Structural Aspects

of

Molecular Recognition

A thesis submitted for the degree of Doctor of Philosophy of the University of London.

Peter Hamilton Walls

October 1992

Biomolecular Modelling Laboratory
Imperial Cancer Research Fund,
44 Lincoln's Inn Fields,
London WC2A 3PX.

and

The Biochemistry and Molecular Biology Department
University College London
Gower Street,
London WC1E 6BT

Abstract

This thesis describes the design, implementation and application of a novel docking algorithm. Chapter 1 reviews some important facts about proteins and protein structure. Several molecular recognition systems are examined in detail. This Chapter also reviews a representative set of recent protein/protein docking methods and discusses their relative merits.

Chapter 2 sets out the aims of the new docking algorithm, called DAPMatch, and gives full details of its implementation on a parallel architecture computer. The testing of the algorithm is also discussed.

Subsequent chapters describe the application of the DAPMatch algorithm to a number of docking problems. DAPMatch is used to reconstruct the known structures of three antibody/lysozyme complexes, using the unbound structure of lysozyme. For the first time a model of the D1.3 antibody is used as a target molecule for a docking algorithm. These results are presented in Chapter 3 and analysed in detail to demonstrate their significance; non-native solutions are also examined. Chapter 4 describes the practical use of the DAPMatch algorithm in a modelling situation, to construct a hypothetical structure for the high molecular weight epidermal growth factor complex. Chapter 5 describes the adaptation of the DAPMatch algorithm to investigate α-helix/α-helix docking, and presents the results obtained. Chapter 6 explains the conclusions that were derived from this work, and suggests possible future enhancements to the algorithm.

Contents

Chapter1

Introduction

Chapter2

Method

Chapter3

Antibody/Antigen Results

Chapter4

A Practical Application of DAPMatch to the modelling of the High Molecular Weight Epidermal Growth Factor complex

Chapter5

α-Helix packing

Chapter 6

Conclusion

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Abbreviations

ASA	Accessible Surface Area
BPTI	Bovine Pancreatic Trypsin Inhibitor
CDR	Complementarity Determining Region
COSY	Correlation Spectroscopy
cpu	Central Processing Unit
DAP	Distributed Array of Processors
DAPFORTRAN	A parallel version of FORTRAN for the DAP computer
DAPMatch	Steric complementarity search program for the DAP computer
DNA	DeoxyriboNucleic Acid
EGF	Epidermal Growth Factor
EGFBP	Epidermal Growth Factor Binding Protein
F_{ab}	Antigen binding fragment of an antibody.
F_c	The antibody fragment that was first to be crystallised.
FORTRAN	FORmula TRANslation language
HMWEGF	High Molecular Weight Epidermal Growth Factor
ICRF	Imperial Cancer Research Fund
I/O	Input/Output
NADPH	Nicotinamide Adenine Dinucleotide Phosphate (reduced form)
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy
PDB	The Brookhaven Protein Databank
rms	Root mean square

rmsd	Root mean square deviation
SCR	Structurally Conserved Regions
SIMD	Single Instruction Multiple Datastream
V_H	Variable Heavy domain of an antibody
V_L	Variable Light domain of an antibody
VR	structurally Variable Regions

Acknowledgements

Special thanks must go to Mike Sternberg (my supervisor) for understanding and Cristiena for waiting.

I would also like to thank all the members of the Biomolecular Modelling Department at Imperial Cancer Research Fund both for their help and for their friendship: Alex Adzhubei, Paul Bates, Jonathan Hirst, Suhail Islam, Richard Lewis, Stephen Pickett, Mansoor Saqi and Paul Snape.

I am grateful to Professor Janet Thornton, my UCL supervisor, for her useful comments and guidance.

I am grateful to the Imperial Cancer Research fund for the three year studentship which funded this work.

Dr A. Lesk and Dr S. Phillips kindly provided some of the protein structures used in this thesis.

Finally, my thanks to my parents, Peter Bratcher, Marcia Hyde, Andrew Mackintosh, Andrew Poole, Dale Rowe, Peter Smith and Helen Weiss for their continuing friendship.

Chapter 1

Introduction

1.1. Synopsis

This thesis describes a protein/protein docking procedure and its application to several different systems. This chapter presents some basic facts about proteins, and in particular describes the levels of protein structure. The forces that mediate inter- and intra- protein interactions are discussed. Three systems that involve molecular recognition are described in detail, antibody/antigen systems, enzyme/inhibitor systems and α-helix/α-helix packing. Finally, previous work on protein/protein docking is reviewed and the relative success of the various methods surveyed.

1.2. Protein Structure

The overall structure of proteins can be divided into four hierarchical levels, called primary, secondary, tertiary and quaternary structure.

1.2.1 Primary Structure.

Proteins are polypeptide chains formed from a sequence of amino acid residues. Twenty amino acids commonly occur in proteins (Table 1.1). The chemical properties of the amino acids are diverse, differing in size, hydrophobicity and polarity (Figure 1.1). Amino acids are joined together by peptide links to form a continuous section of chain, called the protein main-chain. Each amino acid has a different side-chain, which is connected to the main-chain carbon alpha atom. Nineteen of the common amino acids share the same main-chain structure (Figure 1.2b). The side-chain of proline is bonded to the main-chain nitrogen atom, giving it a unique main-chain structure. The carbon alpha atom has two possible stereoisomers, but the D-form is not synthesized on the ribosome and exists only in the rare cases where post-translational modification occurs.

