                   :QSVLTQPPSASG-TPGQRVTISC SGTSSNIGS----STVN WYQQLPGMAPKLLIY RDAMRPS GVPDRFSG
                           | *     |    |  |
  3 N901L/KOL             :QVLMTQTPSSLPVTLGQQASISC RSSQIIIHSDGNTY-LE WFLQKPGQSPKLLIY KVSNRFS GVPDRFSG

  4 KV2F$HUMAN            :DVVMTQSPLSLPVTLGQPASISC RSSQSLVYSDGNTY-LN WFQQRPGQSPRRLIY KVSNRDS GVPDRFSG
                             *   |      |  ||                                   *
  5 N901L/KV2F            :DVLMTQSPLSLPVTLGQPASISC RSSQIIIHSDGNTY-LE WFQQRPGQSPRLLIY KVSNRFS GVPDRFSG

  6 KV4B$HUMAN            :DIVMTQSPDSLAVSLGERATINC KSSQSVLYSSNNKNYLA WYQQKPGQPPKLLIY WASTRES GVPDRFSG
                             *     |        |
  7 N901L/KV4B            :DVLMTQTPDSLPVSLGDRASISC RSSQIIIHSDGNTY-LE WFLQKPGQSPKLLIY KVSNRFS GVPDRFSG
                                                  [        L1       ]               [  L2   ]

                                   80        90        100        110
                           ---------+---------+---- -----+----- ----+----
  1 N901L                 :SGSGTDFTLMISRVEAEDLGVYYC FQGSH--VPHT FGGGTKLEI

  2 KOL                   :SKSGASASLAIGGLQSEDETDYYC AAWDVSLNAYV FGTGTKVTV    ( 44)
                                |   |        |                    *
  3 N901L/KOL             :SGSGTSFTLAISRVEAEDEGVYYC FQGSH--VPHT FGGGTKLEI    (104)

  4 KV2F$HUMAN            :SGSGTDFTLKISRVEAEDVGVYYC MQGTH--WSWT FGQGTKVEI    ( 87)
                                   |        |                     *
  5 N901L/KV2F            :SGSGTDFTLKISRVEAEDVGVYYC FQGSH--VPHT FGGGTKVEI    (101)

  6 KV4B$HUMAN            :SGSGTDFTLTISSLQAEDVAVYYC QQYDT---IPT FGGGTKVEI    ( 71)

  7 N901L/KV4B            :SGSGTDFTLMISRVEAEDLGVYYC FQGSH--VPHT FGGGTKLEI    (109)
                                                   [    L3     ]

Heavy Chain Sequences

                                   10        20        30          40        50         60         70
                           ---------+---------+---------+ ------- --+---------+- --------+--- ------+
  1 N901H                 :DVQLVESGGGLVQPGGSRKLSCAASGFTFS SFGMH-- WVRQAPEKGLEWVA YISSGSF--TIY HADTVKG

  2 KOL                   :EVQLVQSGGGVVQPGRSLRLSCSSSGFIFS SYAMY-- WVRQAPGKGLEWVA IIWDDGS--DQH YADSVKG
                           |         |    | ||        |                 |            ||          |
  3 N901H/KOL             :EVQLVESGGGVVQPGRSLRLSCAASGFIFS SFGMH-- WVRQAPGKGLEWVA YISSDGF--TIY HADSVKG

  4 G36005                :QVQLVESGGGVVQPGRSLRLSCAASGFTFS SYAMH-- WVRQAPGKGLEWVA VISYDGS--NKY YADSVKG
                           |         |    | ||                          |                     |  |
  5 N901H/G36005          :QVQLVESGGGVVQPGRSLRLSCAASGFTFS SFGMH-- WVRQAPGKGLEWVA YISSGSF--TIY YADSVKG

  6 PL0123                :EVQLVESGGGLVQPGGSLRLSCAASGFTFS SYWMS-- WVRQAPGKGLEWVA NIKQDGS--EKY YVDSVKG
                           |                ||                          |                        |
  7 N901H/PL0123          :EVQLVESGGGLVQPGGSLRLSCAASGFTFS SFGMH-- WVRQAPGKGLEWVA YISSGSF--TIY HADSVKG
                                                         [  H1   ]              [    H2      ]

                                   80        90       100        110        120       130
                           ---------+---------+---------+-- -------+--------- +---------+
  1 N901H                 :RFTISRDNPKNTLFLQMTSLRSEDTAMYYCAR MRKGYAM--------DY WGQGTTVTVSS

  2 KOL                   :RFTISRNDSKNTLFLQMDSLRPEDTGVYFCAR DGGHGFCSSASCFGPDY WGQGTPVTVSS   ( 78)
                                  |*            *
  3 N901H/KOL             :RFTISRDDPKNTLFLQMTSLRSEDTAMYYCAR MRKGYAM--------DY WGQGTTVTVSS   (107)

  4 G36005                :RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR DRKDWGWALF-----DY WGQGTLVTVS-   ( 89)
                                   |    |   |   |    |
  5 N901H/G36005          :RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR MRKGYAM--------DY WGQGTLVTVSS   (104)