The primary structure of a protein is simply its amino acid sequence. The sequences of over 40,000 proteins are reported in the April 1992 release of the OWL database (Akrigg, *et al.*, 1988). In comparison, fewer than 800 protein structures are reported in the January 1992 release of the Protein Data Bank (Bernstein, *et al.*, 1977).

Amino acid name	Three letter code	One letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cystine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Table 1.1

The twenty commonly occuring amino acid residues. The three letter and one letter abbreviations for each amino acid are also given.

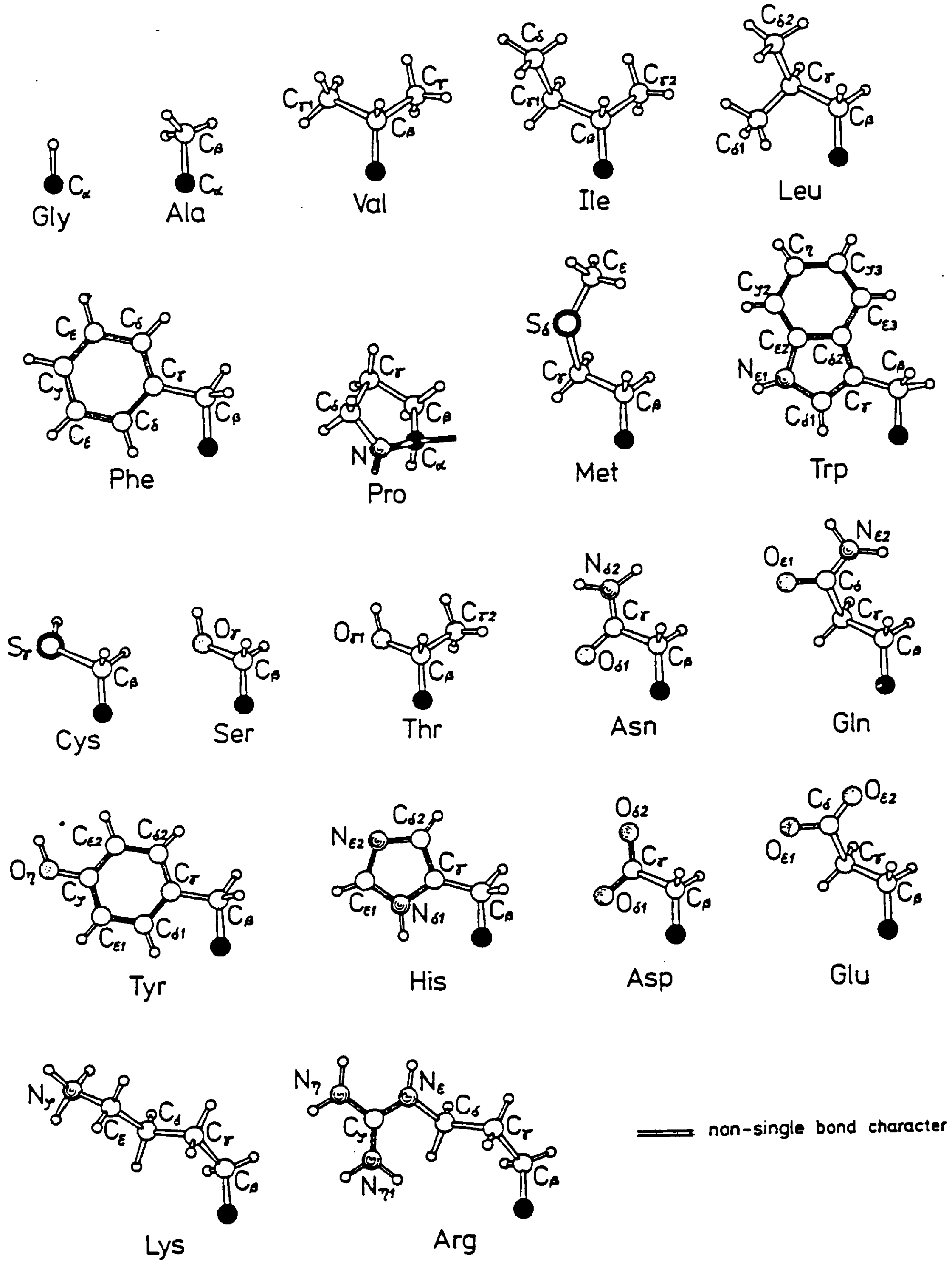


Figure 1.1

The side-chain groups of the 20 commonly occurring amino acids. The atom shaded black is the main-chain Cα atom. A section of the proline main-chain is shown since it is bonded to the proline side-chain. Diagram from Schulz & Schirmer (1979)

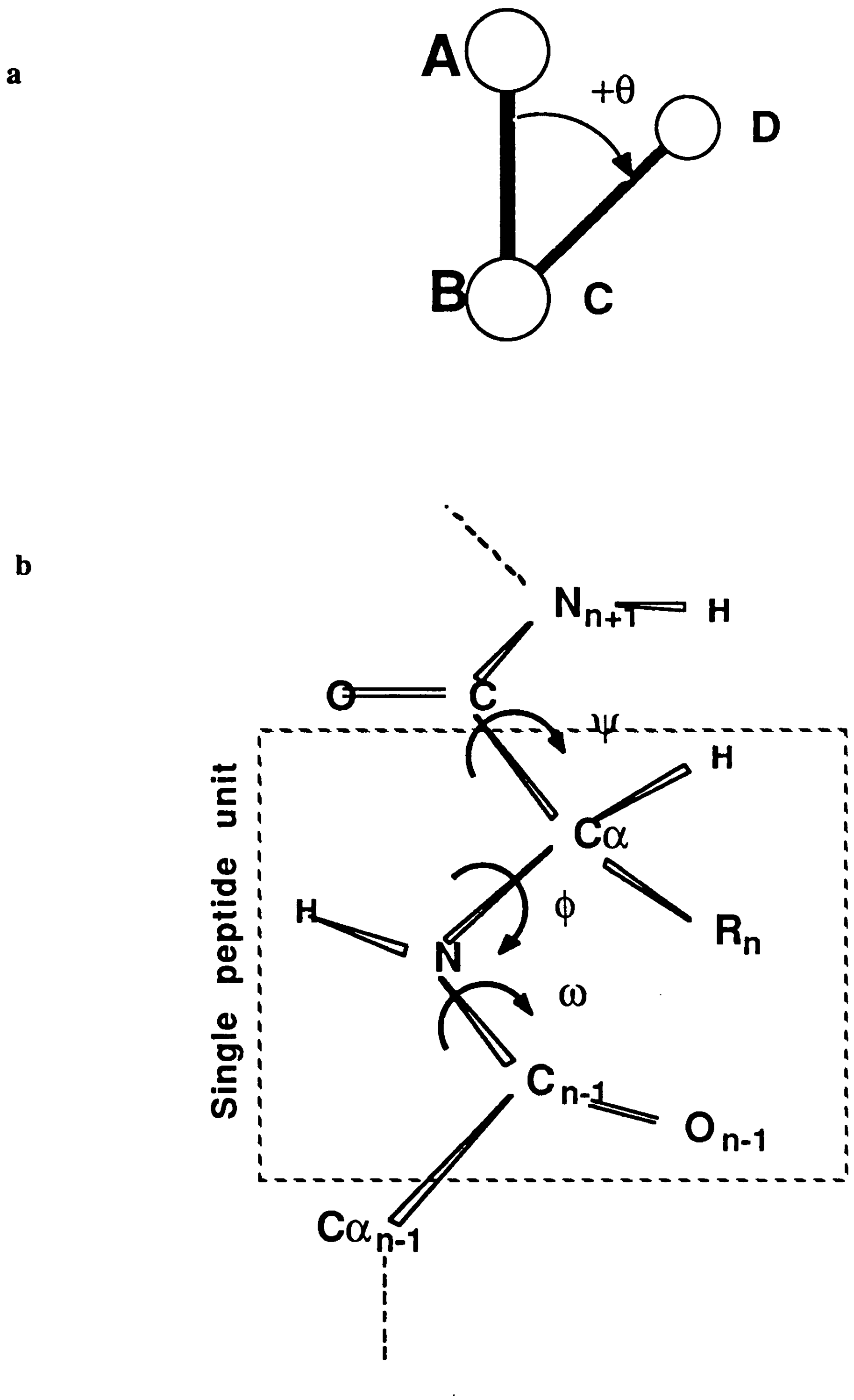


Figure 1.2

a. The dihedral angle θ between atoms A,B,C and D is the angle between the bonds AB and CD measured perpendicular to the bond BC. Clockwise is positive, as shown.

b. The three dihedral angles φ, ψ, and ω determine the conformation of the protein backbone.

1.2.2 Secondary Structure.

The local folding of the protein main-chain produces patterns that are called the secondary structure. The conformation of the protein main-chain can be described using three dihedral angles, ϕ, ψ and ω (Figure 1.2). The peptide bond is planar, due to the partial double bond character of the carbonyl carbon - amide nitrogen bond. This fixes ω at 0° or 180°. Generally, peptides adopt the trans conformation (ω=180°) due to steric hinderances. The exception is proline, where roughly 25% of residues are seen in the cis conformation (ω=0°) (MacArthur and Thornton, 1991).

Ramachandran & Sasisekharan (1968) plotted allowable ϕ, ψ angles (ones which did not lead to steric clashes) (Figure 1.3), and this representation became known as a Ramachandran diagram. The plot clearly shows specific areas of favourable ϕ, ψ angles.

In most protein structures, sections of secondary structure with repeating (ϕ,ψ) angles are seen, the formation of which leads to a compact structure that satisfies hydrogen bonding requirements and shields the hydrophobic residues from the solvent.