  6 PL0123                :RFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR ----------------- -----------   ( 74)
                                   |            |
  7 N901H/PL0123          :RFTISRDNAKNTLFLQMTSLRAEDTAMYYCAR MRKGYAM--------DY WGQGTTVTVSS   (111)
                                                           [        H3       ]
```

Figure 3.10: Alignments of sequences generated using the three methods of humanisation outlined in the text. Sequences are: 1) Original rodent N901. 2+3) KOL (Marquart *et al.*, 1980) and reshaped N901 using KOL surface. 4+5) Most homologous sequences, L (KV2F) (Klobeck *et al.*, 1985b) and H (G36005) (Schroeder and Wang, 1990), and reshaped N901 using these sequences. 6+7) Most homologous with respect to surface residues, L (KV4B) (Klobeck *et al.*, 1985a) and H (PLO123) (Bird *et al.*, 1988a), and reshaped N901 using these sequences. The numbering is the same as used in the antibody modelling program ABM (OML, 1992), which is based on structural conservation and not sequence homology as used by Padlan *et al* (Kabat *et al.*, 1992). The sequence changes which have to be introduced in order to reshape N901 with a given sequence are marked with bars, back-mutations as determined from F_V models are marked with stars. The sequence homology of a given sequences to N901 are shown in brackets after each sequence. Names of database sequences are cited using the OWL (Bleasby and Wouton, 1990) database entry names. The common names for the four sequences used are: KV2F/RPMI6410, KV4B/JI, G36005/M74, PLO123/TD-Vr

chain independently. For N901 these were G36005 (Schroeder and Wang, 1990) and KV2F (Klobeck *et al.*, 1985b) respectively. The identities for the selected sequences are 76 % for the light chain and 68 % for the heavy chain. Surface residues, as indicated in Tables 3.1 & 3.2, were then changed in the murine sequences to match those of the human sequences. Subsequently a model was built of the resurfaced antibody and compared to the model of the original murine antibody. The framework-CDR interface was then inspected and any framework residue within 5 Å of a CDR residue and whose conformation was affected by the changed surface was back mutation to the mouse sequence. In Figure 3.10 sequences 4 and 5 indicate that residues 3 and 52 of the light chain influenced adjacent CDR residues and required restoration of the murine sequence. In the heavy chain no back mutations were necessary. The resurfaced sequences showed a final identity to the selected human sequences of 80 % and 89 % for the heavy and the light chains respectively. These identities include CDR residues.

Most similar surface humanisation In this approach the human heavy and light chains are selected on the basis of their closest identity to the N901 sequences for the surface residues only. The selected sequences were PLO123 (Bird *et al.*, 1988a) and KV4B (Klobeck *et al.*, 1985a) for heavy and light chains respectively. These sequence have 57 % and 62 % identity (not including CDR residues) with the murine heavy and light chain sequences respectively. After construction of the resurfaced model and comparison with the native murine model the only back mutation found to be necessary was at position 3 of the light chain (as in the similar chain method above. The identity of the final sequences (Figure 3.10 sequence 7) are 85 % and 96 % for heavy and light chains respectively (including CDR residues).

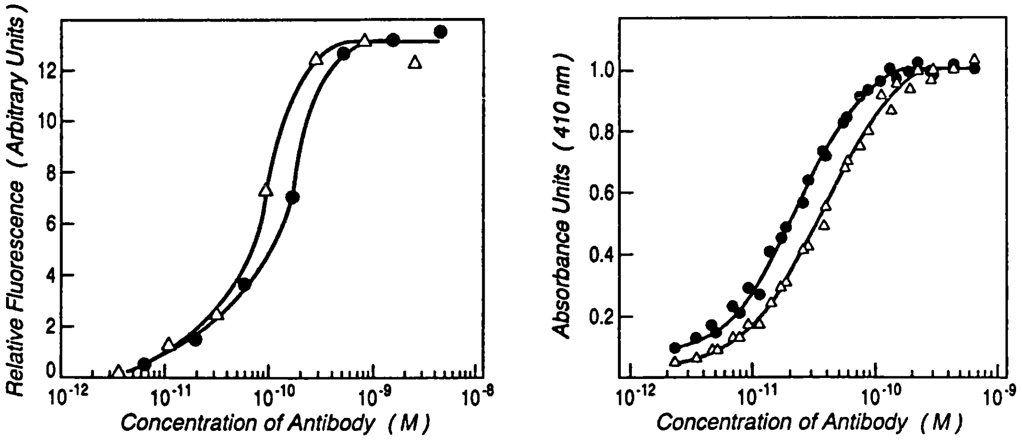


Figure 3.11: A) Binding to SW-2 cells as measured by indirect immunofluorescence. B) Binding to an SW-2 cell membrane preparation in an ELISA assay. These data are produced in collaboration with M. Roguska of ImmunoGen Inc., Cambridge, Mass, USA. Δ: binding curves for resurfaced antibody; •: binding curves for original murine antibody

3.4 Experimental testing of humanised N901

The two latter gene sequences have been synthesised and antibodies expressed. The two humanised antibodies have both been shown to retain binding to the original antigen. In a competitive binding assay, the resurfaced N901 was equal to murine N901 in its ability to inhibit the binding of fluorescein-labeled murine N901 to antigen positive SW-2 cells (FACS assay (Parks *et al.*, 1979)). The apparent K_D values for the resurfaced and grafted antibodies are $9.0x10^{-11}$ M and $1.0x10^{-10}$. The K_D for the murine antibody in the same assay is $1.6x10^{-10}$ M (see Figure 3.11). The result of the binding studies makes it possible to conclude that the framework-CDR interactions in the resurfaced and grafted N901 preserve the native conformation of the CDRs.

3.5 Summary & conclusions

Resurfacing murine F_V's is likely to minimise CDR-framework incompatibilities because a large number of murine surface residues are retained. The total number of differences between the surface residue patterns of the murine N901 V-region and the most identical human V-region was remarkably low so that only a small number of amino acid changes needed to be made to humanise the antibody. This strong conservation of surface accessible amino acid residues and their localisation in the F_V's of murine and human antibodies, together with the fact that the sidechains of surface accessible residues are in general not critical to the structural integrity of the F_V's, may hint at a biological significance for the selective conservation of surface patterns in antibodies.

Chapter 4

Towards antibody design

4.1 Drug - pocket interactions (ligand design)

In drug design a pharmacophore is sought which fits into a binding site (cleft, hole or cavity) of a protein. In traditional drug design an iterative process is used, where series of drug molecule analogs are synthesised and tested. The structural and physiological data obtained from these experiments is then correlated using Qantitative Structure Activity Relationship (QSAR) analysis. In this type of analysis the data is correlated using the Hansch equation (Hansch, 1969) or Free-Wilson method (Free and Wilson, 1964). The correlation can then be used to build a model of the receptor.

Molecular modelling methods provide an alternative method for the generation of complementary shapes. These methods combine various types of shape description with database searches. Kuntz has developed an ingenious sphere generation algorithm (implemented by the author in the program INT see Appendix B.2). The algorithm is able to identify cavities on the surface of proteins and fill these with spheres which have the approximate size of atoms. The "cast" is then used to search a database of small molecule structures which have been determined experimentally. This method has been used to identify possible inhibitor *lead* molecules which bind to the Human Immunodefeciency Virus protease (Desjarlais *et al.*, 1988), although in this instance the docking orientation was not correctly predicted.

Another method, based purely on distance geometry uses a distance matrix to describe the shape of the receptor binding site. This method of identifying complementary shapes has been applied to dihydrofolate reductase (Crippen, 1981), and enzyme binding compounds were identified from a structural database.

Both of the above methods are based on geometry as the primary criterion for complementarity. However, geometric constraints alone frequently fail to identify the molecules from a structural database which are known to bind to a given protein. Potential energy functions, free energy functions and hydrophobic potentials (Desjarlais, 1988) have been included in order to get sufficient discrimination between true and false positives when screening a structural database in the design process.

The above drug design principles are the basis for the receptor design process described in this Chapter. However, the process here attempts to address the reverse question: Which receptor (antibody) will fit a specified ligand (antigen) molecule.

4.2 An approach to *de novo* antibody design

The aim of *de novo* antibody design is to model a complete antibody combining site knowing only the three dimensional structure of the antigen, and obtain an antibody which will bind to the antigen. The process is almost the opposite to that of the antibody modelling presented in Chapter 2 where the aim is to be able to suggest a three dimensional model corresponding to a given sequence. In the *de novo* design the goal is to suggest a sequence of amino acids which is able to bind to a pre-defined antigen, using three dimensional information.

Three different approaches to the design problem are outlined:

1. **Surface matching** Define shape parameters from known three dimensional structure of antigen and search database of molecular surfaces in order to find complementary shapes. Use the complementary shapes (fragments) to build up the antibody CDR loops. This can be tested in the laboratory as cross-reactivity.
2. **Minimal change** In this method a known antibody-antigen complex structure is taken as the starting point for the design. A homologous antigen structure is then docked in the same orientation and the changes which have to be made in the antibody in order to retain binding are determined.
3. **Peptide** *ab initio* This method is related to (2) above, but uses small peptide/hapten antigens. In this process no prior knowledge of paratope shape and orientation of antigen are assumed. Using peptides allows for the easy testing of a large range of homologous compounds, and thus mapping of the individual interactions.

The third method was chosen as the most viable method for testing a given design algorithm. The choice of a small antigen limits the number of interactions and a much larger database of structures is available. Also, using peptides keeps open the possibility of using proteins at a later stage. Furthermore, the synthesis of many different analogs to test a designed antibody binding site is made easier.

4.3 The design process

4.3.1 Antibody selection

The shape of a protein is determined by its backbone fold, and the functionality and physiological properties are mainly determined by the distribution of amino acid sidechains (Padlan, 1990). If this assumption is applied to complementarity of antibodies and antigens, it is deducible that any antigenic site (of a particular size) can be complemented by many different complementary shapes (if not all !). For the design process it is assumed that specificity of the antibody will not be dependent on particular combinations of CDR length. The **basis or platform** of the antibody could be a tight groove accommodating the binding of a small hapten or a small loop on a large multisubunit protein (see FIgure 1.6). Here the **platform** is defined as residues in the CDR's which are not structural, e.g. canonical residues.

One argument for the validity of the assumption that backbone conformation and CDR length is largely independent of the size of the antigen is that the immune system would be vulnerable if there is only one, or very few, complementary shapes to any unknown antigen shape. In order to explore the CDR length dependency on antigen size the CDR length of a number of sequences for which the antigen is known were plotted (see Figures 4.1-4.3). In these plots the molar weight is used as a descriptor of antigen shape, although this is a poor measure it is possible to see that any CDR length is allowed for small antigens. For larger antigens ($M_R > 10000$) it is not possible to deduce anything from these data as molar weight does not accurately describe the shape of the binding epitope. This is substantiated by crystal structure data that show the formation of similar grooves in binding sites by both long and short CDRs. In McPC603 (Rudikoff *et al.*, 1981), which is an anti-phosphocholine antibody, the binding site is a tight hole and the length of the CDR H3 is eleven residues. In Gloop-2 (Jeffrey *et al.*, 1991), which binds to a nine residue peptide fragment of hen egg lysozyme, a groove exists in the center of the binding site although the length of CDR H3 is only four residues. This means that the length independency hypothesis can be applied at least to anti-hapten antibodies. Some antibodies have been reported recently (Rossmann, 1993) which require a special type of CDR in order to be able to bind to a concave (a hole) shaped epitope. These are CDRs which bind by insertion into a hole on the surface of human rhinovirus via a long CDR H3. In these unusual instances the independence hypothesis will clearly break down.

If the backbone conformations of the CDR's of a typical antibody are to be conserved, there are approximately 25 residues out of 50 in an average antibody which can be changed without changing the shape of the CDR's. There are 19 (Pro excluded, and only 18 if Gly is also excluded) different residues which can be substituted at each of these positions. This gives 19^{25} possible combinations of sidechains, which is far more than the $10^7 - 10^9$ antibodies which are thought to be necessary to account for any molecular shape (Perelson, 1989).

The design process derived from these considerations is shown in Figure 4.4

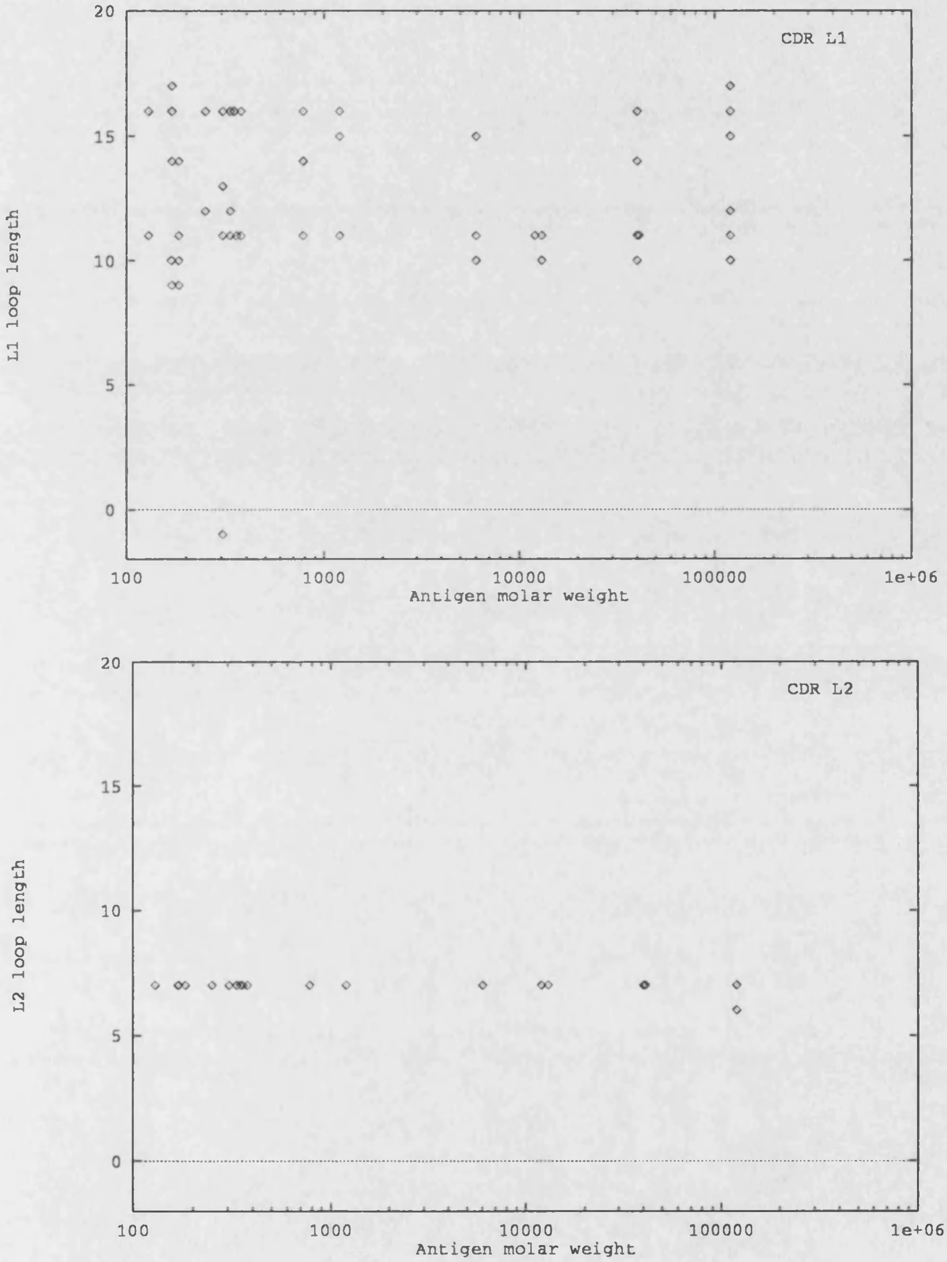
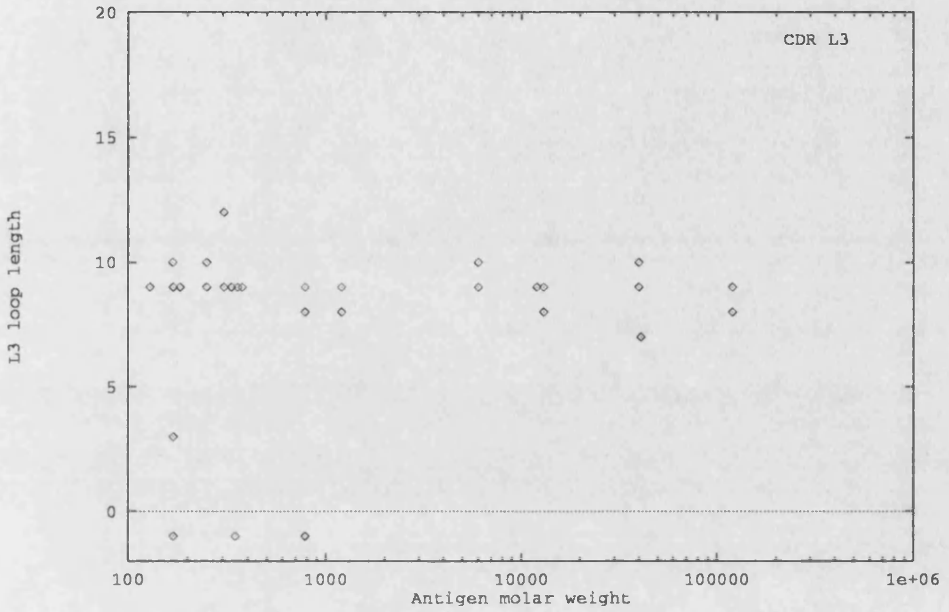


Figure 4.1: CDR length as a function of antigen size, The molar weight is used as the shape descriptor. Plots are for CDR L1 and CDR L2. Data was extracted from the sequence database of (Kabat *et al.*, 1992)



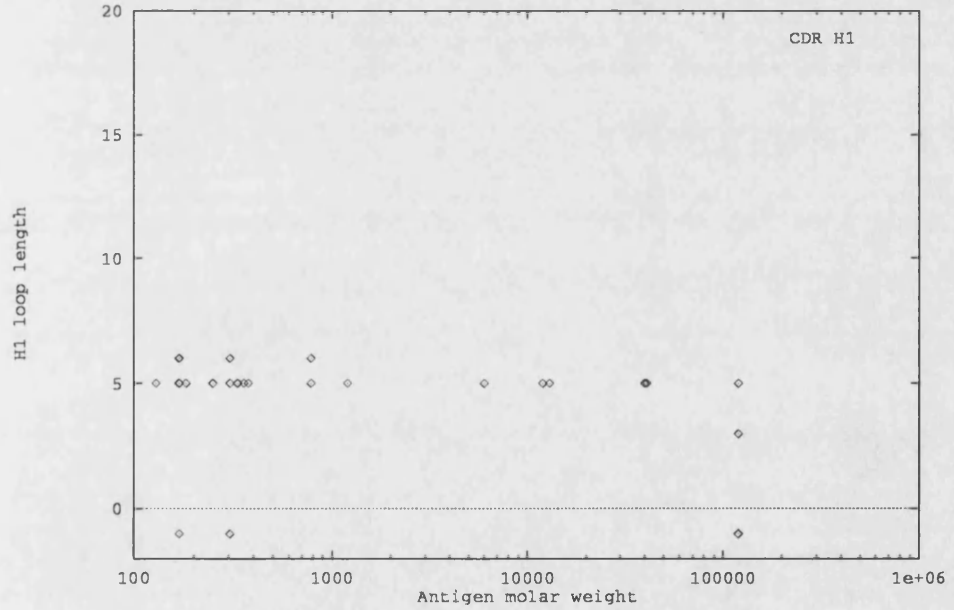


Figure 4.2:
CDR length as a function of antigen size, The molar weight is used as the shape descriptor. Plots are for CDR L3 and CDR H1. Data was extracted from the sequence database of (Kabat *et al.*, 1992)

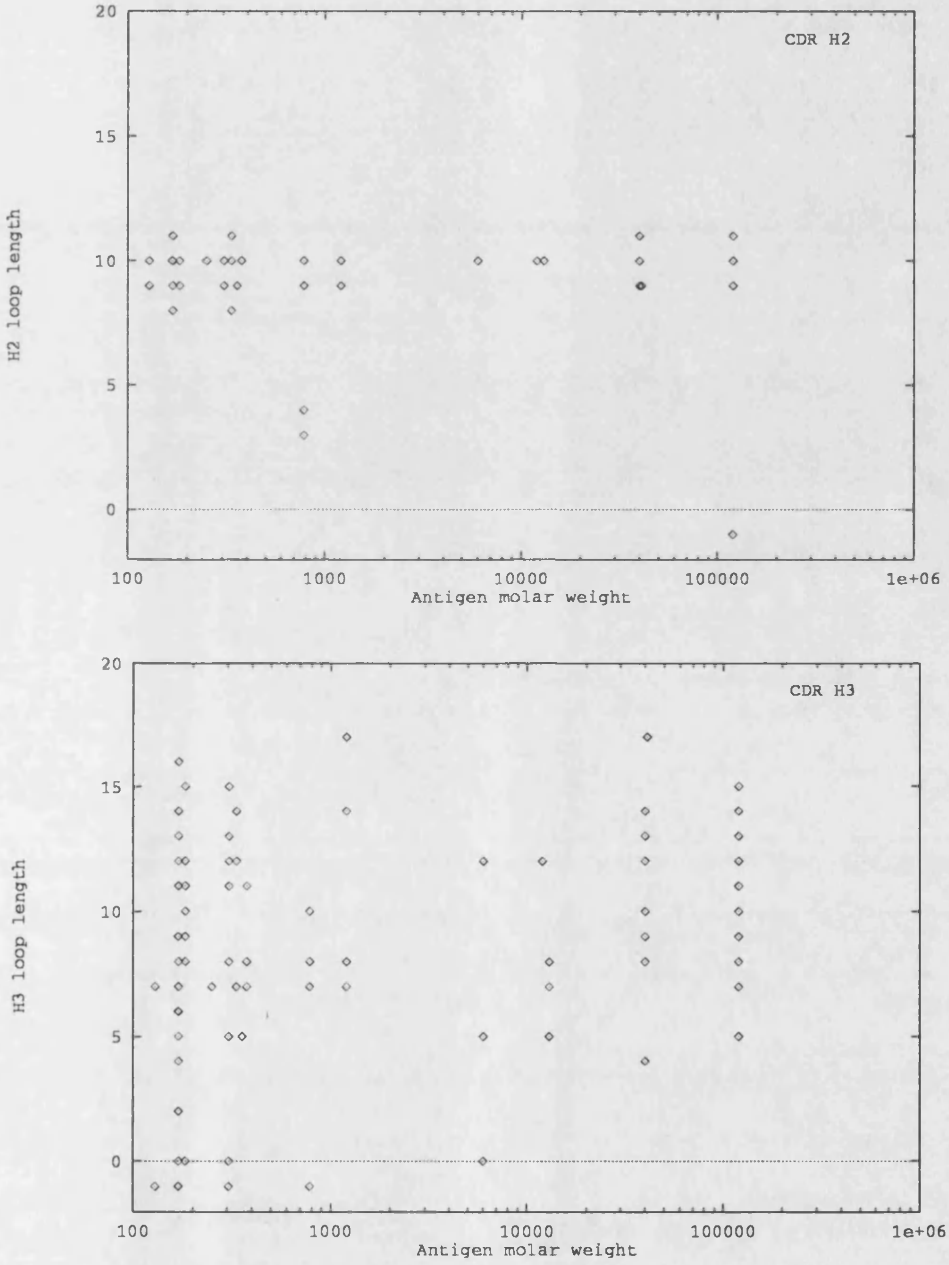


Figure 4.3:
CDR length as a function of antigen size, The molar weight is used as the shape descriptor. Plots are for CDR H2 and CDR H3. Data was extracted from the sequence database of (Kabat *et al.*, 1992)

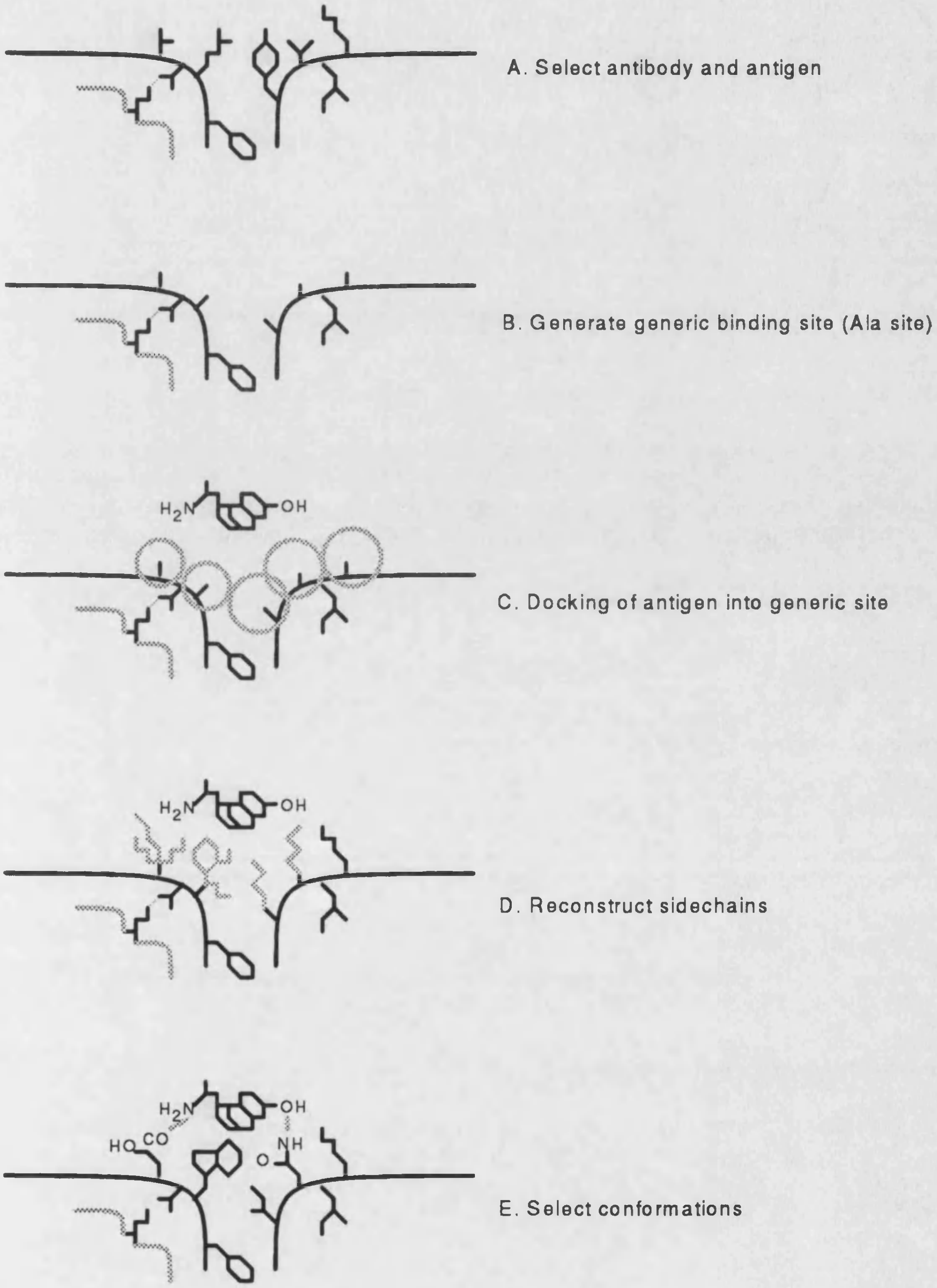

A. Select antibody and antigen
B. Generate generic binding site (Ala site)
H₂N
OH
C. Docking of antigen into generic site
H₂N
OH
D. Reconstruct sidechains
H₂N
OH
HO
CO
NH
O
E. Select conformations

Figure 4.4: In the design process (outlined in detail in the text) an antibody of known structure is chosen by random. A generic (alanine) binding site is generated by replacing all non-structural residues by alanine with an extended VdW radius (R = average length of all 20 amino acids). A tentative docking is performed, and sidechains reconstructed. Using various objective scoring functions the sidechain conformations are evaluated, and a final conformation is selected.

4.3.2 Generation of a generic binding site

From an arbitrarily chosen antibody, the sidechains that do not influence the backbone structure or are not buried by the CDR backbone and framework, are truncated to alanine residues. The resulting structure is termed "the alanine cushion". Alanines are choosen for the generic binding side to allow for pseudo properties to be assigned to the sidechain. The Cβ atoms of the alanines are assigned an extended VdW radius of 4.2 Å which allows for other sidechains to be constructed at each of the alanine positions. The extended radius of 4.2 Å is the average sidechain length for the 20 most abundant amino acids.

4.3.3 Antigen docking

The next step is to dock the antigen in a reasonable initial orientation in the combining site in order to obtain the maximum interaction surface area and the maximum satisfaction of electrostatic interactions. Several different strategies have been tested in this work:

1. **Functionality mapping** of all the known functional groups in one or more analogs of the antigen is attempted. Then distance constraints for optimal liganding geometry are determined. All possible sidechain combinations are then searched using this distance geometry information (Crippen, 1981).
2. **Functionality mapping, minimum perturbation** As (1) above but where only functional groups which are present in the binding site are searched. The search is biased by use of the statistical distribution of amino acids in particular positions, determined from aligned antibody sequences.

3. **Monte Carlo docking** where docking is attempted using the **Autodock** program on an empty (Ala) combining site (Goodsell and Olson, 1990).
4. **Residue-residue interaction preferences** determined with the program SIRIUS (Singh and Thornton, 1990). This procedure is similar to (2).

4.3.4 Sidechain construction

After the antigen has been docked in the binding site all residue positions which can potentially interact with the antigen are reconstructed. A distance cutoff (equal to the length of an extended Arg residue, which is the longest possible residue) is used to determine which residues are to be constructed, and which of the original residues are to be retained. The crude distance criteria is used first in order to reduce the time needed to compute the sidechain conformations. This gives the initial residue location to be constructed. When all sidechain conformations for these positions have been constructed a further reduction of residue positions is obtained by selecting only those residue positions for which a sidechain conformation exists which is capable of interacting with the antigen. If a position does not generate any conformations which can contact the antigen, the original sidechain is retained.

For the residue positions where sidechains pass the above criterion all possible conformations of the 19 amino acids (Pro excluded) are then constructed, where a simple energy function (Equation 4.1) eliminates unfavorable conformations:

$$E_{sidechain} = \sum \varepsilon [(\frac{r^{\star}}{r})^{12} - 2(\frac{r^{\star}}{r})^{6}] + \sum \frac{q_i q_j}{r} + \sum \kappa_o \cdot cos(3\omega) \qquad (4.1)$$

ε is a constant describing the steepness of the Lennard-Jones potential; r is the distance between two atoms; $r^\star$ is the minimum energy distance between two atoms; q_i, q_j is the partial charge of atoms i and j; κ_0 is a constant describing the size of the torsional potential; ω is the angle of a given sidechain torsion.

The sidechains are constructed recursively in a torsional grid (10°-30°), using tree pruning to avoid combinatorial explosions. This sidechain reconstruction algorithm has been implemented by the author in the program MC (see documentation in Appendix B.1). All sidechain conformations for each position are then ranked using the above energy function on the sidechain conformers and the antigen alone. By this procedure the sidechains that interact best with the antigen, that is have the lowest electrostatic interactions, are scored highest. Standard forcefield parameters as contained in the DISCOVER (TM Biosym Inc. San Diego, CA) molecular mechanics program are used. As a second measure of how well the antigen was buried in the surface of the antibody combining site, an accessibility calculation was carried out.

4.3.5 Selection of conformations

Simple free energy equations have been used by (Novotny, 1991) to estimate the binding energy of antibody-antigen complexes. Novotny's free energy of binding, ΔG_{tot} is a function of five terms:

$$\Delta G_{tot} = \Delta G_{\phi} + \Delta G_{EL} - T\Delta S_{CF} - T\Delta S_{TR} - T\Delta S_{CR} \tag{4.2}$$

ΔG_{ϕ} is the free energy contribution from the hydrophobic effect and is proportional to the excluded surface area upon antigen binding; ΔG_{EL} accounts for electrostatic interactions; $T\Delta S_{CF}$ describes the loss of sidechain conformational entropy upon antigen binding; $T\Delta S_{TR}$ is a term which describes the loss of overall rotational and translational entropy; $T\Delta S_{CR}$ is a correction term which accounts for dilute concentrations of proteins encountered in biological systems.

Unfortunately Equation 4.1 does not contain any terms which describe the exclusion of hydrophobic surface area from the antigen species. In order to be able to select and rank sidechain types and conformations the following ranking scheme was used. First all possible sidechains satisfying Equation 4.1 to within a given energetic cutoff were constructed. For all of these conformations the solvent accessible surface area lost by the antigen was calculated. The lists of energies and accessibilities are then sorted independently. From the accessibility list the possible residue types for a given position in the sequence are determined and lowest energy conformation of each sidechain type is extracted. The residue types and order obtained from the two lists is then combined to give a final residue rank for each sequence position.

4.4 The design of an opioid antibody (GlaMor)

In order to test the design process outlined above, Gloop-2 was chosen as the antibody scaffold and the enkephalins/ morphins as the antigen.

Class	Type of secondary structure	Examples
Caffein	Non-peptide	Caffein, Theophyllin
Small neuro peptides	Turn	Enkephalins
Cyclical peptides	Turn	Cyclosporins
Linear peptide hormones	Turn or none	MSH, FSH, ACTH
AIB peptides	Helix	Alimethicin

Table 4.1: Possible antigen target groups selected on the basis of abundance of structural information and of prior knowledge of antigenicity.

4.4.1 Antigen selection

The selection of the antigen target was limited by the available structures in the Cambridge Crystallographic database (Kennard, 1991). Possible candidate molecular classes sampled from the database, are outlined in Table 4.1. It is important that there exists both crystallographic and NMR structures for the selected groups, to ensure that the conformation of the antigen is not flexible in solution. This avoids the complications introduced by induced fit. From this list three groups of antigens were considered.

The first selected group are the non-peptide antigens caffein and theophyllin (Sutor, 1958b; Sutor, 1958a). The second selected group contains the enkephalin neuropeptides (opioids) (Aubry *et al.*, 1988), which consist of 5 to 7 residues, and the non-peptide opioid morphine (Bye, 1976). The third selected group are the helical peptides containing the sterically constrained amino acid α-amino-isobutyric-acid. Structures of many peptides of this type are known (Karle *et al.*, 1990; Karle *et al.*, 1991).

The rationale behind these selections is to provide a spectrum of antigen types, namely, a small non-peptide hapten, a small peptide hapten and a large peptide antigen. The structures and the structural information for each of these are outlined in Figure 4.5.

To begin the process, the opioids, where most structural information is available were chosen as the primary target for the first design project.

The reason for choosing the enkephalins is that there exists a wide range of structural information on a large number of opiates and many QSAR (Quantitative Structure Activity Relationship) analyses are available. Antibodies which have a cross reactivity between different opiates have also been produced (Kussie *et al.*, 1991).

The opiate receptor binding site has been extensively mapped by many different methods (see (Casy and Robert, 1986) for a review). The main features of the binding site are outlined in Figure 4.6. The opiate structures of Leu-Enkephaline, Morphine, Naloxone, Methadone and Nalorphine were overlapped in order to be able to establish distance constraints for strategies 1 and 2 outlined in Section 4.2.

All the opiate structures have been determined by x-ray crystallography of crystals from aqueous solution, and refined to a resolution of < 0.5 Å. All the structures obtained from the crystallographic database were minimised by the author, using the DISCOVER (TM Biosym Technologies) forcefield, without solvent molecules present.

The inter-atomic distances between the seven points defined in Figure 4.6 were calculated for each of the structures. The average distances were then used to search the binding site of the target antibody. During the search, an upper devia-

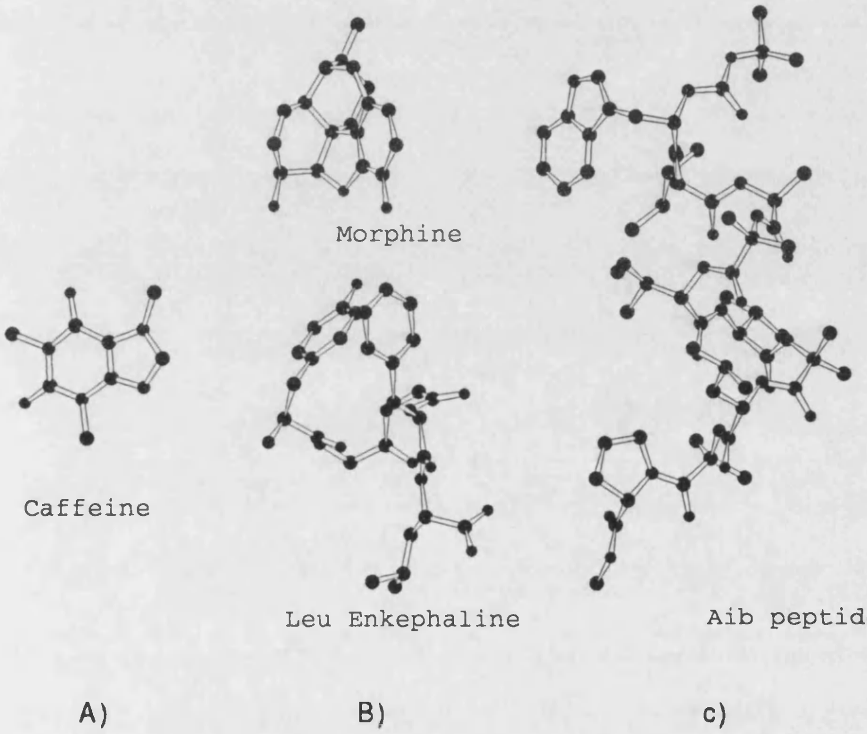