The α-helix is a simple element of secondary structure, it was first predicted by Pauling *et al.* (1951) twelve years prior to the first protein crystal structure being solved. It allows a regular main-chain hydrogen bonding pattern to be set up between each residue and the residue four positions further along in the sequence (Figure 1.4a). In this way, a continuous section of the protein chain can satisfy its main-chain hydrogen bonding constraints. In theory α-helices could have either a right handed or left handed twist. The right handed form has repeating (ϕ,ψ) angles of roughly (-60°,-50°) (Section α in Figure 1.3a) and the left handed form has repeating angles of about (60°,50°). However, L-amino acids impose right-handed

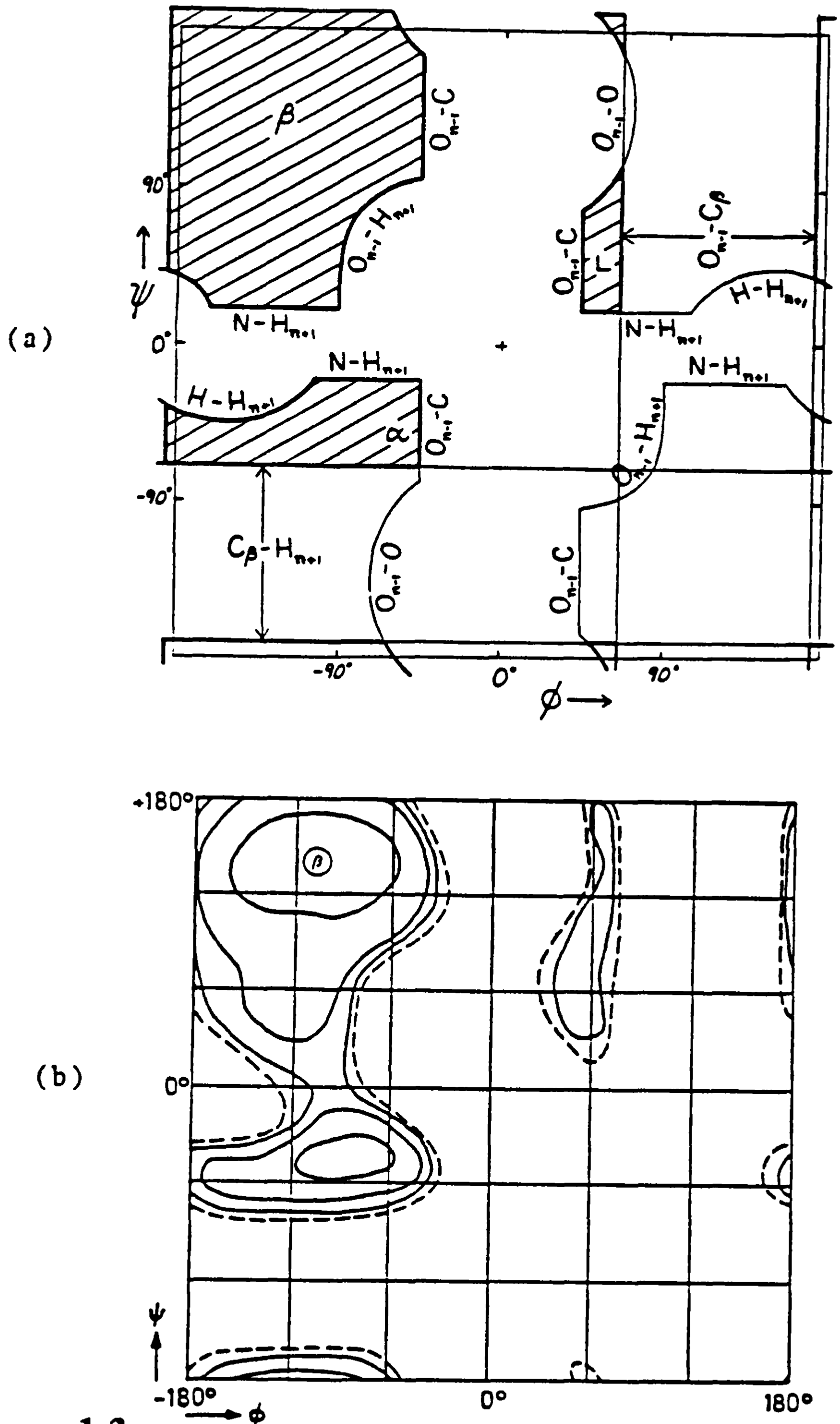


Figure 1.3

(a) A ϕ,ψ plot showing showing which atomic collisions produce restrictions on main chain conformation. The shaded regions are allowed conformations for all residues. (From Richardson, 1981).

(b) A potential energy distribution in the ϕ,ψ plane for a pair of peptide units separated by an alanine residue. The zero kcal/mol contour is dashed, other contours are drawn at -1 kcal/mol intervals. (From Ramachandran *et al.*, 1968).

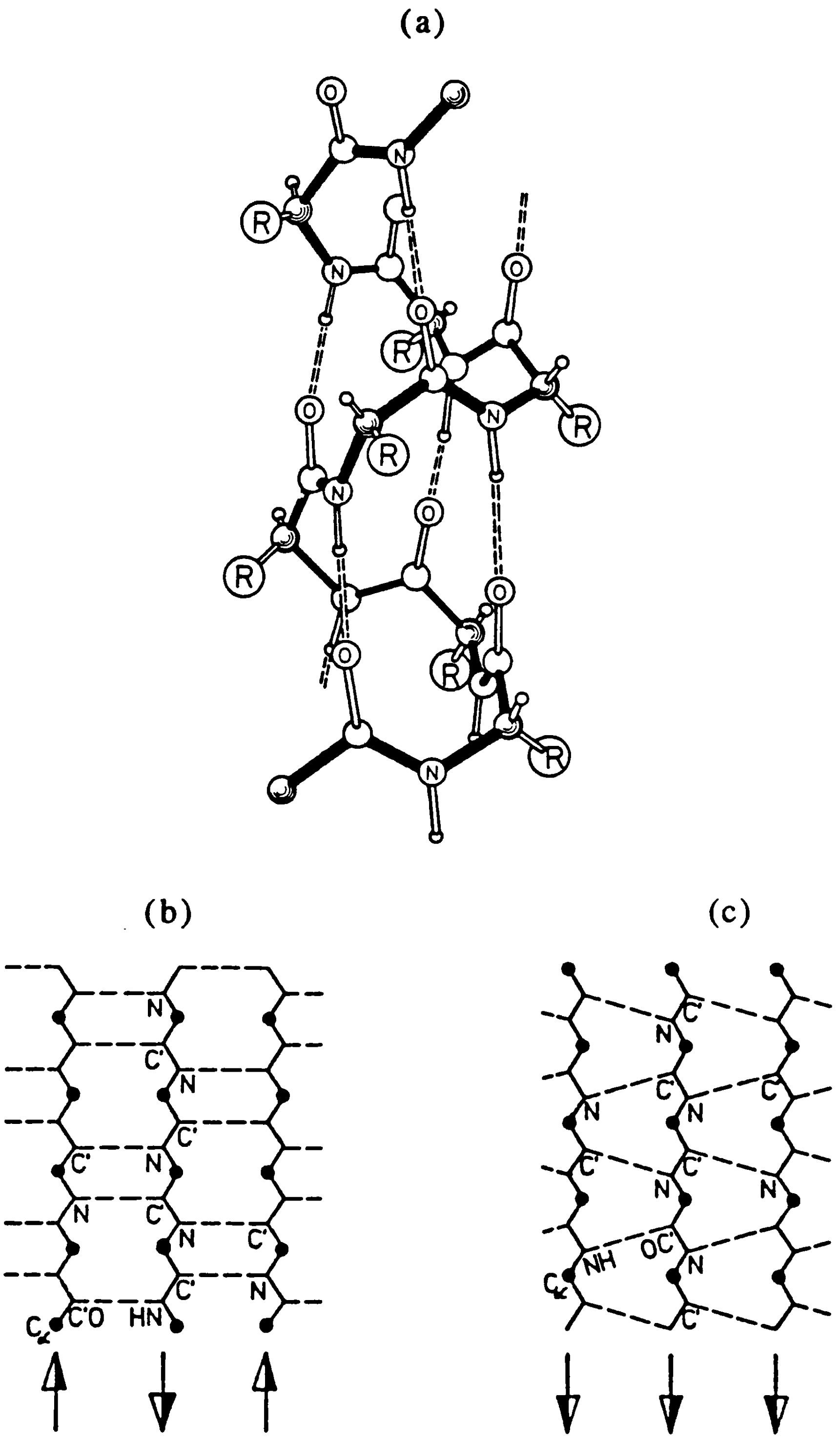


Figure 1.4

The hydrogen bonding patterns (shown as dotted lines) of the α-helix (a), the antiparallel β-sheet (b) and the parallel β-sheet (c). The α-helix forms bonds between residue *i* and *i*+4. Antiparallel β-sheets form hydrogen bonds perpendicular to the strand direction, parallel sheets form staggered bonds between adjacent strands. Diagrams from Schulz & Schirmer (1979).

twist, due to steric restrictions on the protein side-chain. Isolated examples of left-handed α–helix are possible, a single turn of left-handed helix exists in thermolysin (Matthews, *et al.*, 1974).

Few α-helices conform to the predicted ideal geometry. Barlow and Thornton (1988) found only seven regular, linear helices in a sample of 48. The predominate distortion that they found was curvature of the helix axis, induced by side-chain packing and solvent effects. Barlow & Thornton also noted that prolines which occur in helical segments cause kinks. These kinks alter the direction of the helix axis by roughly 25° at a single residue.

Although the α-helix predominates, other classes of helices are observed. The α-helix provides a hydrogen bond between each residue *i* and the residue *i*+4 . Short sections of both 3^{10} helix, which bonds residues *i*,*i*+3, and π helix, which bonds residues *i*,*i*+5, have been observed in proteins (Hendrickson, *et al.*, 1973).

Unlike α-helices, which are formed from continuous sections of protein, β–sheets are built up from several sections, called β–strands. Each residue of the strand has (ϕ,ψ) angles of roughly (-120°,120°) (Section β in Figure 1.3a) The strands pack together, side by side to create the 'sheet'. Hydrogen bonds are formed between adjacent strands. The strands can be arranged in parallel or antiparallel forms (Figure 1.4b). In antiparallel sheets the hydrogen bonds are made perpendicular to the strand direction, whereas parallel sheets have staggered bonds (Figure 1.4c). Mixed β–sheets, with both antiparallel and parallel connections between strands occur (Sternberg and Thornton, 1977). β–sheets are not planar, but generally twist in a well defined manner, with the same handedness for nearly all known structures (Chothia, 1973). The twist of the peptide plane is right handed when viewed along any strand of the sheet. This right-handed twist for each strand means that adjacent β−strands cross each other with a left-handed rotation, as shown in Figure 1.5.