Figure 4.5: MC (Appendix B.1) plots of the three groups of compounds chosen for the design of an antibody.A) Caffein (Sutor, 1958b) B) Morphine (Bye, 1976) and β-turn conformation of Leu-enkephalin (Aubry *et al.*, 1988). C) BOC-Trp-Ile-Ala-Aib-Ile-Val-Leu-Aib-Pro-OMe helical peptide (Karle *et al.*, 1991).

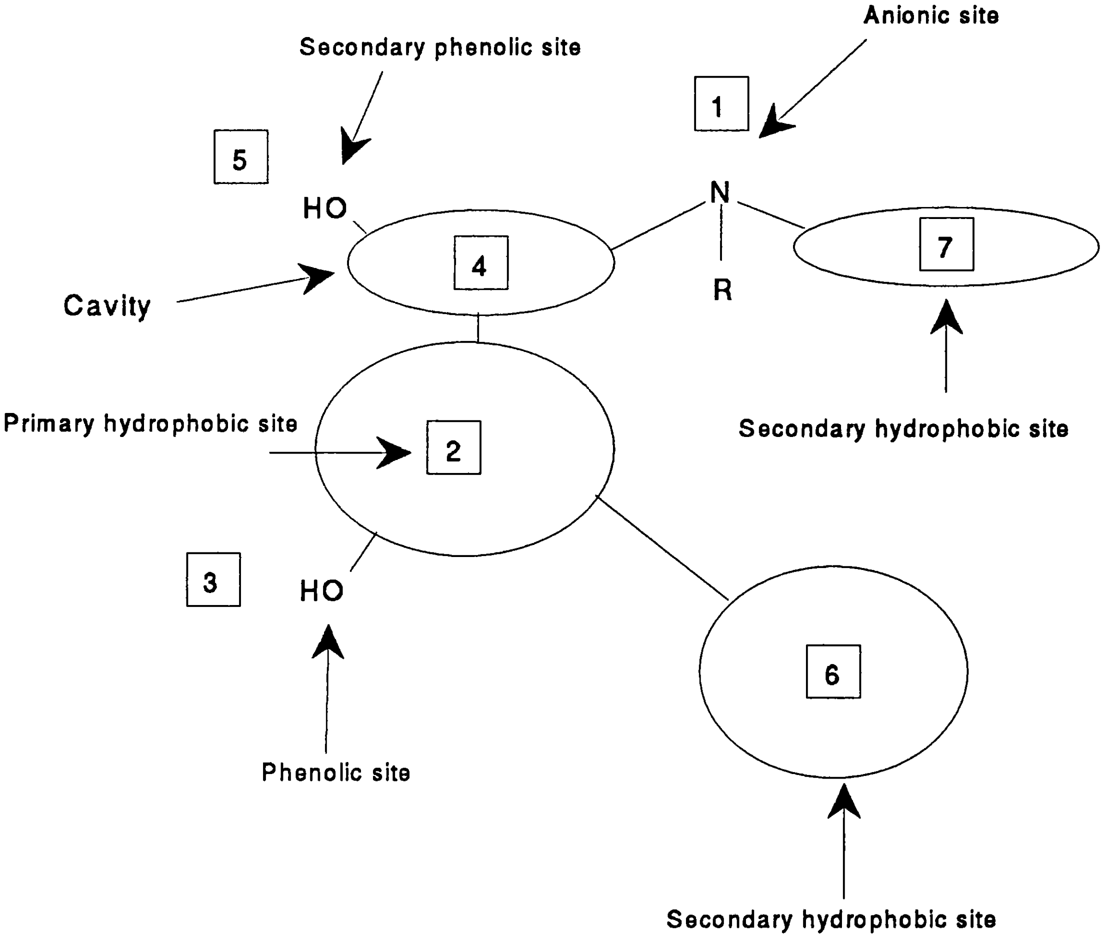


Figure 4.6: Schematic representation of the opioid receptor binding site (Casy and Robert, 1986). The data are obtained from QSAR analysis of the binding properties of a large number of antagonists and agonists. This mapping is the basis for the calculation of distance constraints. Numbered boxes indicates the position of liganding groups in the receptor.

Loop	Canonical Class	Sequence
L1	2	{ R [A] S Q E [I] S G Y [L] S }
L2	1	[I] Y { A A S T L D S } [G]
L3	1	{ L [Q] Y L S Y [P] L T }
H1	1	{ T F G [I] T }
H2	none	{ [E] [I] [F] [P] [G] [N] [S] [K] [T] [Y] }
H3	none	{ [E] [I] [R] [Y] }

Table 4.2: **Canonical classification of Gloop-2. Canonical residues are in [ ] brackets. The loop region is marked with { } brackets. There are 26 amino acid positions which can be changed in order to accommodate the complementarity, and still retain the original backbone conformation.**

tion corresponding to the diameter of a non-hydrogen atom (2-3 Å) was specified, for each of the search distances. This upper deviation is allowed since the constraints obtained from a complementary shape must be larger than constraints obtained from the ligands.

4.4.2 Antibody selection

The antibody platform chosen was the anti-lysozyme antibody Gloop-2 (Darsley and Rees, 1985). The CDR sequences are outlined in Table 4.2.

In the case of Gloop-2 there are three glycines in the CDRs (L1,H1 and H2), but none of these are in positions where a possible sidechain would be able to interact with the antigen. Although the glycines are not classified as canonical residues they should probably be excluded from the set of modifiable sidechains because of their possible structural role.

4.4.3 Search methods

The antibody combining site was searched in three different ways in order to find an initial orientation of the ligand.

First, all the 26 (see Table 4.2) residues are removed and replaced by alanine. A distance map of the Cβ positions is generated. This distance matrix is then searched against the distance constraints determined for the antigen. The constraints are sorted after length, and the longest are searched first. If a hit is found within the allowed variation of the distance constraint, the next constraint is searched for, and so on. If all the constraints are satisfied a hit is found. The hits are ranked according to deviation from mean constraints.

In the second search all the residue types in the 26 positions which can paticipate in specific ligand interactions are retained. For example, Asp and Glu sidechains can participate in charge-charge interactions which are often important for ligand-receptor interactions. These residue positions are then searched with the distance searching procedure described above. Identifying an initial orientation by this procedure reduces the number of residues that have to be changed. This method is attractive because it determines the maximum number of ligand requirements that can be satisfied by the original antibody sidechains (minimum perturbation).

The third method for orientation achieves direct docking using simulated annealing (Goodsell and Olson, 1990). The principle of the docking is simple, and consists of "throwing" the ligand, inside a pre-calculated potential grid (field), into the combining site a large number of times. An average orientation is derived from the lowest energy conformations. This method has been used successfully for the docking of phosphocholine into the combining site of the McPC-603 antibody

(Goodsell and Olson, 1990). The potential grid used in Goodsell's AUTODOCK program is calculated using parameters from the AMBER (Weiner *et al.*, 1984) forcefield.

The last method of orientation, mentioned in Section 4.3.3, was not applied to this problem. The procedure is a knowledge based method which uses the fact that there appears to be a preference for particular residue-residue pairing when proteins interact with each other (Singh and Thornton, 1990). This preference has been exploited previously to design antagonists and agonists (Singh *et al.*, 1991). In the present case the orientation would have to rely solely upon the Leu-enkephaline structure since this is the only peptide structure in the set of analogs. Since there are several different crystal structures of Leu-enkephaline available which all have different conformations (Aubry *et al.*, 1988), this method of initial orientation was considered unsuitable for the design of this antibody.

Distance geometry searching resulted in many (>10000) possible conformations. It was not possible to evaluate all these conformations properly within a reasonable time frame. Sorting the geometric hits using a simple RMS deviation between the search constraints and the geometric hit resulted in 3 main clusters of orientations, two large on the outside of the combining site, and one smaller in the center pocket of the binding site. The center pocket was uniquely identified by the AUTODOCK program. This orientation in the center pocket was therefore selected as the initial orientation (Figure 4.7). This orientation is similar to the orientation of flourescein in the 4-4-20 antibody (Herron *et al.*, 1989).

After the initial orientation has been obtained all the original residues of the set of 26 which are not in contact ($d > 8.0$ Å) with the antigen are restored. All residues (13 positions) which are in contact with or overlapping the antigen are reconstructed, using the MC sidechain replacement program described previously. The sidechain conformations are selected using the scoring function in Equation 4.3:

$$S_i = A_i \cdot w_{burried} + E_i \cdot w_{energy} \tag{4.3}$$

Where A_i is the surface area of the antigen burried by sidechain conformation i; E_i is the energy of the conformation according to Equation 4.1; $w_{burried}$ and w_{energy} represent a relative weighting of the two terms. This results in the ranked list of residues to be constructed as listed in Table 4.3.

From Table 4.3 the lowest energy conformations of the two most probable sidechain types, at each sequence position, are extracted (Table 4.4). From Table 4.4 the ten lowest energy residue combinations for all the residues in the construct are selected (Table 4.5).

4.5 The final GlaMor antibody model

The final ten best models were subjected to energy minimisation (100 cycles of steepest descents followed by 300 cycles of conjugate gradients) in order to validate the conformations. Little change in sidechain conformation was observed (Figure 4.8). Table 4.6 summarises the results of the minimisation and the ten best models.

The RMS deviation of the sidechains of the ten best constructs follow the same trend. The largest RMS deviation (0.4 Å) was observed for residue 203 (His).

Residue	Atom-Atom	Nconf	Dist.	Exclusion order	Energy order	Rank
203	CB - HO12	2594	2.04	R,W,M,K,Y,F,Q,E,L,H,I,N,D,V,A,G	Q,E,M,L,R,K,V,W,N,D,F,Y,H,I	I,H,D,N,V,Y,F,L,K,W,E,Q,R,M
91	CB - H3	2230	3.48	R,W,M,K,F,Y,Q,E,H,L,I,N,D,V,A	Q,E,M,L,R,V,K,N,D,I,H,W,F,Y	I,D,N,H,Y,V,F,L,W,K,E,Q,R,M
204	CB - O1	3094	4.11	R,W,Y,M,K,F,E,Q,H,L,I,N,D,V,A	Q,E,M,L,R,K,V,I,H,Y,N,F,D,W	D,N,V,I,H,F,W,Y,L,K,E,Q,M,R
96	CB - H3	3128	5.05	R,M,K,W,Y,F,H,Q,E,L,N,I,D	M,Q,E,L,R,V,K,I,N,D,H,W,F,Y	D,N,I,Y,F,H,W,L,E,Q,K,R,M
94	CB - H6	2766	5.96	R,W,K,Y,M,F,H,Q,E,I,L	Q,E,M,V,R,L,K,I,N,D,H,W,F,Y	F,Y,H,I,L,W,E,Q,K,M,R
139	CB - H3	2128	6.41	R,W,Y,M,K,F,H,Q,E,I,L	M,E,Q,V,R,K,I,L,N,D,H,F,W,Y	L,F,H,I,Y,W,K,Q,E,R,M
32	CB - H5	3331	6.83	R,W,Y,M,K,F,H,Q,E	Q,E,M,R,K,L,V,I,H,Y,N,D,F,W	F,W,H,Y,E,K,Q,M,R
156	CB - H8	2694	6.97	R,W,M,Y,K,F,H.Q.E	Q,E,M,R,L,V,K,I,H,W,N,D,F,Y	F,Y,H,E,W,K,Q,M,R
154	CB - H8	2536	7.32	R,W,M,K,Y,F,Q,E,H	V,Q,E,M,L,R,K,I,D,N,H,F,Y,W	H,Y,F,W,K,E,Q,R,M
205	CB - O1	2988	7.13	R,W,M,K,Y,F,Q,E	Q,M,E,R,V,L,K,I,D,N,H,Y,F,W	F,Y,W,E,K,Q,R,M
161	CB - H17	3124	7.99	R,W,M,Y,K	Q,E,M,R,V,K,L,I,N,H,D,Y,W,F	Y,W,K,M,R
89	CB - H3	2966	8.01	R,W,M,K,Y	V,Q,E,M,L,R,K,I,D,N,H,W,F,Y	Y,W,K,M,R
136	CB - H11	2040	7.45	R,M,W,K	Q,E,M,R,V,L,K,I,N,D,W,H,F,Y	W,K,R,M

Table 4.3: Residue positions which can possibly interact with the antigen in its initial orientation after sidechain reconstruction. The residue numbers used are the same as used in A*b*M (OML, 1992). The sidechain positions are ranked after number of possible interacting residue types. This order roughly corresponds to distance ordering.

Residue	Type	Energy	Type	Energy
32	F	2663.54	W	3813.32
89	Y	1443.21	W	1386.82
91	I	2776.02	D	1756.31
94	F	1099.97	Y	1209.06
96	D	940.79	N	940.31
136	W	1949.30	K	1864.97
139	L	2607.09	F	1219.11
154	H	1279.43	Y	1356.96
156	F	946.01	Y	1106.48
161	Y	1592.40	W	1354.70
203	I	3523.71	H	2718.69
204	D	1188.85	N	1208.85
205	F	1248.81	Y	1249.98

Table 4.4: The two lowest energy residue types for each of the 13 reconstructed residue positions in the GlaMor F_V The total energy of the sidechain is Kcal. The enegies are relative within the model.

Residue position Number	32	89	91	94	96	136	139	154	156	161	203	204	205
1	F	W	D	F	N	K	F	H	F	W	H	D	F
2	F	W	D	F	**D**	K	F	H	F	W	H	D	F
3	F	W	D	F	N	K	F	H	F	W	H	D	**Y**
4	F	W	D	F	**D**	K	F	H	F	W	H	D	**Y**
5	F	W	D	F	N	K	F	H	F	W	H	**N**	F
6	F	W	D	F	**D**	K	F	H	F	W	H	**N**	F
7	F	W	D	F	N	K	F	H	F	W	H	**N**	**Y**
8	F	W	D	F	**D**	K	F	H	F	W	H	**N**	**Y**
9	F	**Y**	D	F	N	K	F	H	F	W	H	D	F
10	F	**Y**	D	F	**D**	K	F	H	F	W	H	D	F
Original	Y	L	Y	Y	L	F	T	E	F	K	E	I	R

Table 4.5: The ten lowest energy conformations of the complete construct. Residues which differ from the best (lowest energy) conformation are outlined in bold type. The original residues are also shown.

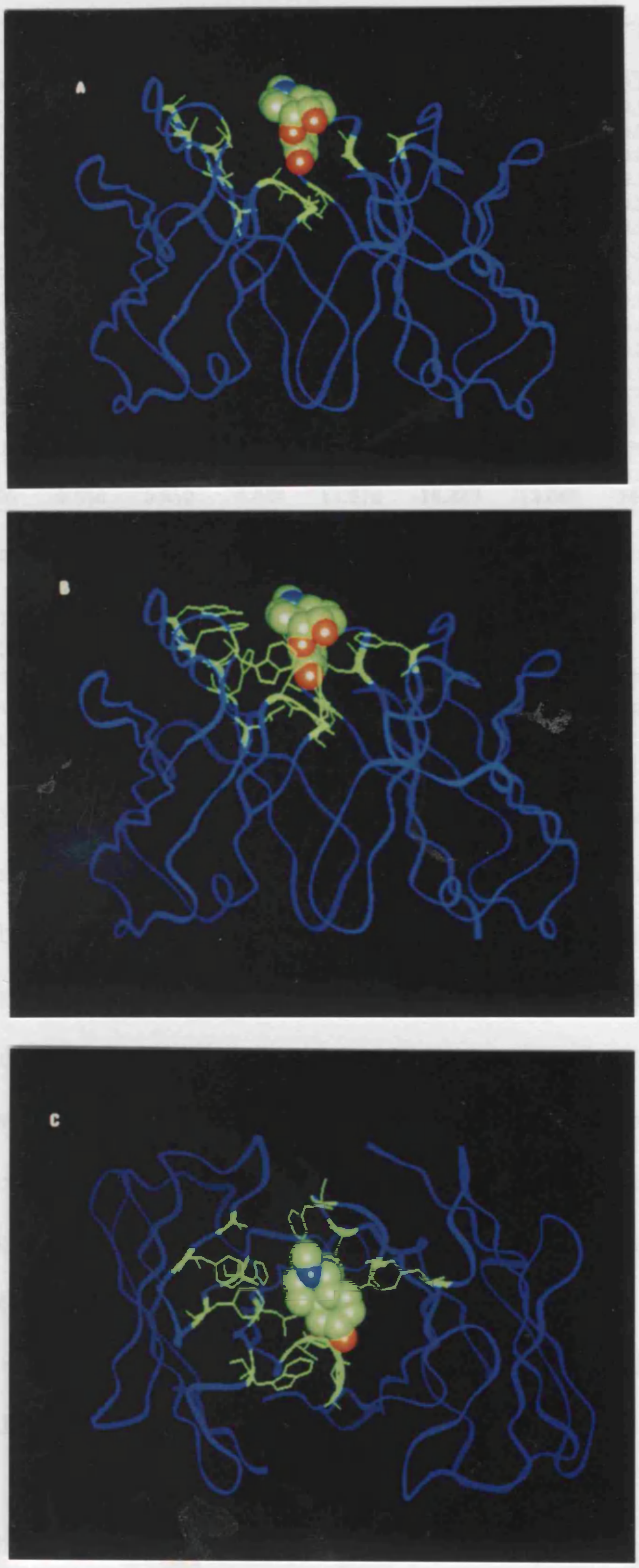


Figure 4.7: Pictures showing the initial (Ala) and final binding site of the GlaMor antibody (Best construct). A). Ala site. B). After sidechain reconstruction, side view. C). After sidechain reconstruction, front view.

Construct	1	2	3	4	5	6	7	8	9	10
Residue										
32	0.254	0.275	0.277	0.271	0.279	0.278	0.274	0.274	0.275	0.280
89	0.160	0.155	0.152	0.153	0.162	0.159	0.160	0.157	0.192	0.181
91	0.287	0.305	0.293	0.305	0.309	0.324	0.309	0.325	0.220	0.301
94	0.357	0.315	0.356	0.312	0.361	0.312	0.358	0.309	0.442	0.423
96	0.215	0.152	0.193	0.152	0.195	0.136	0.194	0.136	0.363	0.265
136	0.274	0.287	0.295	0.294	0.294	0.288	0.297	0.293	0.288	0.313
139	0.194	0.268	0.228	0.266	0.206	0.237	0.210	0.242	0.191	0.207
154	0.129	0.140	0.141	0.144	0.134	0.138	0.140	0.143	0.110	0.138
156	0.128	0.106	0.111	0.105	0.110	0.105	0.109	0.105	0.102	0.110
161	0.116	0.112	0.113	0.107	0.109	0.108	0.106	0.105	0.107	0.111
203	0.416	0.417	0.391	0.400	0.447	0.447	0.431	0.429	0.369	0.445
204	0.275	0.295	0.294	0.317	0.342	0.339	0.365	0.361	0.353	0.345
205	0.173	0.204	0.242	0.254	0.233	0.214	0.279	0.260	0.149	0.226
Average	0.229	0.265	0.237	0.237	0.245	0.237	0.249	0.241	0.243	0.257
Etot+ag	1804.6	1873.4	1850.7	1863.1	1848.5	1861.6	1837.6	1851.1	1807.5	1852.8
Etot-ag	1809.5	1878.1	1846.8	1867.2	1857.5	1873.4	1848.4	1863.7	1819.0	1858.3
Aver. deriv.	0.08	0.511	0.488	0.543	0.513	0.503	0.500	0.517	0.474	0.472
Max. deriv.	-13.400	-9.556	9.652	8.862	14.278	-16.449	-13.065	-16.662	-8.623	10.902

Table 4.6: Summary Table of results from minimisation of ten best constructs. RMS deviation refers to difference between minimised structure with and without antigen. In minimisations where antigen was present the antigen molecule was fixed. All structures were minimised with the same protocol (see text). The total energy of each of the constructs is approximately the same, as is the numerically largest derivative. *Aver.deriv.* is average largest derivative of any atom in the construct after minimisation; *Max.deriv.* is the largest derivative on any of the atoms after the minimisation.

Before the energy minimisation one hydrogen bond between the antibody and the antigen was observed (His 203 ND1 → Morphine O). After the minimisation two other hydrogen bonds were observed (Asp 204 OD1 → Morphine OH1, and His 203 NE1 → Morphine OH1). The largest derivatives indicate that the minimisation has not converged. The best conformation was therefore subjected to an additional 3000 steps of conjugate gradients minimisation which only changed the RMS deviation of residue Phe 32 by 0.2 Å. The largest derivative was also found on an atom of the Phe 32 sidechain. The largest derivative of the Phe sidechain after the 3000 steps of minimisation was 0.03 Kcal and was not decreased by a further 2000 steps of minimisation. The average derivative after the minimisation performed in Table 4.6 is less than 0.6 Kcal.

In Figure 4.9 all the opioid ligands extracted from the crystallographic database have been overlapped in the GlaMor combining site. All the ligands have the same orientation as the docked morphine. None of the non-peptide ligands clash with the backbone of the model, but Leu-enkephaline is overlapping sidechains with the terminal residues of the peptide.

4.6 Experimental test of the design

In order to determine how the design process performs two different experimental systems have been devised and implemented in this laboratory (Elliott, 1992). First, the ten F_V constructs described are now being synthesised as single chain F_V constructs. Second, restricted mutagenesis experiment has been set up, in which all the residue positions which are within range of the morphine molecule (the ten closest positions identified) are randomly changed and the resulting antibodies screened, using a phage display library (Clackson *et al.*, 1991). These two systems will allow for the revison of the design process by providing binding data for a set of mutants.

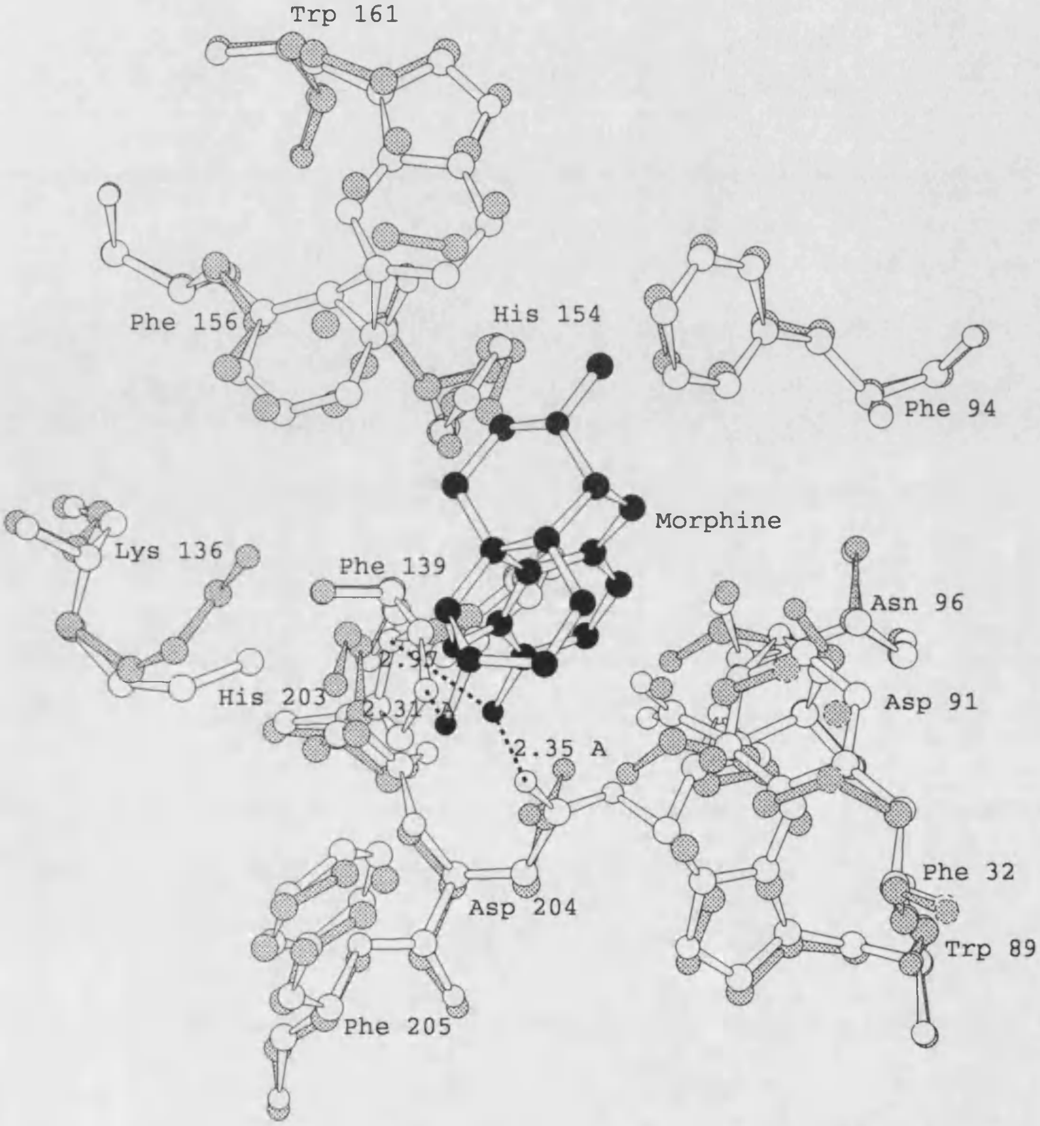


Figure 4.8: The GlaMor binding site before and after energy minimisation, the largest RMS deviation is observed on the sidechain of residue 203 (0.4 Å)

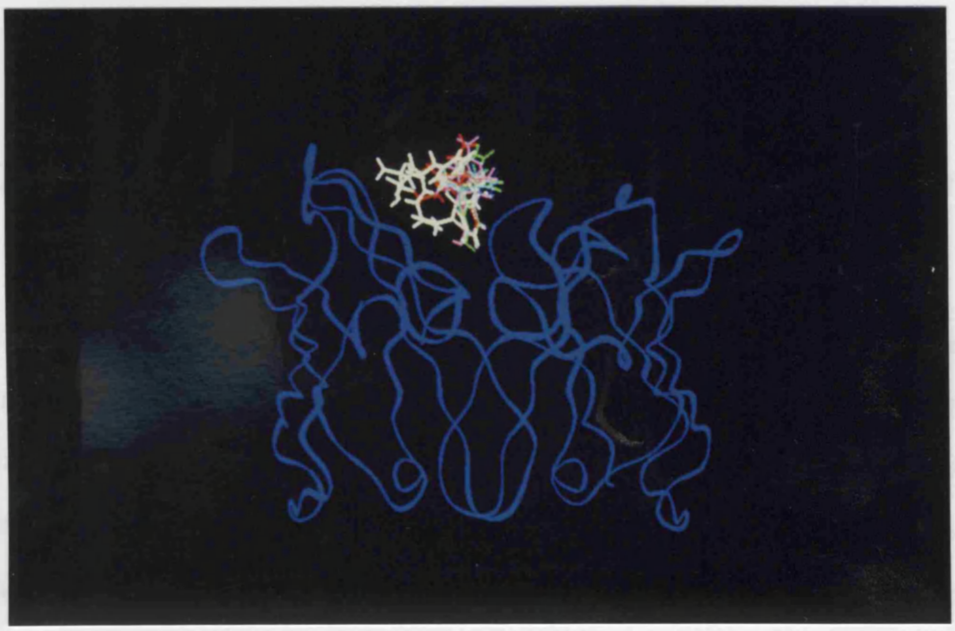


Figure 4.9: Overlap of opioid receptor ligands extracted from the crystallographic database. White: Leu-enkephaline, Blue: Morphine, Red: Methadone, Yellow: Nalorphine, Pink: Naloxone. None of the ligands overlap the F_V backbone, some clashes are observed between sidechains of Leu-enkephaline, and the constructed sidechains of GlaMor.

Chapter 5

Conclusions & Discussion

5.1 Antibody modelling - A retrospective view

The rising number of antibody F_V crystal structures and the use of new *ab initio* methods enables the prediction of F_V structures for which only the primary amino acid sequence is known. Using a combined algorithm, which generates F_V frameworks from a crystallographic database, and predicts CDR conformations by a number of methods, models can be generated which are within the accuracy of medium resolution x-ray crystallography.

From the data in Chapter 2 and data presented in Appendix A.3, several conclusions can be drawn:

1). In the original CAMAL (Martin *et al.*, 1989) algorithm which was developed on the basis of a single antibody structure (Gloop-2), all the CDR's were built using the combined algorithm. This sometimes results in higher RMS deviation values for CDRs where canonical families exist. In A*b*M, loops for which canonical families exist are built using the most homologous (canonical) loop from a database of antibody crystal structures. In Gloop-2 CDR H3 is a four residue loop. As shown by the CDR length distributions (Figure 2.3), this is an unusually short CDR length. As seen from the data in Chapter 2 and Appendix A.3 the accuracy of the prediction is related to the length of the loop. The shorter the loop the better the prediction.

2). For longer CDR H3 loops the RMS deviation varies considerably and for loops longer than 12 residues most accuracy is lost ! The fact that Gloop-2 CDR loops are shorter than the average for most of the six CDR's also explains why the assumption that loops can be modeled independently into an empty combining site is successful. In Gloop-2 there are very few CDR-CDR interactions, and the above assumption that loops can be modelled independently is valid. For models which have longer CDR H2 and H3 loops the independence assumption is invalid, as shown by the model of 1hil (Section 2.8 and Appendix A.3) were CDR H3 and CDR H2 are intertwined in the final model if the CDRs are modelled independently.

3). The emphasis of the heavy chain selection on CDR H3 (see Section 2.8) results in a deterioration of the RMS deviations for CDR's H1 and H2. There is therefore scope for making the loop overlap independent from the selection of a particular framework. This could be achieved defining a set of fixed loop classes in the standard framework orientation used in A*b*M, defined by cartesian fix-points for the centres of mass and takeoff positions of loops. This is currently being investigated by the author.

4). For long CDR H3 loops the confidence in the model loop is low. The poor correlation between models and crystal structures indicates either that conformational space was not saturated during the conformational search, or that the initial orientation of the loop in the F_V model was incorrect or that multiple, low energy conformations of long loops are possible. The second of the three possibilities is substantiated by the variation of takeoff angles described above, and by the fact that when a crystal structure is modelled, the original antibody loops are rarely selected from the structural database. Because of the many selection and ranking criteria used in CAMAL there is a chance of losing conformations during each of the many processing steps. Therefore new algorithms which are capable of saturating conformational space in a rational way for the complete combining site simultaneously are required. Methods such as minimisation (Moult and James, 1986), Monte Carlo (Garel *et al.*, 1991; Covell, 1992) or Genetic Algorithms (Legrand and Merz, 1992) are promising for the future. Furthermore, distance geometric methods used for solving NMR structures can be used to saturate conformational space for a complete combining site, using database constraints (Havel and Snow, 1991).

5). If the main chain is predicted correctly, the sidechain conformations can be predicted with high confidence as shown for the model of 3D6 in Chapter 2. In order to predict the sidechain conformations correctly terms which describe the accessibility of aromatic residue types in solvent conditions need to be included, as used in the MC program (see Section 2.5 and Appendix B.1). The various annealing parameters for this algorithm have to be optimised in order to reduce the run time for the sidechain generation, which currently is the limiting factor in using this method.

The core of the antibody appears to be well conserved structurally, which is shown by the RMS deviation values in Table 2.9 and Appendix A.3. An accurate prediction of the framework is important for the humanisation procedure presented in Section 3.3.

5.2 Antibody resurfacing

It was shown that there are distinct differences between human and murine F_V surfaces, and that this information can successfully be used for the resurfacing of antibodies. In the case of the N-901 antibody the original functionality of the antibody is retained, but the surface has been changed to that of a human antibody. It remains for the clinical trials to show that a reduction of immunogenicity results from resurfacing the F_V fragment. The homology data also indicate that the current method of chain classification of immunoglobulins may have to be revised, merging some of the classes defined by (Kabat *et al.*, 1992). The homology data also suggest that the surfaces of F_V sequences are conserved during the affinity maturation (somatic mutation) of antibodies, since no difference in surface residues were observed between a set of germline sequences and somatic mutant sequences. The conservation of the immunoglobulin surface may be important for the recognition of *self*, preventing the generation of auto-antibodies.

5.3 Antibody design

A method for the *ab-initio* design of antibodies has been developed, yielding a theoretical antibody which utilises sidechain interactions observed in crystal structures of antibody complexes (bifunctional residue types). Bifunctional residue types are here defined as residues which have two physical properties, eg. Tyr which is both hydrophobic and has an active -OH group. These residue types are mainly selected because the average distance between antibody and antigen is approximately the same as the sidechain length (4.2-4.3 Å) of the selected residue types (His, Phe, Trp, Tyr, Gln, Asn).

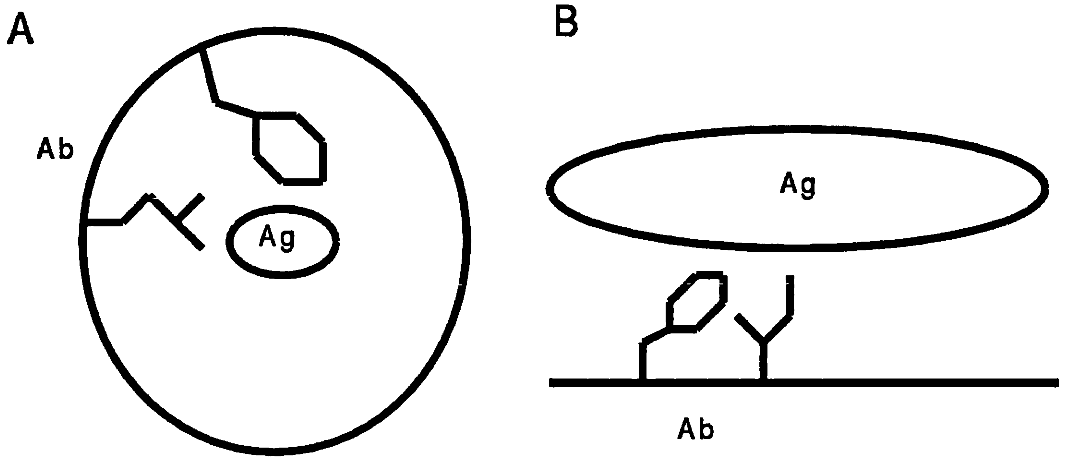


Figure 5.1: Dependence of antibody (Ab) sidechain sidechain interaction on antigen (Ag) size. A) If the antigen is small the interacting sidechains are distributed radially out from the antigen. B) For a large antigen with less curved surface the interacting sidechains are distributed on a planer surface and there is a possibility for sidechain-sidechain interactions.

Energy minimisation of the final ten best constructs of the (GlaMor) F_V model showed that the molecular structure is stable and that no backbone movement is observed. Only small perturbations in sidechain position are seen when the antigen is removed. Surprisingly very few sidechain-sidechain clashes are observed before minimisation. This is probably due to the fact that the antigen is small, that all the sidechain positions are distributed radially from the center of the antigen, and that the sidechain conformations are selected on the basis of antigen burial. Sidechain-sidechain interactions will probably occur as the size of the antigen increases and the antigen-antibody interaction surface area increases and becomes less curved (see Figure 5.1).

The ten F_V constructs described are now being synthesised as single chain F_V constructs (Elliott, 1992). In parallel with this experiment a restricted mutagenesis experiment has been set up, in which all the residue positions which are within range of the morphine molecule are randomly changed and the resulting antibodies screened, using a phage display library.

The main weakness of the design algorithm is the assumption that CDR length is independent of the antigen size. Figures 4.1 to 4.3 show plots of CDR length as a function of antigen molar weight, and indicate that all CDR length are allowed for small antigens but long CDR's are not allowed for large antigens, if good interactions are required. The molar weight is however likely to be a poor indicator of the actual shape of the antigenic epitope which could be an exposed loop, a flat surface or an exposed helix on a large protein. Other structural shape descriptors have to be used in this correlation in order to get a better assessment of the CDR-length/antigen independence hypothesis.

5.3.1 Model quality

The quality of a given antibody (protein) model required depends on the purpose of the model. If the aim of the model is to reshape or resurface an F_V fragment only low resolution data (2-3 Å) is required since only the relative positioning of CDR's and the surface residues is required. The identification of residues which can possibly interact with the CDR's can be obtained from low resolution data. For structural studies, such as protein-protein interactions or F_V design etc, where exact information about sidechain positions is required more accurate and confident models are needed. In the last case structural data of a resolution less than 2 Å is required. Thus, a design should begin with an x-ray structure as the starting scaffold.

Appendix A

Appendix: Immunoglobulin structure & model data

1. Example of sequence entry in the antibody sequence database
2. CDR takeoff angles for the six CDR's of 17 different F_V crystal structures.
3. Results of A*b*M modelling of 16 antibody F_V structures and the comparison to crystal structures

A.1 Sequence database entry

This is an example of a sequnce database entry containing assigned and unassigned descriptors. The format is NBRF (Bleasby, 1990), and descriptors are added as comments.

```
>P1;HHC106
Heavy chains subgroup I V region - Human
- Q I Q L V Q S G G E V K K P G A S V R V S C K A S G Y T F
H S Y G I T - - W V R Q A P G Q G L E W M G W I S G - - Y N
G N T N Y A Q K L Q D R V T M T T D T S T N T V Y M E V R S
L R S D D T A V Y Y C A R D D C S G D N C Y M S - - - - - -
- - - - - - - A Y W G Q G T L V T V S S - - -
*
C;accession:
C;alternate name:
C;antibody-specificity:
C;canon: 1 1 0
C;cdrlength:    5  10  13
C;contains:
C;domain: VH
C;gene name:
C;host:
C;includes:
C;initiation codon:
C;intron:
C;keywords:
C;map position:
C;opttemp:
C;pair_code:
C;pir-name:
C;protein:
C;region:
C;segment number:
C;species: Human
C;superfamily:  Immunoglobulin
```

A.2 Takeoff angles for CDR's of 17 antibody F_V structures

The next six tables contain the tables of takeoff angles of 17 F_V crystal structures. The angles are calculated as the angle between the planes defined by N-terminus and C-terminus and the center of geometry of the backbone of a given loop. The structures have previously been fitted using the multiple fitting program MULFIT which is described in Section 2.

CDR L1

	2fbj	1baf	2fb4	1mam	8fab	1igf	1fdl	5fab	3fab	1hil	2mcp	1dfb	4fab	3hfm	glb2	1f19	2hfl
2fbj	0.000	3.836	7.750	11.299	13.021	18.113	12.988	12.416	7.138	26.760	23.761	11.609	15.236	13.398	9.630	12.903	8.891
1baf	3.836	0.000	5.564	7.489	9.185	14.290	9.390	8.592	5.268	22.950	19.958	7.854	11.541	9.674	5.813	9.142	5.258
2fb4	7.750	5.564	0.000	6.915	9.003	12.925	6.836	7.982	0.775	22.374	19.584	6.590	12.635	7.803	5.933	7.647	4.203
1mam	11.299	7.489	6.915	0.000	2.094	6.820	2.979	1.164	7.498	15.850	12.941	0.903	5.828	2.495	1.679	1.823	2.856
8fab	13.021	9.185	9.003	2.094	0.000	5.229	3.899	1.083	9.593	13.873	10.929	2.544	3.924	2.764	3.509	2.203	4.935
b13i	18.113	14.290	12.925	6.820	5.229	0.000	6.152	5.698	13.639	9.502	6.902	6.646	5.839	5.128	8.483	5.396	9.480
1fdl	12.988	9.390	6.836	2.979	3.899	6.152	0.000	2.975	7.576	15.651	12.972	2.099	7.719	1.294	4.165	1.696	4.133
5fab	12.416	8.592	7.982	1.164	1.083	5.698	2.975	0.000	8.593	14.687	11.783	1.463	4.998	2.023	2.788	1.348	4.004
3fab	7.138	5.268	0.775	7.498	9.593	13.639	7.576	8.593	0.000	23.068	20.259	7.231	13.146	8.513	6.405	8.327	4.713
1hil	26.760	22.950	22.374	15.850	13.873	9.502	15.651	14.687	23.068	0.000	3.014	15.900	11.797	14.578	17.382	14.744	18.677
2mcp	23.761	19.958	19.584	12.941	10.929	6.902	12.972	11.783	20.259	3.014	0.000	13.047	8.793	11.834	14.438	11.938	15.785
1dfb	11.609	7.854	6.590	0.903	2.544	6.646	2.099	1.463	7.231	15.900	13.047	0.000	6.440	1.844	2.204	1.294	2.834
4fab	15.236	11.541	12.635	5.828	3.924	5.839	7.719	4.998	13.146	11.797	8.793	6.440	0.000	6.474	6.751	6.044	8.439
3hfm	13.398	9.674	7.803	2.495	2.764	5.128	1.294	2.023	8.513	14.578	11.834	1.844	6.474	0.000	4.036	0.702	4.532
glb2	9.630	5.813	5.933	1.679	3.509	8.483	4.165	2.788	6.405	17.382	14.438	2.204	6.751	4.036	0.000	3.421	1.746
1f19	12.903	9.142	7.647	1.823	2.203	5.396	1.696	1.348	8.327	14.744	11.938	1.294	6.044	0.702	3.421	0.000	4.105
2hfl	8.891	5.258	4.203	2.856	4.935	9.480	4.133	4.004	4.713	18.677	15.785	2.834	8.439	4.532	1.746	4.105	0.000

MAX ANGLE = 26.760

MIN ANGLE = 0.702

CDR L2

	2fbj	1baf	2fb4	1mam	8fab	1igf	1fdl	5fab	3fab	1hil	2mcp	1dfb	4fab	3hfm	glb2	1f19	2hfl
2fbj	0.000	4.209	9.771	8.940	11.259	5.161	6.808	7.700	96.343	7.619	14.756	11.265	7.845	11.182	12.651	9.691	3.594
1baf	4.209	0.000	5.876	4.765	7.056	2.157	3.949	4.635	92.448	6.269	13.647	9.472	7.431	7.936	10.522	9.087	5.893
2fb4	9.771	5.876	0.000	3.714	3.816	4.689	3.665	2.961	86.614	6.049	11.216	6.907	12.049	8.775	7.085	8.529	11.740
1mam	8.940	4.765	3.714	0.000	2.380	5.108	5.457	5.326	88.422	8.437	14.679	10.281	8.841	5.121	10.685	11.278	9.771
8fab	11.259	7.056	3.816	2.380	0.000	7.013	6.840	6.416	86.105	9.633	14.964	10.707	10.813	5.736	10.761	12.275	12.135
b13i	5.161	2.157	4.689	5.108	7.013	0.000	1.856	2.656	91.184	4.188	11.507	7.314	9.586	9.334	8.388	7.071	7.687
1fdl	6.808	3.949	3.665	5.457	6.840	1.856	0.000	0.904	89.636	2.993	9.913	5.603	11.316	10.256	6.578	5.887	9.527
5fab	7.700	4.635	2.961	5.326	6.416	2.656	0.904	0.000	88.732	3.218	9.596	5.208	11.881	10.307	6.012	5.967	10.341
3fab	96.343	92.448	86.614	88.422	86.105	91.184	89.636	88.732	0.000	89.945	85.764	86.810	96.436	89.800	85.233	89.686	98.141
1hil	7.619	6.269	6.049	8.437	9.633	4.188	2.993	3.218	89.945	0.000	7.438	3.647	13.656	13.234	5.086	2.921	10.948
2mcp	14.756	13.647	11.216	14.679	14.964	11.507	9.913	9.596	85.764	7.438	0.000	4.410	21.070	19.787	4.261	5.089	18.246
1dfb	11.265	9.472	6.907	10.281	10.707	7.314	5.603	5.208	86.810	3.647	4.410	0.000	16.893	15.381	1.578	3.060	14.548
4fab	7.845	7.431	12.049	8.841	10.813	9.586	11.316	11.881	96.436	13.656	21.070	16.893	0.000	6.743	17.878	16.376	5.016
3hfm	11.182	7.936	8.775	5.121	5.736	9.334	10.256	10.307	89.800	13.234	19.787	15.381	6.743	0.000	15.804	16.143	10.234
glb2	12.651	10.522	7.085	10.685	10.761	8.388	6.578	6.012	85.233	5.086	4.261	1.578	17.878	15.804	0.000	4.562	15.837
1f19	9.691	9.087	8.529	11.278	12.275	7.071	5.887	5.967	89.686	2.921	5.089	3.060	16.376	16.143	4.562	0.000	13.218
2hfl	3.594	5.893	11.740	9.771	12.135	7.687	9.527	10.341	98.141	10.948	18.246	14.548	5.016	10.234	15.837	13.218	0.000

MAX ANGLE = 98.141
MIN ANGLE = 0.904

CDR L3

	2fbj	1baf	2fb4	1mam	8fab	1igf	1fdl	5fab	3fab	1hil	2mcp	1dfb	4fab	3hfm	glb2	1f19	2hfl
2fbj	0.000	12.089	14.094	6.790	2.890	4.586	4.953	4.293	3.901	3.292	5.262	3.618	1.449	3.346	5.619	5.682	2.989
1baf	12.089	0.000	2.339	7.133	14.853	8.660	7.731	7.827	8.673	9.875	15.763	8.682	10.649	9.209	17.449	7.448	14.399
2fb4	14.094	2.339	0.000	8.463	16.771	10.295	9.463	9.910	10.864	11.607	17.370	10.820	12.677	11.380	19.288	9.032	16.542
1mam	6.790	7.133	8.463	0.000	8.938	2.225	2.027	4.176	5.544	3.615	8.965	5.081	5.693	5.704	11.174	1.194	9.727
8fab	2.890	14.853	16.771	8.938	0.000	6.740	7.379	7.136	6.785	5.340	3.337	6.507	4.314	6.224	2.733	7.979	3.052
b13i	4.586	8.660	10.295	2.225	6.740	0.000	1.032	2.998	4.152	1.400	7.108	3.600	3.594	4.117	9.067	1.272	7.549
1fdl	4.953	7.731	9.463	2.027	7.379	1.032	0.000	2.302	3.611	2.152	8.039	3.110	3.738	3.711	9.829	0.833	7.812
5fab	4.293	7.827	9.910	4.176	7.136	2.998	2.302	0.000	1.380	3.172	8.721	1.017	2.844	1.666	9.822	3.059	6.645
3fab	3.901	8.673	10.864	5.544	6.785	4.152	3.611	1.380	0.000	3.941	8.861	0.570	2.565	0.572	9.517	4.403	5.743
1hil	3.292	9.875	11.607	3.615	5.340	1.400	2.152	3.172	3.941	0.000	5.922	3.377	2.552	3.715	7.701	2.648	6.281
2mcp	5.262	15.763	17.370	8.965	3.337	7.108	8.039	8.721	8.861	5.922	0.000	8.427	6.324	8.372	3.096	8.351	6.384
1dfb	3.618	8.682	10.820	5.081	6.507	3.600	3.110	1.017	0.570	3.377	8.427	0.000	2.210	0.652	9.230	3.919	5.724
4fab	1.449	10.649	12.677	5.693	4.314	3.594	3.738	2.844	2.565	2.552	6.324	2.210	0.000	2.052	7.028	4.528	4.080
3hfm	3.346	9.209	11.380	5.704	6.224	4.117	3.711	1.666	0.572	3.715	8.372	0.652	2.052	0.000	8.957	4.531	5.194
glb2	5.619	17.449	19.288	11.174	2.733	9.067	9.829	9.822	9.517	7.701	3.096	9.230	7.028	8.957	0.000	10.336	5.111
1f19	5.682	7.448	9.032	1.194	7.979	1.272	0.833	3.059	4.403	2.648	8.351	3.919	4.528	4.531	10.336	0.000	8.585
2hfl	2.989	14.399	16.542	9.727	3.052	7.549	7.812	6.645	5.743	6.281	6.384	5.724	4.080	5.194	5.111	8.585	0.000

MAX ANGLE = 19.288
MIN ANGLE = 0.570