Both the α-helix and the β−strand are linear elements. Changes of direction in the protein chain are necessary to form a compact, globular protein. This is often done by sections of protein with a non-repetitive conformation, called coil. However some turns, particularly short sections of sequence that change the direction of the chain by 180°, can be classified into distinct groups. These turns form a distinct third secondary structure group, the β−turn. Such turns usually occur at the surface of the protein. The classic β−turn, as identified by Venkatachalam (1968), involves four residues which reverse the chain direction and are arranged so that a hydrogen bond exists between the first and last residues of the turn. Venkatachalam proposed three classes of turn, all of which could also exist in mirror image form. Lewis *et al.* (1973) made a study of known protein structures and found that 25% of β−turns do not possess a hydrogen bond. They therefore suggested a broader definition, which recognised a β−turn as four residues, with non-helical conformation, where the distance between the first and fourth residues is less than 7Å. This definition increased the number of recognised classes of turn. Richardson (1981) later rationalised these classes into five clear groups (including mirror images) and one miscellaneous class. Although these definitions have been added to (Wilmot and Thornton, 1988) they are still in regular use.

1.2.3 Tertiary Structure.

The tertiary structure of a protein is a description of the three-dimensional arrangement of an entire polypeptide chain. Proteins can be divided into five general categories, depending on the types of

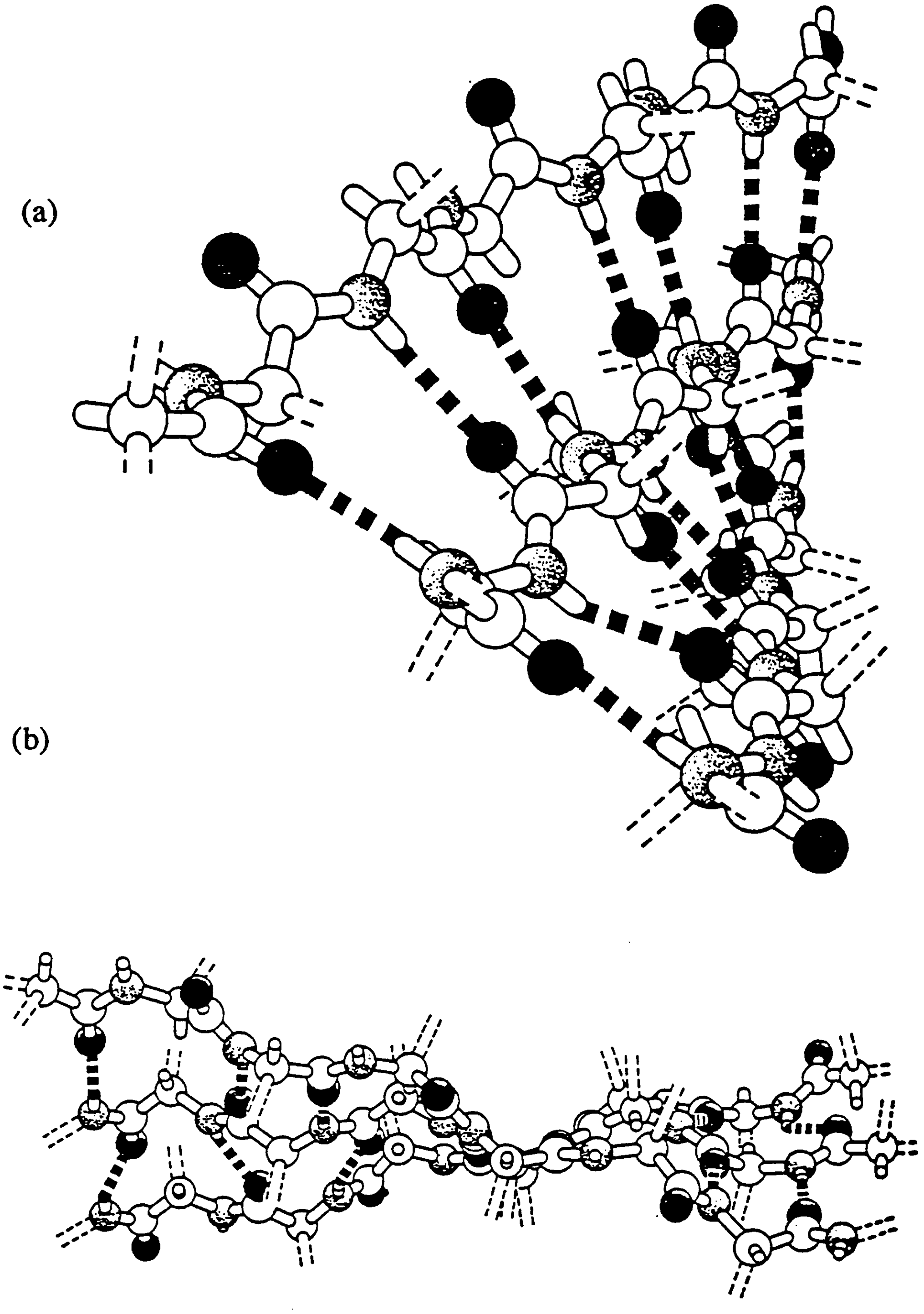


Figure 1.5

The twist of a β-sheet is right handed when viewed along a peptide strand (a) which means that crossing strands give the β-sheet a left handed screw (b). Diagram taken from (Chothia, 1973) secondary structure they form (Levitt and Chothia, 1976). The all-α category contains proteins which primarily contain α-helices with few β-sheets, such as myoglobin (Phillips, 1980) and parvalbumin (Moews and Kretsinger, 1975). Many proteins fall into the all-β category, for example immunoglobulins and chymotrypsin. There are two categories for proteins that contain a mix of the two main secondary structure types. The α/β category includes proteins such as triose phosphate isomerase (Banner, *et al.*, 1976) and phosphofructokinase (Evans and Hudson, 1979) whose secondary structure elements alternate, along the sequence, between α-helices and β–strands. Proteins such as hen egg white lysozyme (Diamond, 1974) form the α+β category as they have both α- and β–structure but without a regular alternation. The final category is the small metal-rich or disulphide-rich class, the structure of these proteins is dominated by either a metal ligand or by a large number of disulphide bonds, such as rubredoxin (Watenpaugh, *et al.*, 1979) and high potential iron protein (Carter, *et al.*, 1974).

Analysis of available protein structures reveals that secondary structure elements often pack together into recognisable motifs. These motifs can involve only two secondary structure elements, such as the αα twisted pair, or many, such as the TIM barrel which is made up from the regular packing of eight helices and eight strands. Other large scale motifs commonly found in proteins are the Greek Key, the Jelly Roll and the Immunoglobulin domain .

Secondary structure elements often pack together into small, regular motifs. Such elements include the βαβ unit, the αα twisted pair and the ββ hairpin. Large motifs are often found to be made up from multiples of these smaller motifs. The TIM barrel is a repeating βαβ unit. The tertiary structure class strictly includes both these large motifs and the smaller ones from which they are constructed. This seems to include two levels of structure in a single classification and so motifs involving only a few elements of secondary structure are often considered to be lower level structure than true tertiary structure, and are classified as supersecondary structure.

Another useful classification, is the protein domain (Wetlaufer, 1973). In a large protein domains are the separate sections that are seemingly independent and they sometimes perform an identifiably different function. The domain structure of a protein is often clearly discernable, as in the immunoglobulins. Domains are generally formed from a single, continuous section of protein sequence. The main-chain of a domain forms a compact region that can be almost totally surrounded by a closed surface. The initial definition of a domain, as proposed by Wetlaufer, stated that only two sections of chain could cross this surface. These were the connections from the previous domain and to the next domain, unless the domain was at the N or C terminus. As more protein structures were elucidated, it became clear that protein with two domains often have crossover connections, as in phosphofructokinase. The N and C terminal sections of the protein form 'arms' which loop over and pack against the opposite domain.

1.2.4 Quaternary Structure.

All the levels of structure discussed so far refer to the organisation of a single chain of protein. Larger proteins often comprise several independent protein subunits which pack together to create the active form of the protein. Each subunit is a single polypeptide chain which can be independently stable, and able to fold. Quaternary 'structure is a description of how these subunits pack together to form a single, stable entity. The subunits involved may be identical, e.g. triosephosphate isomerase, or different, e.g. hæmoglobin. If the subunits are identical then polymerisation is often symmetrical, with each monomer using the same section of its surface in the association. It is also possible for identical subunits to associate in an unsymmetric way, using different sections of surface, e.g. hexokinase (Bennett and Steitz, 1980).

The individual subunits of a polymeric protein may be completely independent, or co-operative effects may occur (Chothia, *et al.*, 1976). Hæmoglobin is an example of a protein that exhibits this effect (Baldwin and Chothia, 1979). Hæmoglobin consists of two different subunits, each occurring twice to form a tetramer. Each subunit binds a hæm group; these groups bind to oxygen, which is then carried around the bloodstream by the protein. When oxygen binds to any one of the subunits the oxygen affinity of the others is increased. This co-operative effect is transmitted across the subunit boundaries by relatively small changes in tertiary and quaternary structure.

1.3. Protein Structure Determination.

1.3.1 X-Ray Crystallography.

The majority of known protein structures have been determined by X-ray crystallography (Blundell and Johnson, 1976). The first stage of crystallography is obtaining a well-ordered protein crystal. Protein crystals are grown from supersaturated solutions. Proteins do not readily form regular crystals; they are generally an ellipsoidal shape and do not pack into space filling geometries. Individual molecules in a protein crystal, therefore, only interact with neighbouring molecules in a few, small regions. The large gaps in the protein crystal are filled with solvent molecules. Although the large amount of disordered solvent present (often 50% of the crystal) makes it difficult to obtain well-ordered crystals, it does mean that the protein adopts a conformation in the crystal which is not markedly different from that adopted in solution.