```
CDR H1

        2fbj    1baf    2fb4    1mam    8fab    1igf    1fdl    5fab    3fab    1hil    2mcp    1dfb    4fab    3hfm    glb2    1f19    2hfl

2fbj   0.000   2.748  25.664  13.119  21.408   8.987  88.692 103.611 108.727  15.280  23.208   2.590  18.144  45.547 160.460 170.140  67.866
1baf   2.748   0.000  26.322  12.078  21.371   6.425  88.115 104.317 108.462  14.902  23.045   1.261  17.519  45.041 161.004 172.504  67.142
2fb4  25.664  26.322   0.000  17.395   7.437  26.081 114.351  78.010 134.210  13.166  48.832  27.337  43.743  71.209 134.827 147.453  93.452
1mam  13.119  12.078  17.395   0.000  10.694   9.141  98.716  93.901 119.544   4.231  34.779  13.336  28.881  56.112 149.585 163.852  77.593
8fab  21.408  21.371   7.437  10.694   0.000  19.780 109.193  83.376 129.791   6.693  44.404  22.535  38.831  66.308 139.669 153.351  88.096
b13i   8.987   6.425  26.081   9.141  19.780   0.000  89.586 103.038 110.468  13.131  26.381   7.381  20.252  47.129 158.204 172.900  68.461
1fdl  88.692  88.115 114.351  98.716 109.193  89.586   0.000 167.333  21.211 102.506  65.600  87.053  70.623  43.145 110.818  97.416  21.126
5fab 103.611 104.317  78.010  93.901  83.376 103.038 167.333   0.000 146.155  90.068 126.511 105.347 121.742 149.032  56.850  69.986 171.458
3fab 108.727 108.462 134.210 119.544 129.791 110.468  21.211 146.155   0.000 123.164  85.527 107.333  90.965  63.490  90.528  76.599  42.186
1hil  15.280  14.902  13.166   4.231   6.693  13.131 102.506  90.068 123.164   0.000  37.933  16.107  32.245  59.674 146.243 160.043  81.405
2mcp  23.208  23.045  48.832  34.779  44.404  26.381  65.600 126.511  85.527  37.933   0.000  21.870   6.293  22.522 175.178 160.427  45.038
1dfb   2.590   1.261  27.337  13.336  22.535   7.381  87.053 105.347 107.333  16.107  21.870   0.000  16.432  43.951 162.107 172.730  66.114
4fab  18.144  17.519  43.743  28.881  38.831  20.252  70.623 121.742  90.965  32.245   6.293  16.432   0.000  27.523 178.456 166.664  49.728
3hfm  45.547  45.041  71.209  56.112  66.308  47.129  43.145 149.032  63.490  59.674  22.522  43.951  27.523   0.000 153.941 139.970  22.544
glb2 160.460 161.004 134.827 149.585 139.669 158.204 110.818  56.850  90.528 146.243 175.178 162.107 178.456 153.941   0.000  14.872 131.673
1f19 170.140 172.504 147.453 163.852 153.351 172.900  97.416  69.986  76.599 160.043 160.427 172.730 166.664 139.970  14.872   0.000 118.532
2hfl  67.866  67.142  93.452  77.593  88.096  68.461  21.126 171.458  42.186  81.405  45.038  66.114  49.728  22.544 131.673 118.532   0.000

MAX ANGLE =  178.456
MIN ANGLE =    1.261
```

```
CDR H2

       2fbj   1baf   2fb4   1mam   8fab   1igf   1fdl   5fab   3fab   1hil   2mcp   1dfb   4fab   3hfm   glb2   1f19   2hfl

2fbj  0.000  7.063 10.237 18.408  6.314  7.406  9.960  9.902  5.461  6.221 17.819 12.820 23.436 10.298  3.100  2.676  5.531
1baf  7.063  0.000 12.113 12.784  5.883  8.716  3.007 11.804  4.598  6.958 13.397  7.185 19.864  3.768  6.457  5.968  9.006
2fb4 10.237 12.113  0.000 15.345  6.230  3.435 13.149  0.342 14.243  5.215 12.772 11.829 16.073 12.317  7.425 12.468 15.765
1mam 18.408 12.784 15.345  0.000 12.755 14.031 10.201 15.284 17.369 13.587  3.717  5.683  9.272  9.066 15.998 18.393 21.741
8fab  6.314  5.883  6.230 12.755  0.000  2.852  7.217  5.921  8.526  1.203 11.649  7.630 17.122  6.670  3.394  7.559 11.217
b13i  7.406  8.716  3.435 14.031  2.852  0.000  9.993  3.111 10.888  1.781 12.191  9.569 16.786  9.317  4.358  9.344 12.824
1fdl  9.960  3.007 13.149 10.201  7.217  9.993  0.000 12.881  7.320  8.416 11.400  4.944 18.035  1.364  8.852  8.972 11.884
5fab  9.902 11.804  0.342 15.284  5.921  3.111 12.881  0.000 13.903  4.892 12.778 11.662 16.213 12.068  7.084 12.127 15.429
3fab  5.461  4.598 14.243 17.369  8.526 10.888  7.320 13.903  0.000  9.186 17.928 11.782 24.296  8.314  7.012  2.955  4.632
1hil  6.221  6.958  5.215 13.587  1.203  1.781  8.416  4.892  9.186  0.000 12.207  8.656 17.359  7.868  3.130  7.855 11.439
2mcp 17.819 13.397 12.772  3.717 11.649 12.191 11.400 12.778 17.928 12.207  0.000  6.459  6.661 10.074 15.041 18.353 21.927
1dfb 12.820  7.185 11.829  5.683  7.630  9.569  4.944 11.662 11.782  8.656  6.459  0.000 13.105  3.635 10.624 12.710 16.077
4fab 23.436 19.864 16.073  9.272 17.122 16.786 18.035 16.213 24.296 17.359  6.661 13.105  0.000 16.694 20.439 24.377 28.030
3hfm 10.298  3.768 12.317  9.066  6.670  9.317  1.364 12.068  8.314  7.868 10.074  3.635 16.694  0.000  8.785  9.644 12.774
glb2  3.100  6.457  7.425 15.998  3.394  4.358  8.852  7.084  7.012  3.130 15.041 10.624 20.439  8.785  0.000  5.046  8.467
1f19  2.676  5.968 12.468 18.393  7.559  9.344  8.972 12.127  2.955  7.855 18.353 12.710 24.377  9.644  5.046  0.000  3.676
2hfl  5.531  9.006 15.765 21.741 11.217 12.824 11.884 15.429  4.632 11.439 21.927 16.077 28.030 12.774  8.467  3.676  0.000

MAX ANGLE =  28.030
MIN ANGLE =   0.342
```

CDR H3

	2fbj	1baf	2fb4	1mam	8fab	1igf	1fdl	5fab	3fab	1hil	2mcp	1dfb	4fab	3hfm	glb2	1f19	2hfl
2fbj	0.000	2.840	3.749	14.126	11.668	14.500	9.501	15.918	21.569	23.308	29.881	28.064	29.222	13.648	17.147	58.537	70.889
1baf	2.840	0.000	6.284	15.816	14.208	16.826	12.285	18.710	24.385	26.142	32.570	30.838	31.808	10.859	14.311	61.355	73.384
2fb4	3.749	6.284	0.000	10.760	7.947	10.757	6.200	12.578	19.004	20.195	27.879	25.712	25.536	16.507	20.411	55.818	69.287
1mam	14.126	15.816	10.760	0.000	6.717	4.898	9.957	11.846	20.183	18.650	29.867	26.710	18.759	23.003	27.712	53.613	70.682
8fab	11.668	14.208	7.947	6.717	0.000	3.424	3.720	5.998	14.116	13.680	23.775	20.879	17.601	23.833	28.061	49.521	65.097
b13i	14.500	16.826	10.757	4.898	3.424	0.000	7.127	7.054	15.352	13.784	25.013	21.821	15.372	25.693	30.152	49.064	65.794
1fdl	9.501	12.285	6.200	9.957	3.720	7.127	0.000	6.425	13.110	13.995	22.401	19.920	20.046	22.707	26.555	49.694	63.935
5fab	15.918	18.710	12.578	11.846	5.998	7.054	6.425	0.000	8.341	7.791	18.034	14.978	14.258	29.053	32.972	43.669	59.161
3fab	21.569	24.385	19.004	20.183	14.116	15.352	13.110	8.341	0.000	4.857	9.696	6.831	15.949	35.213	38.536	36.969	51.020
1hil	23.308	26.142	20.195	18.650	13.680	13.784	13.995	7.791	4.857	0.000	12.317	8.722	11.102	36.701	40.453	35.880	52.045
2mcp	29.881	32.570	27.879	29.867	23.775	25.013	22.401	18.034	9.696	12.317	0.000	3.667	21.781	43.385	46.140	29.785	41.534
1dfb	28.064	30.838	25.712	26.710	20.879	21.821	19.920	14.978	6.831	8.722	3.667	0.000	18.196	41.697	44.776	30.691	44.219
4fab	29.222	31.808	25.536	18.759	17.601	15.372	20.046	14.258	15.949	11.102	21.781	18.196	0.000	41.066	45.499	36.342	56.264
3hfm	13.648	10.859	16.507	23.003	23.833	25.693	22.707	29.053	35.213	36.701	43.385	41.697	41.066	0.000	4.847	72.175	83.835
glb2	17.147	14.311	20.411	27.712	28.061	30.152	26.555	32.972	38.536	40.453	46.140	44.776	45.499	4.847	0.000	75.441	85.640
1f19	58.537	61.355	55.818	53.613	49.521	49.064	49.694	43.669	36.969	35.880	29.785	30.691	36.342	72.175	75.441	0.000	22.898
2hfl	70.889	73.384	69.287	70.682	65.097	65.794	63.935	59.161	51.020	52.045	41.534	44.219	56.264	83.835	85.640	22.898	0.000

MAX ANGLE = 85.640
MIN ANGLE = 2.840

A.3 Modelling of 16 F$_V$ structures

This appendix contains the results from the comparison of 16 crystal structures to the model generated using A*b*M (OML, 1992) v 1.0.

Struct	Loop	CDR Len.	Run Times		Global RMS		Local RMS	
			Csearch	Eureka	Ca	Backbone	Backbone	All
glb2	L1	11	CANONICAL		1.103	1.161	0.801	2.898
	L2	7	CANONICAL		0.631	0.647	0.228	0.654
H - 1baf	L3	9	CANONICAL		1.003	1.031	0.297	1.104
L - 3671	H1	5	CANONICAL		1.884	1.785	0.204	1.664
	H2	10	CANONICAL		1.543	1.609	0.624	2.997
	H3	4	3.6	7.2	1.172	1.273	0.652	3.478
	Total		3.6	7.2	1.222	1.251	0.467	2.132
2hfl	L1	10	CANONICAL		1.140	1.150	0.479	0.812
	L2	7	CANONICAL		0.709	0.712	0.320	1.237
H - 1f19	L3	8	272.7	516.1	2.462	2.524	0.880	2.341
L - 1baf	H1	5	CANONICAL		1.176	1.261	0.696	2.301
	H2	10	CANONICAL		2.202	2.155	0.624	2.455
	H3	7	1.1	10.3	2.315	2.310	1.195	2.349
	Total		273.8	526.4	1.667	1.685	0.699	1.915
2mcp	L1	17	CANONICAL		0.720	0.784	0.546	1.303
	L2	7	CANONICAL		0.613	0.538	0.436	0.955
H - 1mam	L3	9	CANONICAL		0.718	0.739	0.242	1.052
L - 1hil	H1	5	CANONICAL		0.968	1.004	0.275	1.492
	H2	12	CANONICAL		2.022	2.014	0.787	1.623
	H3	11	258.8	824.6	2.238	2.306	0.544	2.520
	Total		258.8	824.6	1.213	1.231	0.471	1.490
4fab	L1	16	39.8	35.3	2.317	2.470	1.694	2.923
	L2	7	CANONICAL		0.768	0.792	0.279	1.144
H - 1mam	L3	9	CANONICAL		1.231	1.255	0.625	1.900
L - b13i	H1	5	CANONICAL		0.672	0.721	0.378	1.867
	H2	12	CANONICAL		1.922	2.028	0.787	3.465
	H3	7	1.5	7.6	2.140	2.132	0.519	2.889
	Total		40.3	42.9	1.508	1.566	0.714	2.364
2fbj	L1	10	CANONICAL		1.681	1.733	0.615	1.052
	L2	7	CANONICAL		0.893	0.867	0.320	2.812
H - 1hil	L3	9	238.7	735.3	1.581	1.724	0.585	2.414
L - 2hfl	H1	5	CANONICAL		0.502	0.515	0.142	1.520
	H2	10	CANONICAL		0.767	0.779	0.455	3.076
	H3	9	986.7	3888.3	4.017	4.171	1.839	4.437
	Total		1225.4	4623.6	1.574	1.632	0.659	2.551
3671	L1	11	CANONICAL		0.848	0.788	0.631	2.429
	L2	7	CANONICAL		0.293	0.304	0.252	1.031
H - 1hil	L3	9	CANONICAL		1.160	1.131	0.930	2.047
L - 1mam	H1	5	CANONICAL		1.358	1.341	0.434	2.002
	H2	10	88.1	583.8	4.198	4.099	1.086	3.398
	H3	12	2072.8	1136.7	4.363	4.192	2.794	5.078
	Total		2160.9	1720.6	2.036	1.752	1.021	2.664
3d_6	L1	10	CANONICAL		0.871	0.834	0.508	3.256
	L2	7	CANONICAL		0.738	0.750	0.228	1.166
H - 8fab	L3	7	0.5	1.2	1.894	1.878	0.994	3.311
L - 1rei	H1	5	CANONICAL		0.736	0.736	0.206	0.935
	H2	10	CANONICAL		1.180	1.202	0.455	2.980
	H3	17	452.0	357.0	5.420	6.163	3.256	5.047
	Total		452.5	358.2	1.807	1.927	0.941	2.782
3hfm	L1	11	CANONICAL		0.801	0.775	0.508	2.880
	L2	7	CANONICAL		0.978	1.021	0.437	2.064
H - 1baf	L3	9	CANONICAL		0.426	0.394	0.234	1.463
L - 3671	H1	5	CANONICAL		2.037	2.012	0.937	2.098
	H2	9	CANONICAL		0.914	0.942	0.561	1.406
	H3	5	0.7	5.5	1.501	1.683	0.845	2.542
	Total		0.7	5.5	1.110	1.138	0.587	2.075

Table A.1:

Struct	Loop	CDR Len.	Run Times		Global RMS		Local RMS	
			Csearch	Eureka	Ca	Backbone	Backbone	All
1mam	L1	11	CANONICAL		1.251	1.302	0.742	2.812
	L2	7	CANONICAL		1.361	1.362	0.252	0.966
H - 2mcp	L3	9	CANONICAL		1.270	1.289	0.260	0.704
L - 1f19	H1	5	CANONICAL		1.849	1.845	0.275	1.218
	H2	12	76.1	189.5	2.852	2.976	1.566	4.016
	H3	8	7.3	61.8	2.448	2.524	1.322	2.499
	Total		83.4	251.3	1.839	1.883	0.736	2.035
b13i	L1	16	31.1	71.5	2.585	2.667	1.478	3.066
	L2	7	CANONICAL		0.749	0.763	0.279	1.144
H - 1hil	L3	9	CANONICAL		0.888	0.877	0.416	1.081
L - 4fab	H1	5	CANONICAL		1.335	1.310	0.266	1.664
	H2	10	CANONICAL		1.185	1.202	0.284	2.695
	H3	10	144.9	400.6	2.894	2.970	1.997	3.201
	Total		176.0	472.1	1.606	1.632	0.787	2.142
2fb4	L1	13	CANONICAL		0.755	0.780	0.235	2.088
	L2	7	CANONICAL		1.172	1.247	0.966	2.308
H - 8fab	L3	11	165.7	1062.8	1.612	1.730	0.920	2.368
L - 2rhe	H1	5	CANONICAL		0.626	0.621	0.096	0.352
	H2	10	CANONICAL		0.708	0.726	0.200	3.352
	H3	17	172.2	967.8	4.024	4.224	3.107	4.137
	Total		337.9	2030.6	1.483	1.555	0.921	2.434
d1_3	L1	11	CANONICAL		0.799	0.799	0.267	3.107
	L2	7	CANONICAL		0.944	0.928	0.502	1.611
H - 8fab	L3	9	CAN0NICAL		1.126	1.138	0.652	2.243
L - 1f19	H1	5	CANONICAL		0.869	0.846	0.448	0.794
	H2	9	CANONICAL		1.369	1.413	0.634	3.336
	H3	8	1501.8	860.6	2.069	2.188	0.627	2.925
	Total		1501.8	860.6	1.196	1.219	0.521	2.336
8fab	L1	11	2.7	18.7	4.022	4.021	2.733	4.925
	L2	7	733.4	4410.2	0.581	0.592	0.347	1.213
H - 2fb4	L3	9	4.7	7.9	1.215	1.218	0.745	4.215
L - 2fb4	H1	5	CANONICAL		1.103	1.106	0.096	0.330
	H2	10	CANONICAL		1.183	1.195	0.200	2.848
	H3	12	218.5	1386.5	5.978	5.974	2.753	4.857
	Total		959.5	5823.4	2.347	2.351	1.145	3.065
1hil	L1	17	CANONICAL		1.148	1.151	0.546	1.467
	L2	7	CANONICAL		0.917	0.922	0.436	1.634
H - b13i	L3	9	CANONICAL		1.278	1.288	0.242	1.035
L - 2mcp	H1	5	CANONICAL		0.565	0.563	0.266	1.083
	H2	10	CANONICAL		0.845	0.829	0.284	2.711
	H3	11	9460.9	2742.9	4.621	4.623	2.779	5.652
	Total		9460.9	2742.9	1.562	1.562	0.758	2.263
1baf	L1	10	CANONICAL		0.889	0.883	0.000	0.000
	L2	7	CANONICAL		1.296	1.266	0.000	0.000
H - 3hfm	L3	10	66.5	259.6	3.519	3.397	1.689	2.973
L - 2hfl	H1	6	CANONICAL		2.040	2.136	1.487	1.373
	H2	9	CANONICAL		1.975	1.952	0.000	0.000
	H3	6	0.9	0.0	3.031	2.794	1.994	3.803
	Total		67.4	259.6	2.125	2.125	0.862	1.358
1f19	L1	11	CANONICAL		1.161	1.226	0.801	2.885
	L2	7	CANONICAL		0.915	0.966	0.502	1.725
H - 2hfl	L3	9	CANONICAL		1.382	1.389	0.931	2.055
L - 1mam	H1	5	CANONICAL		4.444	4.601	0.696	1.280
	H2	10	499.2	1872.7	5.120	5.336	1.056	2.911
	H3	15	188.0	735.5	8.907	9.095	4.075	5.401
	Total		687.2	2608.2	3.655	3.769	1.343	2.710

Table A.2:

Appendix B

Appendix: Program documentation

This appendix contains the program documentation for three of the programs used in the antibody design process:

1. **MC** A complete Monte Carlo simmulate annealing program, used for the reconstruction of sidechain conformations. This package has several other features such as a complete molecular drawing package.
2. **INT** A menu driven protein-protein interaction and protein surface investigation program.
3. **CLUSTER** A menu driven torsional clustering program for the evaluation of large loop ensembles.
4. **FRAMEBUILD** Antibody framework construction program.

B.1 Simulated annealing package for side chain reconstruction

B.1.1 Introduction

This Section contains the documentation to the MC (Monte Carlo) program developed during the course of this PhD thesis. The program has developed into a general tool for molecular modelling. The program is the platform into which most of the programs which I have written have been implemented.

The primary target for the program is in the prediction of sidechain conformations in proteins. The program currently contains three algorithms for sidechain conformation generation, 1) Monte Carlo simulated annealing method. 2) Torsional grid searching. and 3) A Torsional grid searching, with all possible sidechain types. The various methods are outlined in detail in the sections below, and in the main body of the thesis.

B.1.2 Simulated annealing

Simulated annealing is a method which is frequently used for solving the traveling salesman problem, with the mississippi river twist: How does a salesman visit N towns taking the shortest possible route, and only visiting each town once, and giving a penalty for crossing the river. This type of problem is called an NP-complete problem. The time taken to compute the exact solution to this problem is $K{\cdot}e^N$ where N is the number of unknowns in the problem, and K is a machine dependent constant. Even with a small number of variables a combinatorial explosion is observed in the number of possible solutions.

J. von Neumann and S.M. Ulam introduced around 1945 the Monte Carlo method of solving problems which have a large solution space, they showed that a solution could be computed by using a random walk through the solution space, a practical approach was outlined by Metropolis (Metropolis *et al.*, 1953). Instead of computing the analytical solution, a solution is generated by random sampling of the solution space. Metropolis developed the method further by introducing a probability density function, and an objective evaluation function E, introducing the method of simulated annealing or a simulation of a cooling process. The result becomes a biased random walk, having an initial state where all moves are allowed. By slowly lowering the probability for accepting an unfavorable move the system is moved towards a global minimum.

In terms of molecular structure determination the objective evaluation function is an energy function, and the probability function is derived from the Bolzman distribution. Assuming that a given molecular structure is will adopt a conformation which is a global minima and well "packed" (no space between the atoms), a simple energy function can be used for the evaluation of the Metropolis probability:

$$E = \varepsilon_o \sum_{i=1}^{n}((\frac{r_o}{r})^6 - 2(\frac{r_o}{r})^{12}) + \kappa_o \cdot cos(3\omega)$$

Where the first term is a simple *Lennard-Jones* potential which evaluates the non-bonded contacts between the atoms in a given molecule, the second term is a simple torsional term which only applies to C-C bonds. The torsional term biases the function towards 60° rotamers. ε_o and κ_o are constants. The Metropolis function:

$$P = e^{\frac{-\delta E}{T}}$$

is then used to evaluate the energy function. This simple method can be used to search the large conformational space defined by a set of torsion angles in amino-acid sidechains, and find or define the global minima which exist for a given set of sidechains. It is necessary to emphasise that the Metropolis method of simulated annealing not is a minimisation, it is merely a biased random walk. The value T is the simulation parameter which determines how fast the function should approach a minimum. In the case of thermic motion this is a temperature, thus the denotation T. In the following we will call this the simulation temperature.

When searching sidechain conformations using this method the simulation system usually get trapped in an energetic minima well before the global minimum is encountered, at a high temperature, without the solution space having been searched sufficiently. This can be overcome by truncating the *Lennard-Jones* potential, in order to allow atoms to pass through each other. In reality this function would converge towards infinity when the distance r between the atoms goes towards zero.

The torsional potential is precalculated and only updated every 10 steps since the average movement over 10 random steps is no more than $10 \cdot \sqrt{10}$ the precision of the energy calculation is maintained. Why is it necessary to have the torsional potential at all ? The potential does only have little influence on internal side chain conformations, but becomes significant for surface sidechains.

B.1.3 Evaluation of conformations

Evaluation of side chain conformations is done purely on an energetic basis for internal (core) residues, good van der Waals interactions are considered to be equal to a good packing of the residues. The situation gets more complicated when trying to predict the conformation of surface residues. Many low energy conformations are possible on the molecular surface, all which have a good packing.

Using the fact that hydrophobic, bulky residues will be shielded by the hydrophilic sidechains, and be buried in the surface, it is possible to generate simple functions which will take these rules into account. These functions can either be implemented in the objective evaluation function of the MC simulation, or as is done here, added as a post processing step. Including a accessibility/hydrophobicity term in the evaluation function would slow down the calculations to much, this is why this term has been added as a post processing step.

In the functions used here the accessibilities and the hydrophobicities have been scaled appropriately. All accessibilities are relative to the accessibility of an extended conformation of the amino acids, and thus in the range [0;1]. Hydrophobicities are taken from ref (Cornette *et al.*, 1987), but have been normalised to be in the range [-1;1]. The simplest type of function can be either of two:

1. $f_a = -1 \cdot \frac{A_{rel}}{H_{rel}} \qquad f_a \in \,]-\infty;\infty[$

or

2. $f_a = -1 \cdot A_{rel} \cdot H_{rel} \qquad f_a \in [-1;1]$

The main difference between the two functions above is the ranges in which they are defined. In the first case the score for an favorable conformation is exponential, where as in the other case the score is linear to the relative exposed area of a given group. The first function is not defined for H_{rel} or A_{rel} equal to zero. The second function is a continuous function in the range $[-1;1]$. The two functions have been implemented in the **HPHACCESSIBILITY** option in the calculation option. Both values will be calculated if this option is chosen. The surface area is calculated using the tessellated icosahedron approach, which is not too exact, but it is quick.

Similar semi-analytical expressions have been suggested by Still *et al* (Still *et al.*, 1990). These have been included in energy calculations and have been shown to be able to generate conformations of sidechains which are close in conformation to what is observed in crystal structures. The traditional (Still *et al.*, 1990) perception of solvation free energy (G_{sol}), as consisting of the term:

$$G_{sol} = G_{cav} + G_{vdW} + G_{pol}$$

G_{cav} is a solvent cavity term, G_{vdW} is a solute van der Waals term, and G_{pol} is a solute solvent electrostatic term. For saturated hydrocarbons in water G_{sol} is linearly related to the solvent-accessible surface area A_s.

B.1.4 Program documentation

The program has been written such as to be as flexible a possible since I had several ideas with the basic program, and the program is developing all the time.

The program is also an attempt to write a good parser - in this case I have used a three dimensional command space defined by the array *com* in the include file

"Par.h". There are three arrays the first *com* contains the command mask passed to the parser, the second (*defaults*) contains a static mask defining any defaults, this is not used a lot at the moment - but is necessary in order to avoid any syntactical mistakes. The third array *command* contains the actual commands which the program understands.

The first word is the key command, and any subsequent commands are children of this command. There are no *required* sub-commands, for example has the **ANNEAL** command at the moment twelve sub-commands any of these sub-commands are optional. The only requirement is that, if multiple sub-commands are given they have to be given in the order stated in the documentation or the builtin **HELP** command. All file names have to be in quotes. The reason for this is that the program is not case sensitive, all passed words are capitalised (except for arguments which are lowercased). All the sub-commands in the **ANNEAL** command are independent commands and can all be stated at the same time. Each level of sub-commands can have several different optional modifiers, e.g. **READ** can have **PDB**, **VDW**, etc as the format modifier, but only one of these can be stated at the time.

The following section contains a summary of all the currently supported commands.

B.1.5 Command summary

BYE,QUIT,EXIT Any of these commands will halt the program and exit;

HELP This help is a hard-wired help which dumps the current *command* array - so this is where to look if you are in doubt about any given command is supported on the machine you are actually on. Any new commands will also occur here. This help will also give the order of sub commands.

MALLINFO This facility will dump the *mips* supported mallinfo structure which gives information about the current program arena, this facility is at the moment only supported for the ESV workstation and the SGI machines. The NeXT and HP700 does unfortunately not have this facility - but you can use MallocDebug on the NeXT which is much better.

READ **READ** is the command to read data into the program. Following types of data can be loaded:

The valid sub commands are:

- **PDB** This is the default, expects the file name to be a valid brookhaven file, which exists in the search path. The name of the structure object generated by this command is by default the name of the file specified, if an other name is required the modifier **OBJECT** $< objectname >$ may be used to set the name of the object.
- **ORDER** Reads an atom order file, specifying the order of the atoms required by the program. his order is C,O,N,Cα,-Sidechain. This order makes the programming more easy.
- **CHI** Reads chi angle definition file, which defines the number of chi angles in each of the 20 amino-acids.

- **VDW** Reads atom van der Waal radii and the atom pair constant ε_o.

- **CONFORMATIONS** Reads a set of conformations generated by the program. The conformations are read into a dynamically allocated structure list - and there is no check whether you actually have the required space - so with a lot of conformations you might be swapped out. the command takes two file type modifiers.

  - **COORDINATES** Expects the conformations to be in ascii coordinate form. I have never tried this so I do not know how well it works. The problem is that it takes up a lot of space to store conformations this way. If the modifier **NUMBER** n is added only conformation n is extracted.
  - **TORSION** Expects the conformations file to contain the torsions of each of the conformations, the residue numbers and the total energy of this conformation. This is the default.

- **RADII** Reads the single atom radii used for the generation neighbor lists.

- **ACCESS** Expects a file containing the residues and the accessibility of an extended conformation of each of the 20 amino acids. This is used for the calculation of the relative exposed surface area. Although this is should not be used for exact calculations. The **CALCULATE RELATIVE ACCESSIBILITY** command will not use this data, but will calculate the actual accessibility of a blocked amino acid in a given conformation.

- **HPH** Contains the scaled hydrophobicities for each of the 20 amino-acids. The hydrophobicities have been scaled such that they lie in the range $[0, 1]$.

- **ATOMICHPH** A file containing atomic charge or hydrophobicity parameters.

- **FRAGLIB** A set of rigid sidechain fragments used for searching and building sidechain conformations.

- **DAYHOFF** A Dayhoff mutation matrix.
- **CSSR** Coordinate file from the Cambridge Crystallographic database, in ASCII format.
- **PSDAT** PostScript plot data file. This file contains all the information relevant for a plot. this is only data relevant for the layout of the plot.
- **PLDAT** Plot data file. Data relevant for the presentation of the plotted molecule, such as bond width atomic radii and colours.
- **SEARCHITS** Reads a file of distance geometry search hits. This file is generated by the **SEARCH** command.

Syntax :

READ [subcommand] < *filename* > [file type] [object] < *objectname* >

WRITE This is the main output command and has the following sub-commands:

- **COORDINATES** Specifies that the following type of file will be in coordinate format. **OBJECT** < *objectname* > specifies the object from which the coordinates should be written. **PDB** specifies that the object is written to a Brookhaven Database format file. **PDBACCESS** will write an extended Brookhaven format file which has potential parameters and accessibility data added in extra columns.
- **TORSIONS** Specifies that the coordinate format is torsions. Selecting this option will write or initiate the writing of conformations as torsions only, in the range $[0, 2{\cdot}\pi]$.

- **DISTMATCH** The name of a file to which hits should be written during a distance geometry search.
- **POSTSCRIPT** A file to which a specified plot object should be written.

Each of these format specifiers can be applied to the two output types :

- **PDB** Write a brookhaven file
- **CONFORMATIONS** This will not write anything, but will allocate a file pointer to a conformations file, which will be used for the dumping of conformations in an annealing run.

Syntax:

WRITE [format] [file type] < *filename* >

ORDER **ORDER** will order the atoms in each residue according to the atom order specified in the *order* file. If the atom order is unknown, or you are unsure of the order always use this command after having read in the coordinates of a structure. At the moment the only supported format is brookhaven. order can also be used to order a file of distance geometry search hits. The hits are ordered after RMS deviation between the search constraints and the atoms in the hit, this is done with the **TRANSFORMATION** modifier.

Syntax:

ORDER [format]

SETUP **SETUP** will initialise various data structures and set them up appropriately for the calculation routines to handle them. The possible **SETUP** modifiers are :

- **TORSION** Will setup the torsional potential with a pre-defined grid, the default is 10° and the energy constant A is 1.5 kcal. These two parameters can be modified with the modifiers:
    - **ANGLE** New angle.
    - **ENERTOR** New torsional energy constant.
- **NAYBLIST** Sets up a neighbor list with a given cutoff, specified by the modifier **CUTOFF**, the default cutoff is 5 Å.
- **SELNAYB** Setup of neighbor list for the selected residues - this is used when processing many conformations, the routine should be called each time a new conformation is generated. The list cutoff is specified by adding the **CUTOFF** modifier.
- **ICOSAHEDRON** Setup the unit icosahedron, and use this in the following surface and accessibility calculations. The precision of following calculations depends on the tessellation frequency γ. There are two modifiers to this setup command:
    - **CUTOFF** This is a generation parameter which is used to eliminate overlapping vertices generated by the algorithm, the default value is 0.1 Å- it should not be necessary to fiddle around with this parameter.
    - **TESSELATION** This is the specification of precision. The number is a value in the range $[1,\infty]$, normally this value is 4 and will generate an icosahedron with 162 vertices.

- **RADII** Assigns vdw radii to the current atom list - this always has to be done before an annealing run.
- **DISTANCEMATRIX** Generates a distance matrix from a molecular object. The matrix can the be used to search another distance matrix.
- **CSSR** Generates a molecular object in Cambridge database format. This format is required for the **PLOT** command to work.

Syntax:

SETUP [setup item] [setup param 1] < *value*1 > [setup param 2] < *value*2 >....

SELECT The select command is used to select the atoms which are to be used in any calculation or annealing run. At the moment the the second modifier is sort of redundant, the selection of residues and not be replaced by anything else, the reason for this peculiarity is that I had the intention of implementing some sort of selection stack, but it did not get further than to the thinking stage. The residues can be selected in three different ways:

- **FILE** A file containing the numbers of the residues to be selected the. The format of the file is free, the numbers just have to be separated by spaces or control characters such as newlines or tabs.
- **ALL** Selects all the atoms currently in the list.
- **RESIDUE** Selects all the residues which have have a sidechain within a certain distance cutoff of the specified sidechain. This is probably the most useful of the selection commands.

The command can also be used for the colour assignment of atoms used in the **PLOT** command.

Syntax :

SELECT [RESIDUE] $< method, rgbcolours >< cutoff >$

SELECT [COLOUR] $< rgbcolours >$ **ATOM** $< atno >$ **TO** $< atno >$ **OBJECT** $< objectname >$

<u>ANNEAL</u> **ANNEAL** is the actual annealing command, which initialises the annealing run, no command specifiers are necessary, but any of the parameters in the simulation can be changed, **ANNEAL** takes any of the following commands. Several, commands or all can be specified at the same time. The only requirement is that the commands should be given in the order they are mentioned in the list below.

- **RESTART** is a restart flag - this is set to '1' in order to restart the program after system crash or other stops of the program. You have to extract the last written conformation from the conformations file, and set **RESTART** to '1'. Default is '0'.
- **ETRUNC** VdW potential truncation energy, the default value is 7 kcal.
- **NMOV** Number of move steps per T step the default is 10.000
- **NUPD** Number of steps between each update of the torsion potential.
- **DT** Temperature gradient in percent, default is 2 percent.

- **AMOV** Move angle per step, default is 10°.
- **PACCI** Initial acceptance probability, default is 75 percent.
- **PACCF** Final acceptance probability, default is 50 percent.
- **DCUTT** Distance cutoff for neighbor list updating, 5 Åis default.
- **NEQ** Number of random steps per T step in equilibration, default is 100. Higher number of steps might be desirable, it seems that the actual annealing starts at a to high temperature, this can be remedied by using a higher number of steps in the equilibration.
- **TINIT** Initial temperature, this is just a high temperature, which by default is 10.000 K°, this temperature does not really have any significance, except that the program will spend a lot of time in the equilibration, in stead of in the annealing.

Syntax :

ANNEAL [modifier 1] $< modifiervalue >$ [modifier 2] $< modifiervalue >$ ...

CALCULATE This command supplies a analysis interface which is probably going to develop quite a lot, since this is where the actual selection of conformations is happening. At the moment two types of calculations are possible:

- **ACCESSIBILITY** Calculate the accessibility of a given conformation, either the exact accessibility in Å^2 or as a relative fraction compared to the accessibility of an extended conformation.

- **HPHACCESSIBILITY** This is an attempt to develop a function which will honor the fact that hydrophobic residues will be buried, and hydrophilic are exposed. Both of the previously defined functions are calculated. The function will be calculated for each of the individual residues involved, but also for the conformation as a whole.

Each of the evaluation functions can be evaluated as either **EXACT** or **RELATIVE**, followed by the number of the conformation.

If the **ALL** specifier is used both of the functions will be evaluated for all of the conformation. In this case the **TESSELATION** frequency of the icosahedron used and the distance **CUTOFF** must be specified.

Syntax for single conformation:

CALCULATE [function] [exact/relative] $< conformationnumber >$

Syntax when using **ALL** option.

CALCULATE [function] **ALL TESSELATION** $< tesselation >$ **CUTOFF** $< cutoff >$

<u>**EXTRACT**</u> Conformation extraction routine. This routine will take a specific conformation from the current ensemble and regenerate it in the molecular structure. This is currently only supported for the torsion type of conformations. Implementation for the coordinate type of conformations is simple its just to patch the bits of lists into the main list, and free up the old bits. But since we

are working in torsion space, I do not want to do it in cartesian space.

Syntax :

EXTRACT CONFORMATION < *conformationnumber* >

CONDENSE This is a redundant command which can be used to condense the list of atoms to one which only includes atoms which are actually involved in the molten zone. A much better routine is implemented in the *Anneal* function.

The modifier **CUTOFF** can be used here.

DELETE This command is used to delete an object from the index list of the program. The objects which can be deleted are: **PDB,CONFORMATION,SELRES** and **SIDECONF**.

Syntax:

DELETE < *objecttype* >

SIDECONSTRUCT With this command sidechains can be reconstructed in a torsional grid. The following modifiers can be used:

- **ETRUNC** Truncation energy for VdW repulsion.
- **MAXCONF** The maximum number of conformations to generate. The search will terminate after **MAXCONF** is reached.

- **AMOV** The angular grid to searched.
- **DCUTOFF** Distance cutoff to be used in the update of neighbor lists.
- **ECUTOFF** Energetic cutoff to be used for the rejection of conformations and termination of search tree.
- **BONDLENGTH** the length of a bond in an aliphatic sidechain branch.
- **VDWRADIUS** The radius of carbon atoms in the aliphatic sidechain branch.

Syntax:

SIDECONSTRUCT [modifier 1] $< modifiervalue >$ [modifier 2] $< modifiervalue >$ ...

<u>DUMP</u> **DUMP** Is used to inquire the program about I'ts state with respect to held objects and their contents.

- **INDEX** Dumps the current index list.
- **FRAGLIB** Lists the contents of the currently held fragment library.

Syntax:

DUMP [list]

SEARCH A general command for the searching of distance geometry. There are two searches which can be done at the moment. **PEPTIDE** will search a distance matrix for matching amino acids in 3D space. The use of this is in design of peptide mimetics, which are similar to patches of protein surfaces. **STRUCTURE** will search two distance matrices against each other.

The following specifiers are available:

- **DCUT** Distance cutoff for acceptance of hit.
- **MINLENGTH** The minimum length of a given peptide to match a peptide specified in a **PEPTIDE** search.
- **SCUT** Score cutoff for acceptance of a hit.
- **ACCESS** Specifies whether accessibility should be included in the scoring scheme of a **PEPTIDE** type of search.

Syntax:

SEARCH [PEPTIDE,STRUCTURE] [modifier 1] $< modifiervalue >$ [modifier 2] $< modifiervalue >$ ...

PLOT The molecular plotting routines implemented in MC are accessed through this command. Three types of objects can be plotted with this command. **OBJECT** plots a specified molecular object, **LABEL** a label, and **HBOND** a hydrogen bond. The basic plots can then be modified using the following modifiers:

- **PSDAT** Specifies which PostScript data is to be associated with a given plot.
- **PLDAT** Specifies which plot data is to be associated with a given plot.
- **RESIDUE** Specifies which residue a label or a hydrogen bond is associated to.
- **ATOM** Is the atom to which the above label or hydrogen bond is associated.
- **ARRANGE** Specifies whether a given plot should be used for scaling of the global plot.
- **BSCALE** Specifies that temperature factors should be used for the scaling of atomic radii.
- **XOFFSET,YOFFSET** Is used to offset a label.
- **CPK** Makes nice CPK presentation type of plots.

Syntax:

PLOT [OBJECT,LABEL,HBOND] [modifier 1] $< modifiervalue >$ [modifier 2] $< modifiervalue >$ ...

MERGE Merge is used to merge two plot objects. The purpose of this command is to enable superimposition of plotted structures. The plots are sorted according to depth such that real superimposition is obtained.

Syntax:

MERGE PLOT $< plot1 >$ **AND** $< plot2 >$

TRANSFORM **TRANSFORM** is used to transform a molecular object with one or a list of transformation matrices. The transformed objects are then written to a set of numbered files. The format of these files is Brookhaven format.

Syntax:

TRANSFORM TYPE OBJECT < *objectname* > **FILEPREFIX** < *prefix* >

B.1.6 The data files

Several data files are needed in order to run the program they are listed below and the format of each type is described.

- **acces.dat** A file of containing the accessibility for a set of amino acids the format is free. each line consists of the three letter code of an amino acid and a value.
- **chi.dat** A file containing the number of chi-angles in a given amino acid. The format is the same as above.
- **pdb.dat** Residue atom order file. This file contains the reference order required for the program to run. The order is C,O,N,Cα, sidechain. The format is a header followed by a star and then the atom order for for each of the residues in A5 type of format.
- **radii.dat** File of real van der Waal radii. the format is described in the file, and consists of : atom, radius, charge in A5 type of format. The atom information is preceded by a header terminated by a star.
- **vdw.dat** Extended radii data file - containing the information which is used to evaluate the Lennert-Jones potential. The squared values of R_o are kept in this file. The format is an atom pair followed by R_o and E_o for this atom pair. The format is free.
- **pldat.dat** Plotting data file containing information about how to plot a specific structure. The two first parameters in the file are a radius scaling factor, and a bond scaling factor. The next line contains atom radii and atom types. After this follows the RGB colours for these atoms and finally the colour of the bonds.

- **psdat.dat** PostScript data file containing information used to position the figure on the paper, resolution scale etc..

- **mdm78.dat** Dayhoff mutation data matrix used for the distance searching of protein surfaces.

B.2 Documentation for protein interaction investigation program

B.2.1 Introduction

The basic idea of the program is to create a platform for different surface and protein analysis programs and algorithms described in the literature.

The program runs with a scrolling menu, and should be very portable and it should be easy to append new routines to the main, but this statement will always remain subjective to the programmer who wrote the code.

The program uses an index list which handles any generic pointer list, when used the pointer is CAST to the right type. Any new object which is read in or generated will be added to the main list. The main list can have any length. The working set list is an array of NSET generic pointers. The program will only allow the user to have one list of each type in the list at any time. This can however be overridden when using the handling routines in the selection and display menu. There will also at any time be two lists of the type VECTOR, one which handles surface points, and one which holds the unit icosahedron. Se also section on the Selection menu.

B.2.2 Hole filling

The program implements an algorithm for filling of holes on a molecular surface, described by Kuntz et al (Kuntz *et al.*, 1982)

The algorithm uses surface normals generated by a surfacing program, in this case a normal hard sphere surface generated using a tessellated icosahedron with user specified tessellation frequency (default 4). This method of generating surfaces was first used by C.Sander et.al. in the DSSP program (Kabsch and Sander, 1983), who used the method for determination of sidechain accessibilities.

The rules currently implemented are:

1. Only pairs of surface points are considered for which the dot product of normal i and the vector ij is larger than or equal to zero.
2. Only spheres obeying Rmin and Rmax criteria are included.
3. For a given point only the smallest sphere generated is kept.
4. For a given atom only the largest of the above spheres is kept.
5. Only spheres touching residues farther apart than "sepres" residues apart will be considered.

The Cutoff's are usually set to Rmin = probe radius, and Rmax = 5.0 Angstrøm.

B.2.3 Clustering

The program uses a method of clustering described by Oriochi (Lazlo, 1975). The idea behind the clustering is to connect datapoints which are close in the N-dimensional space, using the square of the euclidian distance as the expression for "closeness" of conformations. It should be pointed out that the euclidian distance does not have any physical meaning in this case it is purely an expression which relates N x M parameters. The implication of this is that if the distance between conformation A and B is equal to the distance between A and C then it is not possible to say that B and C are equally close in "conformation".The difference between A and C might arise from the difference in one torsion angle, whereas the difference between A and B may come from the small difference between more torsion angles.

The automated procedure (default value) assumes that the clusters are well resolved !! - so be careful when using this way of clustering. Initially datapoints are excluded when they are more the 3.5 SD units away from the mean shortest distance. The mean distance between the closest and next closest datapoint is used as the step size. The initial distance is set to the shortest distance within the dataset, and the clustering is stopped when the distance reaches $\rho + 3.5\cdot$ SD.

After version 2 the automatic clustering should be all right I have spend some time optimising the code and parameters. The program no longer writes a file of clusters in the Jancy classification routine. This routine has been optimised. Its running quite a bit faster.

Use the display of the cluster-matrix to determine how well resolved the clusters are. It will also give some impression of the deviation of the datasets.

B.2.4 Surface generation

The surface generation uses a tessellated icosahedron as an approximation to the sphere, this makes the program very fast.

This is an approximated solution to the the "golf ball" problem. How do you distribute N points on a sphere , such that the position of each point represents an equal surface area ?.

In this program we assume that the area represented by each vertice in the tessellated icosahedron is equal. This is however not true - but the approximation is close. See Kabsch and Sander (1983) for description of algorithm and source code to DSSP. The algorithm can actually be improved quite a lot by using a recursive generation of the surface points, this way the two edges spanning a triangle will be known as vectors, the area of a given triangle will then be half the length of the cross product vector. This way the area of any individual triangle becomes known. The precision of this method should be around 0.995.

If you want to read more about tessellated icosahedra and other of the beautiful polyhedra - I recommend that you read Pugh (Anthony, 1976), and Chau *et al* (Chau and Dean, 1987)

There is the freedom to choose the frequency with which you want to fragment the faces of the icosahedron. The number of faces, vertices and edges is given by:

$N_{fac} = 20 \cdot \gamma^2$

$N_{ver} = 2 + 10 \cdot \gamma^2$

$N_{edge} = 3 \cdot \gamma^2$

B.2.5 Triangulation

The triangulation used is a tricky one, and unfortunately a bit slow. The problem is that the points generated above, on the surface are not equally distributed on the surface. This means that a normal nurb surface or nearest neighbor triangulation would fail, by generating holes in the surface.

The scheme i use is to generate a list of MEANNAYB nearest neighbors, of a slightly overlapping surface. A surface with a tolerance in the VdW radius when generating the list of exposed surface points.

For each vertice in the surface the vectors to the nearest neighbors are calculated and sorted according to relative angle between the i'th nearest neighbor. This means that a sorted list is made for each of the surface points. Then the triangle to the nearest neighbor of a given edge is generated. The problem with this is that a given triangle will be generated a maximum of three times. This can however be remedied by a cleanup routine which checks all the triangles. This check routine has been left out at the moment since it is quite slow.

B.2.6 The graphics interface

In the current version it is possible to display the hydrophobic or electrostatic potential on a triangulated surface, and to display a set of spheres on it. This is currently being developed such that it will be possible to display a given cluster.

You can also display generated spheres, if there are any spheres in the list you will be prompted whether you want to display or not.

B.2.7 The protein-protein interface programs

The idea to these programs comes from the paper by Vellarkad Viswanadhan (Viswanadhan, 1987), investigating crystal structure packing in a long range of proteins. And from the fact that there seems to be a connection between the excluded surface area of an interface and the binding constant. The calculation of the excluded surface is done as a simple subtraction :

$$(A_{ifree} + A_{jfree}) - (A_{ibound} + A_{jbound}) = A_{excl}$$

A probe radius of approx 0.1 Angstrøm is appropriate.

Excluded volume is calculated in the same fashion, using the volumes in stead of areas, thus :

$$(V_{ifree} + V_{jfree}) - (V_{ibound} + V_{jbound}) = V_{excl}$$

The volume calculation is grid search which does an integration over the grid in order to determine grid cubes which are included in the molecule, and which are not. The precision of this method is naturally dependent on the grid size.

B.2.8 Selection,display and list handling

This menu is the control menu for the program.

The colour map assignment is only used when displaying things with INSIGHT (b) - it assigns colours to dots.

Sphere selection is an option to select only spheres which are generated from a specified range of atoms, or to select spheres which are within a certain distance from a specified atom, specified by its atomnumber.

Debug level is described below.

Allocation information display is a dump of the mallinf structure supported by *MIPS* and can be excluded if the program is ported to other machines. Note that the total space in the Arena always corresponds to the maximum number of blocks used - even if old lists have been freed up - this shows how useless unix is.

List handling is a dangerous - but a very useful option. As described above all information is kept as lists of a certain type (an object). You can free the space occupied by any list. You always have a working set of pointers. This list will always attempt to have only one list of each type in it at the time in order to avoid confusion. But remember ! - None of the objects are linked, thus a given PDB list does not know which VECTOR list it belongs to. Unfortunately I have not been able to find a good way to add the Brookhaven file name to the descriptor of lists derived from this (eg. neighbour lists etc.). This type of object handling will probably be confusing in the beginning. However if you are confused - only handle one set of molecules at the time.

When a Brookhaven file containing several different chains, is read into the program each of the chains will be added to the index list as individual objects. The list in the working set will always be the whole structure. If you want to work on any of the subchains as individual objects use this menu entry to change it.

If you want to use the Protein investigation module you should add all the structures you want to look at to the working set, although this will give a warning. E.g. if you want to calculate the excluded surface area or excluded volume of a protein-protein complex you should have each of the unbound molecules as separate objects and the complex in the working set list.

The final option of this menu is an option to keep track of the time spend in different parts of the program. The timer keeps two values, the first is the time used since the program was initialised and the second holds the time spend doing the last command.

B.2.9 Installation

There comes two makefiles with the program one which generates an optimised version, and one which does not.

To install normal version type:

```
make -f Intnormal.make
```

To install optimised version type:

```
make -f Intopt.make
```

To run the program type : Int

B.2.10 Datafiles

The datafiles used are :

- VDW.dat File of VDW radii for elements , There is an example of this file in the directory. [JP90VDWRADII]
- data.dat Data file containing information for the generation of surface and spheres.

  The number of data parameter required changes from version to version, the list below is updated after each version, and contains the right number of parameters for current version.

  - *Rmax* maximum radius of spheres generated 5.0
  - *Rmin* minimum radius of spheres generated 1.4 - 1.8
  - *Rprobe* probe radius 1.4 - 1.8
  - *Tf* Tesselation frequency 4
  - *cutoff* cutoff for generation of icosahedrons 0.1
  - *sepres* residue separation 3 - 4
  - *col1col2col3* charge colours for insight display 20 100 180
  - *Rcutoff* Distance cutoff for hydrophobic potential calculations. 10.0
  - *SCALE* Scaling factor for Hydrophobic potential 100
  - *MEANNAYB* Mean number of neighbors from which number of neighbors should be calculated in triangulation routine. 7 - 10
  - *Rtolerance* Radius tolerance for surface generation, This is zero for calculations, but is set to 0.05-0.1 Angstrøm when generating triangulated surface.

B.2.10 Datafiles

The datafiles used are :

- VDW.dat File of VDW radii for elements , There is an example of this file in the directory. [JP90VDWRADII]
- data.dat Data file containing information for the generation of surface and spheres.

  The number of data parameter required changes from version to version, the list below is updated after each version, and contains the right number of parameters for current version.

  - *Rmax* maximum radius of spheres generated 5.0
  - *Rmin* minimum radius of spheres generated 1.4 - 1.8
  - *Rprobe* probe radius 1.4 - 1.8
  - *Tf* Tesselation frequency 4
  - *cutoff* cutoff for generation of icosahedrons 0.1
  - *sepres* residue separation 3 - 4
  - *col1col2col3* charge colours for insight display 20 100 180
  - *Rcutoff* Distance cutoff for hydrophobic potential calculations. 10.0
  - *SCALE* Scaling factor for Hydrophobic potential 100
  - *MEANNAYB* Mean number of neighbors from which number of neighbors should be calculated in triangulation routine. 7 - 10
  - *Rtolerance* Radius tolerance for surface generation, This is zero for calculations, but is set to 0.05-0.1 Angstrøm when generating triangulated surface.

The default file is data.dat,and an example is present in the directory. The file contains max 30 elements.

- cvffa.rlb VFF residue library. This is used for the assignment of charges to atoms and charges to spheres. Spheres are assigned opposite charges an colors. Currently it is not possible to assign other colors than one for each of the three charge possibilities: col1 = positive charge col2 = zero charge col3 = negative charge This will hopefully be changed in the future, such that gradient colouring becomes possible.
- hph.rlb Atomic hydrophobicities library values are calculated after (Fauchere *et al.*, 1988)

  hydrogen is set to zero. The format is the same as cvffa.rlb.

B.2.11 Io

This section describes the coordinate files generated and read by the program, except for datafiles which are described above.

- Brookhaven file Standard PDB file. Does a file contain more than one chain ID then these will be treated as individual structures.
- surface file File containing the dots on the generated surface in simple XYZ ASCII format 3F10.
- surface normal file File containing the same information as the above file, but additionally contains the surface normal to each point format 6F10.
- sphere file File of generated spheres containing XYZ coordinates to sphere center, radius, and the number of the atom to which the sphere is associated; format 4F10,I5.
- insight surface normal file File which contains all the surface normals used in the generation of hole spheres. The format is as a Biosym user (LINE) file. The normals can be displayed using the command: "get user <insight normal file> as <name> using <mol-object>" This will display the vectors with respect to the molecular object. (Remember to associate !)

- insight sphere file File of dotted spheres for Insight this has the Biosym user (DOT) format. Display this in the same manner as the above type of file.
- insight surface file File of surface dots. The format is the same as the sphere file for insight.
- triangle file File containing coordinates for triangulated surface as three sets of X Y Z coordinates and a fourth parameter - eg hydrophobic potential.

The program contains both routines for reading and writing. These have not yet been tested properly.

B.2.12 Examples

Install the program and type Int.

Example 1 A typical run of the program would look something like This :

generate a set of spheres.

1. Read the VDW datafile
2. Read your pdb file
3. Read your data file
4. Set the debug level- optional
5. Generate the surface
6. Generate the spheres
7. write sphere file or any other file you might fancy.

Example 2 Select a set of spheres, and colour according to charge.

1. Read the sphere file.
2. Read the data file.
3. "Generate surface" - this will setup the icosahedron such that all the spheres can be regenerated.
4. Read pdb coordinate file.
5. Read charge file (Discover cvffa.rlb)
6. Assign charges to atoms
7. Assign charges to spheres
8. select spheres
9. write new sphere file

The normal values for data are outlined in the file data.dat.

Example 3 How to display water channel in subtilisin (1cse).

The trick is to allow for rather small spheres to be generated, which however makes the run rather longer to run. The other trick is to use something close to the VdW surface for the generation of the spheres, in order to make cavities more visible.

use a datafile which look like this:

```
5.00          <R max>
0.50          <R min>
```

```
0.50          <Probe radius>
4             <Tesselation>
0.05          <Cutoff>
3             <Residue separation>
0 120 240     <Colours for spheres>
10.0          <Potential calculation cutoff distance>
100           <HPH Potential scaling factor>
10.0          <Mean number of neighbors in triangulation>
0.1           <VdW radius tolerance>
```

1. Generate the surface and the spheres.
2. Select all the spheres which are within a radius of 10 Angstrøm of atom 234 (OD2,ASP 32). This removes all the redundant spheres on the surface.
3. Display the spheres in insight.

Note the nice line of spheres from water 410 and out of the channel.

B.2.13 Debug and other useful hints

A debug option is incorporated in the program If you wish to speed up the program you can take out all the debug stuff from the code.

The level can be set between 0 and 5. 0 gives no information and 5 fills up your disk i no time. The debug level is approx an indication of routine level from main routine, although this is not always true.

The timer option is quite useful when running a lot of stuff, the unix time command is used, thus the time is system time.

B.3 Program for cluster analysis of loop conformations

B.3.1 Introduction

The cluster program is a small menu driven tool for the analysis of a set of loop conformations.

The current version can read a CONGEN .cga file and a cluster datafile. The cluster datafile is a free format type of file which contains a N x M matrix. N is the number of variables to be used in the clustering, and M is the number of datasets.

The program is now also able to handle any type of data for clustering - just use the cluster data type of file for your data.

B.3.2 Clustering

Since the number of variables that have to be handled usually is very large, as is the number of datasets, the normal clustering routines can not be used.

The program uses a method of clustering described by L. Orioci (Lazlo, 1975). The idea behind the clustering is to connect datapoints which are close in the N-dimensional space, using the square of the Euclidian distance as the expression for "closeness" of conformations. It should be pointed out that the Euclidian distance does not have any physical meaning in this case it is purely an expression which relates N x M parameters. The implication of this is that if the distance between conformation A and B is equal to the distance between A and C then it is not possible to say that B and C are equally close in "conformation".The difference between A and C might arise from the difference in one torsion angle, whereas the difference between A and B may come from the small difference between more torsion angles.

The automated procedure (default value) assumes that the clusters are well resolved !! - so be careful when using this way of clustering. Initially datapoints are excluded when they are more the 3.5 SD units away from the mean shortest distance. The mean distance between the closest and next closest datapoint is used as the step size. The initial distance is set to the shortest distance within the dataset, and the clustering is stopped when the distance reaches mean $\pm 3.5 \cdot SD$.

After version 2 the automatic clustering should be all right I have spend some time optimising the code. The program no longer writes a file of clusters in the Jancy classification routine. This routine has been optimised. Its running quite a bit faster.

Use the display of the cluster-matrix to determine how well resolved the clusters are. It will also give some impression of the deviation of the datasets.

In version 2.2 and onwards its possible to cluster a set of brookhaven files (e.g. loops), You have to have an ffile and all the brookhaven files have to be ordered according to order of atoms in the datafile [jan/progs/framebuild/data/order.dat].

B.3.3 How to use the program

Initial clustering of database loops:

1. Read in all the conformations (pdb files).
2. Determine the bounds of the dataset.
3. Do the clustering.

Processing of conformational ensemble (CONGEN conformations)

1. Read in you data - either as CONGEN conformation file or as cluster datafile.
2. Read in your Gromscan energy file - this is the free format X Y file containing a column with conformation number and a column with energies. (eg the output file from tabc)
3. Determine the bounds of the dataset. Read in the Gromscan energies.
4. Use the automatic clustering to start with - I have done a couple of tests on the routine now and it seems to do the clustering in a satisfactory way. The automated routine is also quicker. If you want better discrimination of you data use the old protocol. Use a step size of 0.2 and 100 to 200 iterations, this value is dependent of the size of the dataset and deviation within the dataset. Large datasets with long loops from many different parent loops require fewer cycles of clustering since the clusters are better resolved than large datasets with short loops. You will have to experiment a little bit here - there is no clear answer to this problem at the moment.
5. Write out the newly sorted conformations and use these for filtering.

B.3.4 Documentation for antibody framework building program. Version 3

Introduction

The purpose of this package is to provide an easy building tool for the antibody modelling programs developed by A.Martin. The objective is to fully automate the building of antibody CRD's.

The *frambuild* program consists of approx 30 subroutines contained in the C library *framework.a.* The routines can be used in any program by including the library in the top of your source file. The subroutines are commented and an explanation of how to use individual routines is given in the beginning of each of the routines.

The *framebuild* program will build a suitable framework for modelling of antibody combining sites by choosing frameworks from a database of X-ray crystallographic structures of antibody fragments(F_{AB}'s and complexes, and dimers). The program chooses the framework structures from a sequence homology score. The program then compares the sequence of the database structure with the sequence of the required structure. The sidechains of the database structures are then replaced using a maximum overlap approach. The sidechains are replaced by standard conformations, adjusting equivalent chi-angles in the new sidechain to the same as the chi-angles of the database structure sidechain.

The package also provides programs to setup the database of structures used for the building.

V.2.0 This version contains the subunit placer. All the beta-barrels of 8 antibodies have been fitted onto each other by a multiple fit. First fitting all the structures to the barrel of Gloop2. A mean set of coordinates have been derived and all the structures are then fitted onto this one. This is repeated, until the sum RMS converges. This fit is the considered being the best overall fit. The orientation of the barrel is such that the conjugate axis of the best hyperboloide is the X-axis and the focus of the hyperboloide is (0.0.0).

Quick guide to use the framework builder

The building of the framework is done in 4 steps:

Type *package*

1. **Reading sequence:**

   This small sequence editor will give the sequences of equivalent sequences in the framework database. This list of sequences has to be updated if new framework structures are added to the database.

   type: *readseq* , and type in your sequence, additional information is given when you run the program.

   Remember to match the sequence alignment for each fragment, using "-" as deletion and "$\star$" as end of fragment.

2. **Choosing frameworks:**

   This program calculates the sequences alignment score between each of the framework structures and the sequence you have typed in.

   type: *chooser* , and give the base name of your sequence files.

3. **Building framework:**

This program reads the scoring files generated by CHOOSER and builds the framework structure,by replacing sidechains,using a maximum overlap approach.

The program can either be run interactively or be submitted as a background job.

type : *framebuild* <file.inp> <file.out>

If no file names are stated on the command line the program runs interactive. If only input file is given on the command line the input is read from input file. If both files are stated the output is written to <file.out>.

The input file looks like this:

- JP90COOR ! Coordinate library
- JP90RES ! Residue nomenclature library
- JP90CHILINK ! Link table of chi angles
- UDB: ! Framework library
- <base name of sequence files>
- <CHOOSER output file for L-chain>
- <CHOOSER output file for H-chain>
- <prefix for output files>
- JP90FVFITL ! Fitting coordinates l-chain
- JP90FVFITH ! Fitting coordinates h-chain

When you run interactive -just hit return for the first four files, since they have been set up as symbolic links.

4. **Fitting of L and H Chains.**

The last step is to fit the L and H chains onto a suitable framework. This is done using the LSQ option of a graphics system such as FRODO or HYDRA.

V.3.1.0 : This is now done automatically , using conserved positions in the framework regions for the fitting h and l subunit. But watch out for exceptions - some F_V's are not very well represented by this dataset. So run the CLASH program to check for clashes between l and h chain, and have a look on a picture system.

The framework database

The framework database can be found in the *udb* directory. The directory contains the framework structures in PDB format. The structures have been separated in L and H chains. To add a new structure to the database you have to read in the sequence of the new structure,using *readseq*. The .PDB file then has to be truncated to match the sequence which you have typed in. The next step is to run the program *prepare*,which matches the sequence to the .PDB file and adds DEL entries in the .PDB file where these occur in the sequence. The last step is to separate the .PDB file in two .UDB file - one which contains the L chain and one which contains the H chain.

Remember to add the sequence and the name of the new framework structure to the sequence files:

- FRAMEL.DAT
- FRAMEH.DAT
- DB_V⋆.DAT < 14 files >

Probing a new structure

When a new structure is added to the database it is necessary to check the structure for any possible misplacements. This is done in two possible ways:

1. Plotting the B-values of the backbone together with other structures. This gives a qualitative impression of how good the new structure is compared to other structures in the database. The program PDB2CURVY in *Tools* converts a .PDB file into a free format file of two columns containing the residue number and the B-value of CA of the given residue. This free format file can be plotted with suitable graph plotting programs JPLOT (Pedersen, 1992),MATLAB (TM Stardent).
2. Comparing the B-values of the structure, flagging any residue which has a B-value higher than mean $\pm 3 \cdot SD$ units.
3. If no B-values are present the framework has to be compared with other framework structures by least squares fitting, and flagging any non-CDR residue which has a deviation of $\pm 3 \cdot SD$ units.

Comparison of mutation procedures

The different versions of the *framebuild* program have been tested and compared to other existing methods available. These methods are:

- **HYDRA** mutate option in build menu. Uses maximum overlap replacement.
- **FRODO** Replace option, followed by Refi. This method uses a maximum overlap approach, married to a method which optimizes the position of the

remaining atoms to be placed, such that bond angles and length are optimized. The method also checks for VdW clashes.

- **COMPOSER** M.Sutcliffe's (Sutcliffe *et al.*, 1987a) sidechain replacement program which uses a database of homologous structures for the placing of sidechains.

- **BUILD** V1.0 replaces residues with standard conformation. V1.1 replaces sidechains with standard conformation, but retains backbone. This method also includes equivalent chi-angles. V1.2 same as V1.1 but does true overlapping, using side chain for fitting if equivalent chi-angles are present.

The framework of Gloop2 light chain was build using the l chain from the database structure with the highest sequence homology (1REI 61 %) and from the database structure with lowest sequence homology (1FB4 41 %). The results are comprised in table B.1.

The upper value in each box is the mean RMS deviation and the lower value is the Max deviation.

It is evident that the crude replacement of residue, simply taking standard conformations never is going to give satisfactory results, no matter how large the statistical material is which has been used for establishing the standard conformation.

Build V1.1 simply gets the backbone right, but there must still be some misplaced sidechains, as the max deviation is larger then the difference between the backbones alone(8.59 for 1REI and GLOOP2).

Method	Compared to Gloop-2		Compared to 1FB4
	1FB4	1REI	1FB4
BUILD V1.0	6.23	6.12	1.53
	17.11	10.23	5.23
BUILD V1.1	5.00	5.23	0.90
	14.29	8.93	2.49
BUILD V1.2	4.56	4.26	0.33
	11.10	8.59	1.65
FRODO	3.58	4.06	0.28
	10.73	8.59	1.58
HYDRA	5.20	4.53	0.45
	10.91	8.59	1.78
COMPOSER	3.20	3.34	0.41
	10.84	8.59	1.20

Table B.1: Comparison of several traditional Molecular Modelling Packages sidechain replacement methods. RMS Values are backbone and all atoms respectively for the complete construction of an F_V fragment.

Build V2.0 gives a proper overlap and matches very well the values obtained by FRODO and HYDRA. The method is slightly better than HYDRA. The method could probably be improved by driving the sidechain about the terminal chi-angel and testing for clashes, and minimising nonbonded energies.

The COMPOSER method is the superior method, specially when it comes to placing sidechains which have no equivalent chi-angles.

The placement of subunits - with regard to each other(Pairing)

To determine the best pairing, all the beta barrels of 8 antibody structures were fitted by a multiple fit. By this fit all the structures are fitted iteratively, deriving a mean framework for each iteration, which is used for the fitting in the next cycle.

The regions fitted are the conserved regions determined by Chothia *et al*(Chothia *et al.*, 1985). The best fitting hyperboloide is derived by the method described by Chothia (Chothia *et al.*, 1985). The mean deviation (MD) was then plotted from the bottom of the barrel and up. The residues in each strand of the barrel which are closest to the antibody combining site (AC) are denoted as residue 1 an so on. As an expression of the *disorder* the sum of the squared inter atom distances, for each atomic position in the multiple fit, is plotted. The strands 2,7 and 8 seem to be significantly more disordered than the remaining strands. The most disordered strand (MD = 3 A) is strand 7. This strand is excluded in the framework fitting. Strand 2 and 8 have been kept. A more thorough analysis is required to determine whether exclusion of these strands is justified. The same plots have been made - but by plotting the MD and *disorder* as function of distance between the projection onto the conjugate axis and the focus.

Loop numbering

In order to facilitate the construction of F_V domains a standard numbering of loop regions has been adopted, table B.2

There are obvious disadvantages by using this method. The numbering has to be changed if new antibody crystal structures contain longer CDRs than any other known structure. This disadvantage can be remedied by allowing for huge(50 AA's) in the middle of the CDRs.

Note that the latest change is the insertion of two DEL entries in CDR H1 to accommodate for 7 residue loops.

CDR residues	First residue	Last residue
L1	24	40
L2	56	62
L3	95	105
H1	148	154
H2	170	180
H3	220	236

Framework residues	First residue	Last residue
LFR1	41	64
LFR2	51	55
LFR3	91	92
LFR4	106	110
HFR1	155	158
HFR2	164	169
HFR3	215	219
HFR4	237	238

Table B.2: Standard CDR loop numbering, and numbering of framework β-strands (UDB numbering)

References

DISCOVER and INSIGHT are trademarks of Biosym Technologies, San Diego, California, USA.

Abbas, A., Lichtman, A. and Pober, J., (1991). *Cellular and Molecular Immunology.* W.B. Saunders company, Hartcourt Brace Jovanovich, Inc.

Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. and Poljak, R. J., (1986). The three-dimensional structure of an antigen-antibody complex at 2.8Å resolution. *Science*, **233**,747–753.

Amzel, L. and Poljak, R., (1979). Three-dimensional structure of immunoglobulins. *Annu. Rev. Biochem.*, **48**,961–997.

Anthony, P., (1976). *Polyhedra ,a visual approach.* University of California press, first edition.

Åqvist, J., Gunsteren, V. W. F., Leijonmarck, M. and Tapia, O., (1985). A molecular-dynamics study of the c-terminal fragment of the l7/l12 ribosomal-protein - secondary structure mo tion in a 150 picosecond trajectory. *J. Mol. Biol.*, **183**.3,461–477.

Arnold, N., Wienberg, J., Ermert, K. and Zachau, H. G., (1991). Evolution of v-kappa immunoglobulin genes in human and primates analyzed by molecular cytogenetics. *Am. J. Hum. Gen.*, **49**.4,332.

Atassi, M., (1975). Antigenic structure of myoglobin: The complete immunochemical anatomy and conclusions relating to antigenic structures of proteins. *Immunochemistry*, **12**,423–438.

Atassi, M., (1978). Precise determination of the entire antigenic structure of lysozyme. *Immunochemistry*, **15**,909–936.

Aubry, A., Birlirakis, N., Sakarellos-Daitsiotis, M., Sakarellos, C. and Marraud, M., (1988). Relationship of the crystal and molecular structure of leucine-enkephalin trihydrate to that of morphine. *J. Am. Chem. Soc.*, C,963–964.

Bagshawe, K. D., (1987). Antibody directed enzymes revive anti-cancer prodrugs concept. *Bri. J. Cancer*, **56**.5,531–532.

Bank, R. A., Russell, R. B., Tenkate, R. W. and James, M. N. G., (1990). Comparative molecular modeling of human pepsinogens - an attempt to explain its high sieving through the glomerular-basement-membrane. *Kidney international*, **38**.2,360.

Barton, G. J., (1990). Protein multiple sequence alignment and flexible pattern-matching. *Meth. Enz.*, **183**.403,428.

Barton, G. J. and Sternberg, M. J. E., (1987). A strategy for the rapid multiple alignment of protein sequences - confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, **198**.2,327–337.

Barton, G. J. and Sternberg, M. J. E., (1990). Flexible protein-sequence patterns - a sensitive method to detect weak structural similarities. *J. Mol. Biol.*, **212**.2,389–402.

Baum, R., (1991). Catalytic antibody functions invivo. *Chemical & engineering news*, 69.42,27–28.

Benjamin, D., Berzofsky, J., East, I., Gurd, F., Hannum, C., Leach, S., Margloash, E., Michael, J., Miller, A., Prager, E., Reichlin, M., Sercarz, E., Smith-Gill, S., Todd, P. and Wilson, A., (1984). The antigenic structure of proteins - a reappraisal. *Annu. Rev. Immunol.*, **2**,67–101.

Benkovic, S. J., Adams, J., Janda, K. D. and Lerner, R. A., (1991). A catalytic antibody uses a multistep kinetic sequence. *Ciba foundation symposia*, **159**.4,12.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. and Tasumi, M., (1977). The protein databank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, **112**,535–542.

Bird, J., Galili, N., Link, M., Stites, D. and Sklar, J., (1988a). Continuing rearrangement but absence of somatic hypermutation in immunoglobulin genes of human b cell precursor leukemia. *J. Exp. Med.*, **168**,229–245.

Bird, R. E., Hardman, K. D., Jacobson, J. W., Johnson, S., Kaufman, B. M., Lee, S. M., Lee, T., Pope, S. H., Riordan, G. S. and Whitlow, M., (1988b). Single-chain antigen-binding proteins. *Science*, **242**.4877,423–426.

Bleasby, A., (October 1990). Seqnet user guide. version 2.0. UIG (User Group), Daresbury Laboratories.

Bleasby, A. and Wouton, J., (1990). Construction of validated, nonredundant composite protein-sequence database. *Protein Eng.*, **3**.3,153–159.

Blundell, T. L. and Sternberg, M. J. E., (1985). Computer-aided design in protein engineering. *Trends Biotechnol.*, **3**,228–235.

Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M., (1983). Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, **4**,187–217.

Browne, W., North, A., Phillips, D., Brew, K., Vanaman, T. and Hill, R., (1969). A possible three dimensional structure of bovine α-lactalbumin based on that of hen's egg lysozyme. *J. Mol. Biol.*, **42**,65–68.

Bruccoleri, R. E. and Karplus, M., (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, **26**,137–168.

Bruck, C., Co, M. S., Slaoui, M., Gaulton, G. N., Smith, T., Fields, B. N., Mullins, J. I. and Greene, M. I., (1986). Nucleic-acid sequence of an internal image-bearing monoclonal anti-idiotype and its comparison to the sequence of the external antigen. *Proc. Natl. Acad. Sci. USA*, **83**.17,6578–6582.

Brunger, A., Leahy, D., Hynes, T. and Fox, R., (1991). 2.9 ångstrøms resolution structure of an anti-dinitrophenyl spin label monoclonal antibody F_{AB} fragment with bound hapten. *J. Mol. Biol.*, **221**,239.

Bye, E., (1976). The crystal structure of morphine hydrate. *Acta Chemica Scandinavica B*, **30**,549–554.

Byrn, R. A., Mordenti, J., Lucas, C., Smith, D., Marsters, S. A., Johnson, J. S., Cossum, P., Chamow, S. M., Wurm, F. M., Gregory, T., Groopman, J. E. and Capon, D. J., (1990). Biological properties of a CD4 immunoadhesin. *Nature (London)*, **344**.6267,667–670.

Casy, A. and Robert, T., (1986). *Opiod analgesics : Chemistry and receptors.* Plenum Press, New York, London.

Chau, P. and Dean, P., (1987). Molecular recognition: 3D surface structure comparison by gnomonic projection. *J. Mol. Graph.*, **5**,97–100.

Chothia, C., Lesk, A., Levitt, M., Amit, A., Mariuzza, R., Phillips, S. and Poljak, R., (1986). The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science*, **233**,755–758.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D. R., Tulip, W. R., Colman, P. M., Alzri, P. M. and Poljak, R. J., (1989). Conformations of immunoglobulin hypervariable regions. *Nature (London)*, **342**,877–883.

Chothia, C., Novotný, J., Bruccoleri, R. E. and Karplus, M., (1985). Domain association in immunoglobulin molecules—the packing of variable domains. *J. Mol. Biol.*, **186**,651–663.

Clackson, T., Hoogenboom, H. R., Griffiths, A. D. and Winter, G., (1991). Making antibody fragments using phage display libraries. *Nature*, **352**.6336,624–628.

Colman, P. M., (1988). Structure of antibody antigen complexes - implications for immune recognition. *Adv. Immunol.*, **43**.99,132.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. and Delisi, C., (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, **195**.3,659–685.

Covell, D. G., (1992). Folding protein alpha-carbon chains into compact forms by monte-carlo methods. *Proteins: Struct., Funct., Genet.*, **14**.3,409–420.

Crippen, G., (1981). *Distange Geometry and Conformational Calculations.* Research Studies Press, John Wiley & Sons Ltd. NY.

Crowe, J. S., Hall, V. S., Smith, M. A., Cooper, H. J. and Tite, J. P., (1992). Humanized monoclonal-antibody campath-1H - myeloma cell expression of genomic constructs nucleotide-sequence of c-DNA constructs and comparison of effector mechanisms of myeloma and chinese-hamster ovary cell-derived material. *Clin. and Exp. Immunol.*, **87**.1,105–110.

Dalgleish, A. G., Habeshaw, J., Manca, F., Jameson, B. and Hounsell, E., (1992). Modeling of gp120 reveals an alpha-helix structure with mhc class-ii homology containing a known alloepitope - mechanism of graft versus host immune-response in hiv-infection. *Aids research and human retroviruses*, **8**.5,947.

Darsley, M. J., Phillips, D. C., Rees, A. R., Sutton, B. J. and de al Paz, P., (1985). *An approach to the study of anti-protein antibody combining sites. In investigation and exploitation of antibody combining sites*, chapter #A-4, pages 63–68. Plenum Press, first edition.

Darsley, M. and Rees, A., (1985). 3 distinct epitopes within the loop region of hen egg lysozyme defined with monoclonal-antibodies. *EMBO J.*, **4**,383–392.

Dauber-Osguthorpe, P., Campbell, M. and Osguthorpe, D., (1991). Conformational analysis of peptide surrogates. *Int. J. Peptide Protein Res.*, **38**,357–377.

Dayhoff, M., Barker, W. and Hunt, L., (1983). Establishing homologies in protein sequences. *Meth. Enz.*, **91**,524–545.

de la Paz, P., Sutton, B., Darsly, M. and Rees, A., (1986). Modelling of the combining sites of three anti-lysozymemonoclonal antibodies and of the complex between one of the antibodiesand its epitope. *EMBO J.*, **5**,415–425.

Desjarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D. and Venkataraghavan, R., (1988). Using shape complementarity as an initial screen in designing ligandsfor a receptor-binding site of known 3-dimensional structure. *Journal of medicinal chemistry*, **31**.4,722–729.

Desjarlais, R., (1988). *Molecular shape complementarity: A method for finding new lead molecules.* D. Phil. Thesis, UCLA departement of pharmaceutical chemistry.

Elliott, G., (1992). Personal communications.

Ely, K., Herron, J., Harker, A. and Edmunson, A., (1989). Three-dimensional structure of a light chain dimer crystalised in water. conformational flexibility of a molecule in two crystal forms. *J. Mol. Biol.*, **210**,601.

Ely, K., Wood, M., Rajan, S., Hodsdon, J., Abola, E., Deutch, H. and Edmunson, A., (1985). Unexpected similarities in the crystal structures of the mcg light-chain dimer and its hybrid with the WEIR protein. *Mol. Immunol.*, **22**,93.

Engvall, E. and Pesce, A., (1978). Quantitative enzyme immunoassay. *Scand. J. Immunol.*, **8, suppl 7**.

Fauchere, J. L., Quarendon, P. and Kaetterer, L., (1988). Estimating and representing hydrophobicity potential. *J. Mol. Graph.*, **6**.4,203.

Fine, R., Wang, H., Shenkin, P., Yarmush, D. and Levinthal, C., (1986). Predicting antibody hypervariable loop conformations II: Minimisation and molecular dynamics studies of McPC603 from many randomly generated loop conformations. *Proteins: Struct., Funct., Genet.*, **1**,342–362.

Free, S. and Wilson, J., (1964). A mathematical contribution to structure-activity relationships. *J. Med. Chem.*, **7**,395–399.

Furey-Junior, W., Wang, B., Yoo, C. and Sax, M., (1983). Structure of a novel bence-jones protein (RHE) fragment at 1.6 angstroms resolution. *J. Mol. Biol.*, **167**,661.

Garel, T., Niel, J. C., Orland, H. and Velikson, B., (1991). A new monte-carlo method to study protein structures. *J.Chim. Phys. et de Phys-Chem Biol. (Canada)*, **88.1**,2473–2478.

Gibbs, R. A., Posner, B. A., Filpula, D. R., Dodd, S. W., Finkelman, M. A. J., Lee, T. K., Wroble, M., Whitlow, M. and Benkovic, S. J., (1991). Construction and characterization of a single-chain catalytic antibody. *Proc. Natl. Acad. Sci. USA*, **88.9**,4001–4004.

Go, N. and Sheraga, H., (1970). Ring closure and local conformational deformations of chain molecules. *Macromolecules*, **3**,178–187.

Goodsell, D. S. and Olson, A. J., (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., Genet.*, **8.3**,195–202.

Gorman, S., Clark, M., Rutledge, E., Cobbold, S. and Waldman, H., (1991). Reshaping a therapeutic CD4 antibody. *Proc. Natl. Acad. Sci. USA*, **88**,4181–4185.

Greer, J., (1990). Comparative modeling of proteins in the design of novel renin inhibitors. *Biophysical journal*, **57**.2.

Greer, J., (1991). Comparative modeling of homologous proteins. *Methods in enzymology*, **202**.239,252.

Gregory, D., Staunton, D., Martin, A., Cheetham, J., Pedersen, J. and Rees, A., (1990). Antibody-combining sites: Prediction and design. *Biochem. Soc. Trans. (London)*, **57**,147–155.

Griffin, J., Hercend, T., Beveridge, R. and Schlossman, S., (1983). Characterization of an antigen expressed by human natural killer cells. *J. Immunol.*, **130**.6,2947–2951.

Hajdu, J., Machin, P., Campbell, J., Greenhough, T., Clifton, I., Zurek, S., Gover, S., Johnson, L. and Elder, M., (1987). Millisecond x-ray diffreaction and the first electron density map from laue photographs of a protein crystal. *Nature (London)*, **329**,178–181.

Hale, G., Dyer, M. J. S., Clark, M. R. and Waldmann, H., (1991). Development and clinical-experience with humanized monoclonal-antibodies. *Developments in Biotherapy*, **1**.1,195–199.

Hansch, C., (1969). A quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.*, **2**,232–239.

Havel, T. and Snow, M., (1991). A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol.*, **217**,1–7.

He, X., Ruker, F., Casale, E. and Carter, D., (1992). Structure of a human monoclonal antibody F$_{AB}$ fragment against gp41 of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. USA*, **89**.15,7154–7158.

Herron, J., He, X., Mason, M., Voss, E. and Edmunson, A., (1989). Three-dimensional structure of a fluorescein-F_{AB} complex crystallised in 2-methyl-2.4-pentanediol. *Proteins: Struct., Funct., Genet.*, **5**,271–280.

Ikeda, S., Weinhouse, M. I., Janda, K. D., Lerner, R. A. and Danishefsky, S. J., (1991). Asymmetric induction via a catalytic antibody. *J. Am. Chem. Soc.*, **113**.20,7763–7764.

Jackson, D. Y., Prudent, J. R., Baldwin, E. P. and Schultz, P. G., (1991). A mutagenesis study of a catalytic antibody. *Proc. Natl. Acad. Sci. USA*, **88**.1,58–62.

Jacoby, S., Kowalik, J. and Pizzo, J., (1972). *Iterative Methods for Nonlinear Optimization Problems.* Englewood Cliffs, New Jersey : Prentice Hall.

James, H. L., Kumar, A., Girolami, A., Hubbard, J. G. and Fair, D. S., (1991). Variant coagulation factor-x and factor-vii with point mutations in a highly conserved motif in the substrate binding pocket - comparative molecular modeling. *Thrombosis and haemostasis*, **65**.6,937.

Jeffrey, P. D., Griest, R. E., Taylor, G. L. and Rees, A. R., (1991). Crystal structure of the F_{ab} fragment of the anti-peptide antibody Gloop2 and 2.8Å. *Manuscript in Preparation.*

Jerne, N., (1973). The immune system. *Sci. Am.*, **229**.1,52–60.

Jones, T. and Thirup, S., (1986). Using known structures in protein model building and crystallography. *EMBO J.*, **5**,819–822.

Kabat, E. and Wu, T., (1971). Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. *Ann. N.Y. Acad. Sci.*, **190**,382–393.

Kabat, E. A., Wu, T. T., Reid-Miller, M., Perry, H. M. and Gottesman, K. S., (1992). *Sequences of Proteins of Immunological Interest.* U.S. Department of Health and Human Services, Fifth edition.

Kabsch, W. and Sander, C., (1983). Dictionary of protein secondary structure. *Biopolymers*, **22**,2577–2637.

Kang, C. Y., Brunck, T. K., Kieberemmons, T., Blalock, J. E. and Kohler, H., (1988). Inhibition of self-binding antibodies (autobodies) by a vh-derived peptide. *Science*, **240**.4855,1034–1036.

Karle, I. L., Flippenanderson, J. L., Sukumar, M., Uma, K. and Balaram, P., (1991). Modular design of synthetic protein mimics - crystal-structure of 2 7-residue helical peptide segments linked by epsilon-aminocaproic acid. *J. Am. Chem. Soc.*, **113**.10,3952–3956.

Karle, I. L., Flippenanderson, J. L., Uma, K. and Balaram, P., (1990). Apolar peptide models for conformational heterogeneity, hydration, and packing of polypeptide helices - crystal-structure of heptapeptides and octapeptides containing alpha-aminoisobutyric-acid. *Proteins: Struct., Funct., Genet.*, **7**.1,62–73.

Kennard, O., (1991). The cambridge crystallographic databank. crystal structure data for about 90.000 organic and organo-metallic compounds. Cambridge Crystallographic Data Cetre, Released bi-annual in Jan and July.

Kettleborough, C., Saldanha, J., Heath, V., Morrison, C. and Bendig, M., (1991). Humanisation of mouse monoclonal antibody by cdr-grafting: the importance of framework residues on loop conformation. *Protein Eng.*, **4**.7,773–783.

Khalaf, A. I., Proctor, G. R., Suckling, C. J., Bence, L. H., Irvine, J. I. and Stimson, W. H., (1992). Remarkably efficient hydrolysis of a 4-nitrophenyl ester by a catalytic antibody raised to an ammonium hapten. *Journal of the chemical society-perkin transactions I*, **12**.1475,1481.

Klobeck, H., Bornkamp, G., Combriato, G., Mocikat, R., Pohelnz, H. and Zachau, H., (1985a). Subgroup IV of human immunoglobulin κ light chains is encoded by a single germline gene. *Nuc. Ac. Res.*, **3**,6515–6529.

Klobeck, H., Meindl, A., Combriato, G., Solomon, A. and Zachau, H., (1985b). Human immunoglobulin kappa light chain genes of subgroups II and III. *Nuc. Ac. Res.*, pages 6499–6513.

Kroemer, G., Helmberg, A., Bernot, A., Auffray, C. and Kofler, R., (1991). Evolutionary relationship between human and mouse immunoglobulin kappa light chain variable region genes. *Immunogenetics*, **33**,42–49.

Kuntz, I., Blaney, I., Oatley, S., Langridge, R. and Ferrin, T., (1982). A geometric approach to macromolecule-ligan interactions. *J. Mol. Biol.*, **161**,269–288.

Kussie, P., Anchin, J., Subramaniam, S., Glasel, J. and Linthicum, D., (1991). Analysis of the binding-site architecture of monoclonal antibodies to morphine by using competitive ligand-binding and molecular modeling. *J. Immunol.*, **146**.12,4248–4257.

Kyle, V., Roddy, J., Hale, G., Hazleman, B. L. and Waldmann, H., (1991). Humanized monoclonal-antibody treatment in rheumatoid-arthritis. *Journal of Rheumatology*, **18**.11,1737–1738.

Lascombe, M., Alzari, P., Boulot, G., Salujian, P., Tougard, P., Berek, C., Haba, S., Rosen, E., Nisonof, A. and Poljak, R., (1989). Three-dimensional structure of F_{ab} R19.9, a momoclonal murine antibody specific for the p-azobenzenearsonate group. *Proc. Natl. Acad. Sci. USA*, **86**,607.

Lazlo, O., (1975). *Multi Variate Analysis in Vegetation Research.* Dr.W.Junk Publishers, first edition.

Lee, C. and Levitt, M., (1991). Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature (London)*, **352**.6334,448–451.

Lee, C. and Subbiah, S., (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.*, **217**.2,373–388.

Legrand, S. and Merz, K., (1992). The application of genetic algorithm to conformational search. *Faseb journal*, **6**.1.

Lerner, R. A., Benkovic, S. J. and Schultz, P. G., (1991). At the crossroads of chemistry and immunology - catalytic antibodies. *Science*, **252**.5006,659–667.

Lewis, A. P. and Crowe, J. S., (1991). Immunoglobulin complementarity-determining region grafting by recombinant polymerase chain-reaction to generate humanized monoclonal-antibodies. *Gene*, **101**.2,297–302.

Lifson, S., Hagler, A. and Dauber, P., (1979). Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. carboxylic acids, amides, and the $C = O...H$-hydrogen bonds. *JACS*, **101**,55111.

Lobuglio, A. F. and Saleh, M. N., (1992). Monoclonal-antibody therapy of cancer. *Critical reviews in oncology/hematology*, **13**.3,271–282.

Maggiora, G., Mao, B., Chou, K. and Narasimhan, S., (1991). Theoretical and emperical approaches to protein-structure prediction and analysis. *Meth. Biochem. Anal.*, **35**,1–86.

Mainhart, C. R., Potter, M. and Feldmann, R. J., (1984). A refined model for the variable domains (F_V) of the J539 β (1,6)-d-galactan-binding immunoglobulin. *Mol. Immunol.*, **21**,469–478.

Marquart, M., Deisenhofer, J. and Huber, R., (1980). Crystallographic refinement and atomic models of the intact immunoglobulin molecule KOL and its antigen-binding fragment at 3.0Å and 1.9Å resolution. *J. Mol. Biol.*, **141**,369–391.

Martin, A. C. R., (1990). *Molecular Modelling of Antibody Combining Sites.* D. Phil. Thesis, University of Oxford.

Martin, A. C. R., Cheetham, J. C. and Rees, A. R., (1989). Modelling antibody hypervariable loops: A combined algorithm. *Proc. Natl. Acad. Sci. USA*, **86**,9268–9272.

Martin, A. C. R., Cheetham, J. C. and Rees, A. R., (1991a). Modelling antibody hypervariable loops using a 'combined algorithm'. *Meth. Enz.* In the press.

Martin, M. T., Napper, A. D., Schultz, P. G. and Rees, A. R., (1991b). Mechanistic studies of a tyrosine-dependent catalytic antibody. *Biochemistry*, **30**.40,9757–9761.

Martin, M. T., Schantz, A. R., Schultz, P. G. and Rees, A. R., (1991c). Characterization of the mechanism of action of a catalytic antibody. *Ciba foundation symposia*, **159**.188,200.

Mas, M. T., Smith, K. C., Yarmush, D. L., Aisaka, K. and Fine, R. M., (1992). Modeling the anti-cea antibody combining site by homology and conformational search. *Proteins-structure function and genetics*, **14**.4,483–498.

McCammon, A. and Harvey, S., (1987). *Dynamics of Proteins and Nucleic Acids.* Cambridge University Press, first edition.

Mcgregor, M. J., Islam, S. A. and Sternberg, M. J. E., (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular-proteins. *J. Mol. Biol.*, **198**.2,295–310.

McLachlan, A., (1979). Gene duplication in the structural evolution of chymotrypsin. *J. Mol. Biol.*, **128**,49–79.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E., (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**,1087–1091.

Mian, I. S., Bradwell, A. R. and Olson, A. J., (1991). Structure, function and properties of antibody-binding sites. *J. Mol. Biol.*, **217**.1,133–151.

Mosimann, S. C., Johns, K. L., Ardelt, W., Mikulski, S. M., Shogen, K. and James, M. N. G., (1992). Comparative molecular modeling and crystallization of p-30 protein - a novel antitumor protein of rana-pipiens oocytes and early embryos. *Proteins-structure function and genetics*, **14**.3,392–400.

Moult, J. and James, M. N. G., (1986). An algorithm which predicts the conformation of short lengths of chain in proteins. *J. Mol. Graph.*, **4**.3,180.

Moult, J., Yonath, A., Traub, W., Smilansky, A., Podjarny, A., Rabinovich, D. and Saya, A., (1976). The structure of triclinic lysozyme at 2.5 å resolution. *J. Mol. Biol.*, **100**,179–195.

Needleman, S. and Wunsch, C., (1970). A general method applicaple to the search for similarity in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**,443–453.

Newton, S. I., (1729 (1960)). Principia *(Mathematical Principles of Philosophy and his Sistem of the World)*. University of California Press, Berkley California, fourth printing edition. English translation by Andrew Motte.

Northrup, S. and Erickson, H., (1992). Kinetics of protein-protein association explaned by brownian dynamics computer simulation. *Proc. Natl. Acad. Sci. USA*, **89**,3338–3342.

Novotny, J., (1991). Protein antigenecity: A thermodynamic approach. *Mol. Immunol.*, **28**.3,201–207.

Novotny, J., Bruccoleri, R., Newell, J., Murphy, D., Haber, E. and Karplus, M., (1983). Molecular anatomy of the antibody-binding site. *J. Biol. Chem.*, **258**.23,14433–14437.

Novotny, J., Bruccoleri, R. and Newell, J., (1984). Twisted hyperboloid (*strophoid*) as a model of β-barrels in proteins. *J. Mol. Biol.*, **177**.3,567–573.

Novotny, J., Handschumacher, M., Haber, E., Bruccoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A. and Rose, G. D., (1986). Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl. Acad. Sci. USA*, **83**.2,226–230.

OML, (1992). The antibody modelling program **a***b***m**, (TM) Oxford Molecular Ltd., Oxford Science Park, Oxford, UK.

Padlan, E., (1990). On the nature of antibody combining sites: Unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins: Struct., Funct., Genet.*, **7**,112–124.

Padlan, E., (1991). A possible procedure for reducing the immunogenecity of antibody variable domains while preserving their ligand-binding properties. *Mol. Immunol.*, **28**,489–498.

Padlan, E., Davies, D., Pecht, I., Givol, D. and Wright, C., (1976). Model building studies of antigen-binding sites:the hapten-binding site of MOPc-315. *Cold Spring Harbor Quant. Symp. Biochem.*, **41**,627–637.

Padlan, E., Silverton, E., Sheriff, S., Cohen, G., Smith-Gill, S. and Davies, D., (1989). Structure of antibody-antigen complex : crystal structure of the HyHEL-10 F_{AB}-lysozyme complex. *Proc. Natl. Acad. Sci. USA*, **86**,5938–5942.

Palm, W. and Hilschmann, N., (1975). Die Primärstruktur einer kristallinen monoklonalen immunoglobulin-L-Kette vom κ-Typ, Subgruppe I (Bence-Jones-Protein Rei), Isolierung und Charakterisierung der tryptischen Peptide; die vollständige Aminosäuresequenz des Proteins. *Hoppe-Seyler's Z. Physiol. Chem.*, **356**,167–191.

Palmer, K. and Sheraga, H., (1991). Standard-geometry chains fitted to x-ray derivedstructures: Validation of the rigid-geometry approximation.i. chain closure through a limited search of loop conformations. *J. Comp. Chem.*, **12**,505–526.

Parks, D., Bryan, V., Oi, V. and Herzenberg, I., (1979). Antigen-specific identification and cloning of hybridomas with a fluorecence-activated cell sorter. *Proc. Natl. Acad. Sci. USA*, **76**,1962–1966.

Paul, P., Burney, P., Campbell, M. and Odguthorpe, D., (1990). The conformational prefrences of γ–lactamand its role in constraining peptide structure. *J. Comp.-aided. Mol. Des.*, 4,239–253.

Pedersen, J. T., (1992). MUL : Iterative multiple fitting program; JPLOT : A multipurpose graph displaying program. Unpublished.

Pedersen, J., Campbell, R., Carter, C., Martin, C., Rose, D., Ruker, F., Strong, R., He, X. and Rees, A., (1991). Modelling antibody combining sites : A method for prediction of the entire variable domain structure. *Document in preparation.*

Perelson, A. S., (1989). Immune network theory. *Immun. Rev.*, **110**.5,36.

Pimm, M., (1988). Drug-monoclonal antibody conjugates for cancer therapy: Potentials and limitations. *CRC Critical Reviews in Therapeutic Drug Carrier Systems*, **5**.3,189–225.

Ponder, J. and Richards, F., (1987a). Internal packing and protein structural classes. *Cold Spring Harbor Quant. Symp. Biochem.*, **52**,421–428.

Ponder, J. and Richards, F., (1987b). Tertiary tempaltes for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, **193**,775–791.

Porter, R., (1958). Separation and isolation of fractions of rabbit gamma-globulin containing the antibody and antigenic combining site. *Nature (London)*, **182**,670–671.

Porter, R., (1959). The hydrolysis of rabbit γ-globulin and antibodies with crystalline papain. *Biochem. J.*, **73**,119–127.

Press, W., Flannery, B., Teukolsky, S. and Vetterling, W., (1990). *Numerical Recipes in C - The Art of Scientific Computing.* Cambridge University Press, 1'th third printing edition.

Rees, A. R. and de la Paz, P., (1986). Investigating antibody specificity using computer graphics and protein engineering. *Trends Biochem. Sci.*, **11**,144–148.

Rees, A. R., Martin, A. C. R., Roberts, S. and Cheetham, J. C., (January 1989). Combining sites and epitopes defined by molecular modelling, protein engineering and NMR. In *Proceedings of the UCLA Symposia on Molecular and Cellular Biology: Protein and Pharmaceutical Engineering.*

Reichman, L., Clark, M., Waldmann, H. and Winter, G., (1988). Reshaping human antibodies for therapy. *Nature (London)*, **332**,323–327.

Rini, J. M., Schulzegahmen, U. and Wilson, I. A., (1992). Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science*, **255**.5047,959–965.

Roberts, S., Cheetham, J. and Rees, A., (1987). Generation of an antibody with enhanced affinity and specificityfor its antigen by protein engeneering. *Nature (London)*, **328**,731–734.

Rose, D. R., Strong, R. K., Margolis, M. N., Gefter, M. L. and Petsko, G. A., (1990). Crystal structure of the antigen-binding fragment of the murine anti-arsonate monoclonal antibody 36-71 at 2.9Å resolution. *Proc. Natl. Acad. Sci. USA*, **87**,338–342.

Rose, D., Przybylska, M., To, R., Kayden, C., Oomen, R., Vorberg, E., Young, N. and Bundle, D., (1992). Document in preparation. Preliminary structure entry in the Brookhaven database.

Rossmann, M., (January 1993). Structure of human rhinovirus complexed with its receptor molecule. In *Protein Eng.*, page 1. Miami Bio/Technology winter symposium.

Rudikoff, S., Satow, Y., Padlan, E. A., Davies, D. R. and Potter, M., (1981). Kappa chain structure from a crystallized murine F'_{AB}: The role of the joining segment in hapten binding. *Mol. Immunol.*, **18**,705–711.

Sastry, L., Mubaraki, M., Janda, K. D., Benkovic, S. J. and Lerner, R. A., (1991). Screening combinatorial antibody libraries for catalytic acyl transfer-reactions. *Ciba foundation symposia*, **159**.145,155.

Saul, F. A., Amzel, L. M. and Poljak, R. J., (1978). The preliminary refinement and strcutural analysis of the F_{AB} fragment from human immunoglobulin New at 2.0Å resolution. *J. Biol. Chem.*, **253**,585–597.

Saul, F. and Poljak, R., (1992). Crystal structure of the F_{AB} fragment from the human myeloma immunoglobulin IgG HIL at 1.8 ångstrøms resolution. *Document in preparation.* Preliminary structure entry in the Brookhaven database.

Schiffer, M., Ainsworth, C., Xu, B., Carperos, K., Olsen, A., Solomon, F., Stevens, C. and Chang, H., (1989). Structure of a second crystal form of Bence-Jones protin LOC: Strikingly different domain association ti two crystal forms of a single protein. *Biochemistry*, **28**,4066.

Schilling, J., Clevinger, B., Davie, J. M. and Hood, L., (1980). Amino acid sequence of homogeneous antibodies to dextran and DNA rearrangements in heavy chain V-region gene segments. *Nature (London)*, **283**,35–40.

Schroeder, H., Hillson, J. and Perlmutter, R., (1989). Structure and evolution of mammalian V_H families. *International Immunology*, **2**,41–49.

Schroeder, H. and Wang, J., (1990). Preferential utilization of conserved immunoglobulin heavy chain variable gene segments during human fetal life. *Proc. Natl. Acad. Sci. USA*, **87**.

Searle, S. M. J., (1992). SR: (Sequence Reader) a program for the analysis of sequence alignments. Unpublished.

Segal, D., Padlan, E., Cohen, G., Rudikoff, S., Potter, M. and Davies, D., (1974). The three-dimensional structure of a phosphorylcholine binding mouse immunoglobulin F_{AB} and the nature of the antigen binding site. *Proc. Natl. Acad. Sci. USA*, **71**,4298.

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. and Davies, D. R., (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci. USA*, **84**,8075–8079.

Shokat, K. M. and Schultz, P. G., (1991). The generation of antibody combining sites containing catalytic residues. *Ciba foundation symposia*, **159**.118,144.

Singh, J., Saldanha, J. and Thornton, J., (1991). A novel method for the modeling of peptide ligands to their receptors. *Protein Eng.*, **4**.3,251–261.

Singh, J. and Thornton, J., (1990). Sirius - an automated method for the anaslysis of the preffered packing arrangements between protein groups. *J. Mol. Biol.*, **211**.3,595–615.

Stanfield, R. L., Fieser, T. M., Lerner, R. A. and Wilson, I. A., (1990). Crystal-structures of an antibody to a peptide and its complex with peptide antigen at 2.8 å. *Science*, **248**.4956,712–719.

Still, C., Tempczyk, A., Hawley, R. and Hendrickson, T., (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **122**,6127–6129.

Strohal, R., Helmberg, A., Kroemer, G. and Kofler, R., (1989). Mouse V_H gene classification by nucleic acid sequence similarity. *Immunogenetics*, **30**,475–493.

Strong, R., Campbell, R., Rose, D., Petsko, G., Sharon, J. and Margolies, M., (1991). Three-dimmensional structure of murine anti-p-azophenylarsonate F_{AB} 36-71. 1.x-ray chrystallography,site-directed mutagenesis, and modeling of the complexwith hapten. *Biochemistry*, **30**,3739–3748.

Suckling, C. J., Tedford, M. C., Bence, L. M., Irvine, J. I. and Stimson, W. H., (1992). An antibody with dual catalytic activity. *Bioorganic & medicinal chemistry letters*, **2**.1,49–52.

Summers, N. L., Carlson, W. D. and Karplus, M., (1987). Analysis of sidechain orientations in homologous proteins. *J. Mol. Biol.*, **196**,175–198.

Sutcliffe, M. J., Hayes, F. R. F. and Blundell, T. L., (1987a). Knowledge based modelling of homologous proteins part II: Rules for the comformation of substituted sidechains. *Protein Eng.*, **1**,385–392.

Sutcliffe, M., Haneef, I., Carney, D. and Blundell, T., (1987b). Knowledge based modelling of homologous proteins, part I:three-dimmensional frameworks derived from the simultaneous superimposition of multiple structures. *Protein Eng.*, **1**,377–384.

Sutor, D., (1958a). The structures of the pyrimidines: VI the structure of theophyllin. *Acta Crystallogr.*, **11**,83–87.

Sutor, D., (1958b). The structures of the pyrimidines: VII the structure of caffein. *Acta Crystallogr.*, **11**,453–458.

Tainer, J., Getzoff, E., Paterson, Y., Olson, A. and Lerner, R., (1985). The atomic mobility component of protein antigenicity. *Annu. Rev. Immunol.*, **3**,501–535.

Taub, R., Gould, R. J., Garsky, V. M., Ciccarone, T. M., Hoxie, J., Friedman, P. A. and Shattil, S. J., (1989). A monoclonal-antibody against the platelet fibrinogen receptor contains a sequence that mimics a receptor recognition domain in fibrinogen. *J. Biol. Chem.*, **264**.1,259–265.

Thornton, J., Sibanda, B., Edwards, M. and Barlow, D., (1988). Analysis,design and modification of loop regionsin proteins. *BioEssays*, **8**,63–69.

Tomlinson, I., Walter, G., Marks, J., Llewelyn, M. and Winter, G., (1992). The repertoire of human germline V_H sequences reveals about fifty groups of V_H segments with different hyper variable loops. *J. Mol. Biol.*, **227**,776–798.

Tramontano, A., (10th-11th of September 1992). Presented at: An international meeting of the bichemical society & the royal society of chemistry. In *Engeneering Antibodies for Therapy.*

Tramontano, A., Chothia, C. and Lesk, A., (1989). Structural determinants of the conformations of medium sized loops in proteins. *Proteins: Struct., Funct., Genet.*, **6**,382–394.

Verhoeyen, M., Milstein, C. and Winter, G., (1988). Reshaping human antibodies: Grafting an antilysozyme activity. *Science*, **239**,1534–1536.

Verhoeyen, M. E., Saunders, J. A., Broderick, E. L., Eida, S. J. and Badley, R. A., (1991). Reshaping human monoclonal-antibodies for imaging and therapy. *Disease markers*, **9**.3,4.

Vila, J., Williams, R., Vasquez, M. and Scheraga, H., (1991). Emperical solvation models ca be used to differentiate native from nera-native conformations of bovine pancreatic trypsin inhibitor. *Proteins: Struct., Funct., Genet.*, **10**.10,199–218.

Viswanadhan, V. N., (1987). Hydrophobicity and residue-residue contacts in globular-proteins. *International journal of biological macromolecules*, **9**.1,39–48.

Wang, D., Liao, J., Mitra, D., Akolkar, P., Gruezo, F. and Kabat, E., (1991). The repertoire of antibodies to a single antigenic determinant. *Mol. Immunol.*, **28**.12,1387–1397.

Ward, E. S., Gussow, D., Griffiths, A. D., Jones, P. T. and Winter, G., (1989). Binding activities of a repertoire of single immunoglobulin variable domains secreted from escherichia-coli. *Nature (London)*, **341**.6242,544–546.

Weber, I. T., (1990). Evaluation of homology modeling of hiv protease. *Proteins-structure function and genetics*, **7**.2,172–184.

Weiner, S., Kollman, P., Singh, U., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, **106**,765–784.

Wenger, T. L., Butler, V. P., Haber, E. and Smith, T. W., (1985). Treatment of 63 severely digitalis-toxic patients with digoxin-specific antibody fragments. *J. Am. Coll. Car.*, **5**.5.

Williams, W. V., Kieberemmons, T., Vonfeldt, J., Greene, M. I. and Weiner, D. B., (1991). Design of bioactive peptides based on antibody hypervariable region structures - development of conformationally constrained and dimeric peptides with enhanced affinity. *J. Biol. Chem.*, **266**.8,5182–5190.

Williams, W. V., London, S. D., Weiner, D. B., Wadsworth, S., Berzofsky, J. A., Robey, F., Rubin, D. H. and Greene, M. I., (1989a). Immune-response to a molecularly defined internal image idiotope. *J. Immunol.*, **142**.12,4392–4400.

Williams, W. V., Moss, D. A., Kieberemmons, T., Cohen, J. A., Myers, J. N., Weiner, D. B. and Greene, M. I., (1989b). Development of biologically-active peptides based on antibody structure. *Proc. Natl. Acad. Sci. USA*, **86**.14,5537–5541.

Winter, G. and Milstein, C., (1991). Man-made antibodies. *Nature (London)*, **349**,293–299.

Wu, T. and Kabat, E., (1970). An analysis of the sequences of the variable regions of bence jones proteins and myeloma light chains and their implications for antibody coplementarity. *J. Exp. Med.*, **132**,211–250.

Zachau, H., (1990). The human immunoglobulin κ locus and some of its acrobatics. *Biol.Chem. Hoppe-Seyler*, **371**,1–5.