Once a crystal has been obtained it is irradiated with monochromatic X-rays. Most X-rays pass straight through the crystal, but some interact with atoms in the protein molecule. This interaction causes X-rays to be scattered in all directions. The scattered X-rays interfere with each other, generally cancelling each other out. However, in certain positions, and for certain angles, the X-rays interfere constructively and a diffracted X-ray beam leaves the crystal. Such beams are recorded as a spot on a photographic film or electronic detector, placed a known distance from the crystal. Recording the amplitude of the diffracted beam loses a vital piece of information, the phase of the beams which leave the crystal. This problem is circumvented by introducing a small number of extra atoms into the protein crystal. The atoms which are added are 'heavy', metal atoms. These atoms make a large contribution to the X-ray pattern, and so only a few atoms need be added, thus preserving the protein structure. By examining the change in the diffraction pattern it is possible to determine the position of the metal atoms, and their phases. Use of two heavy metal derivatives enables the phase contributions from the protein to be estimated.

Knowledge of both the amplitude and the phase of each spot of the X-ray diffraction pattern allows an electron density map to be calculated. This map is imperfect. Errors arise due to disorder in the crystal, incorrect phase determination and inaccuracy from the X-ray diffraction experiment itself. The electron density map is then interpreted. This involves tracing the path of the protein main-chain through the electron density, and positioning side-chains so that they agree with the observed density as well as possible. This interpretation is often subjective, and is made more difficult by portions of the map which contain little or no observed density, a particular problem at the protein surface, where the side-chains are most mobile. The amino acid sequence of the protein is essential for the correct tracing of the protein backbone.

The next step is to refine the interpreted structure. This is a computerized procedure to minimise the difference between the observed diffraction amplitudes and those calculated for the interpreted structure. This is done by making small changes in the position of atoms, and is often linked to a procedure which normalises the bond lengths and bond angles seen in the protein. During the refinement process it is often possible to assign a B value to each atom. This value indicates the relative mobility of that atom (Artymiuk, *et al.*, 1979).

It is impossible to state the error in the atomic co-ordinates of a protein which has been determined by X-ray crystallography. Two measures of quality that are often quoted are the resolution of the structure and its R-factor. The resolution is simply a measure of the amount of diffraction data collected. More observations means the electron density map is more detailed, and hence the final structure will be better. Resolution is quoted in Ångstroms, a high value for the resolution implies low quality. A resolution of 5Å allows only the general shape of the protein to be established, possibly with some secondary structure assignment. Resolutions of around 3Å allow side-chain position to be determined, many large proteins are determined at this level. Resolutions of between 1Å and 3Å produce very well defined electron maps, with spheres of electron density at atom positions. Small proteins are often resolved at this level. The resolution has several drawbacks as a measure of quality. It is a global measure and hence incorrect areas of the structure cannot be highlighted and it is a measure of the quality of the electron density map. Incorrect interpretation can still result in incorrect structures. A better measure, which overcomes this second problem is the R-factor. The refinement process attempts to minimise the difference between observed and calculated diffraction intensities, the R-factor is a percentage disagreement between the two sets of data. A R-factor of 0% implies that all the observed data is completely explained by the protein structure obtained. Random agreement produces a R-factor of roughly 60%. Most proteins have an R-factor of between 10% and 20%, after refinement. In minimising the difference between observed and expected diffraction intensities the refinement stage is effectively minimising the R factor. This could lead to a good R-factor for a structure that is incorrect. It has been suggested that some of the diffraction data be withheld from the refinement procedure and that this data be used to calculate the R-factor (Bruenger, 1992).

1.3.2 Nuclear Magnetic Resonance.

The other method that has been used to determine precise three dimensional protein structures is nuclear magnetic resonance spectroscopy (N.M.R.) (Wut~~h~~rich, 1990). This method has only recently been extended to proteins from small molecule work. Over thirty protein structure have been solved, mostly with sizes of around 50 residues, but larger structures such as Interleukin 8 with 144 residues have been solved (Clore, *et al.*, 1990).

N.M.R. relies on the magnetic moments (or spin) of the hydrogen nucleus (i.e. a proton). When proteins are placed in a strong magnetic field the spins of the hydrogen nuclei tend to align parallel or antiparallel to the imposed field. It is then possible to excite the spins by subjecting the protein to a pulse of electromagnetic radiation. This excitation will only occur at a certain frequency, which will be in the same broad area of the spectrum for all protons but which also depends upon the local environment of the proton. The variation of characteristic frequency due to environment is called the chemical shift (Figure 1.6a). Conventional one dimensional N.M.R., as used for small molecules, measures the absorption spectra of the sample, in the relevant frequency range. Although this, in principle, gives a unique signal from each non-identical hydrogen, the absorption peaks overlap. This is because the difference in chemical shift is smaller than the resolving power of the experiment. To overcome this problem two dimensional N.M.R. was introduced. Here the experiment measures interactions between two separate spins at different frequencies, ω_1 and ω_2. These interactions are represented as cross-peaks on a 2-D N.M.R. map, as shown in Figure 1.6b. The exact nature of the N.M.R. experiment determines the type of spin-spin interaction examined. COSY (correlation spectroscopy) shows hydrogen atoms which are connected by three or fewer covalent bonds. Hence COSY spectra show hydrogen atoms which belong to the same residue. The NOESY (nuclear Overhauser effect spectroscopy) experiment shows interaction between hydrogen atoms that are less than 5Å apart.

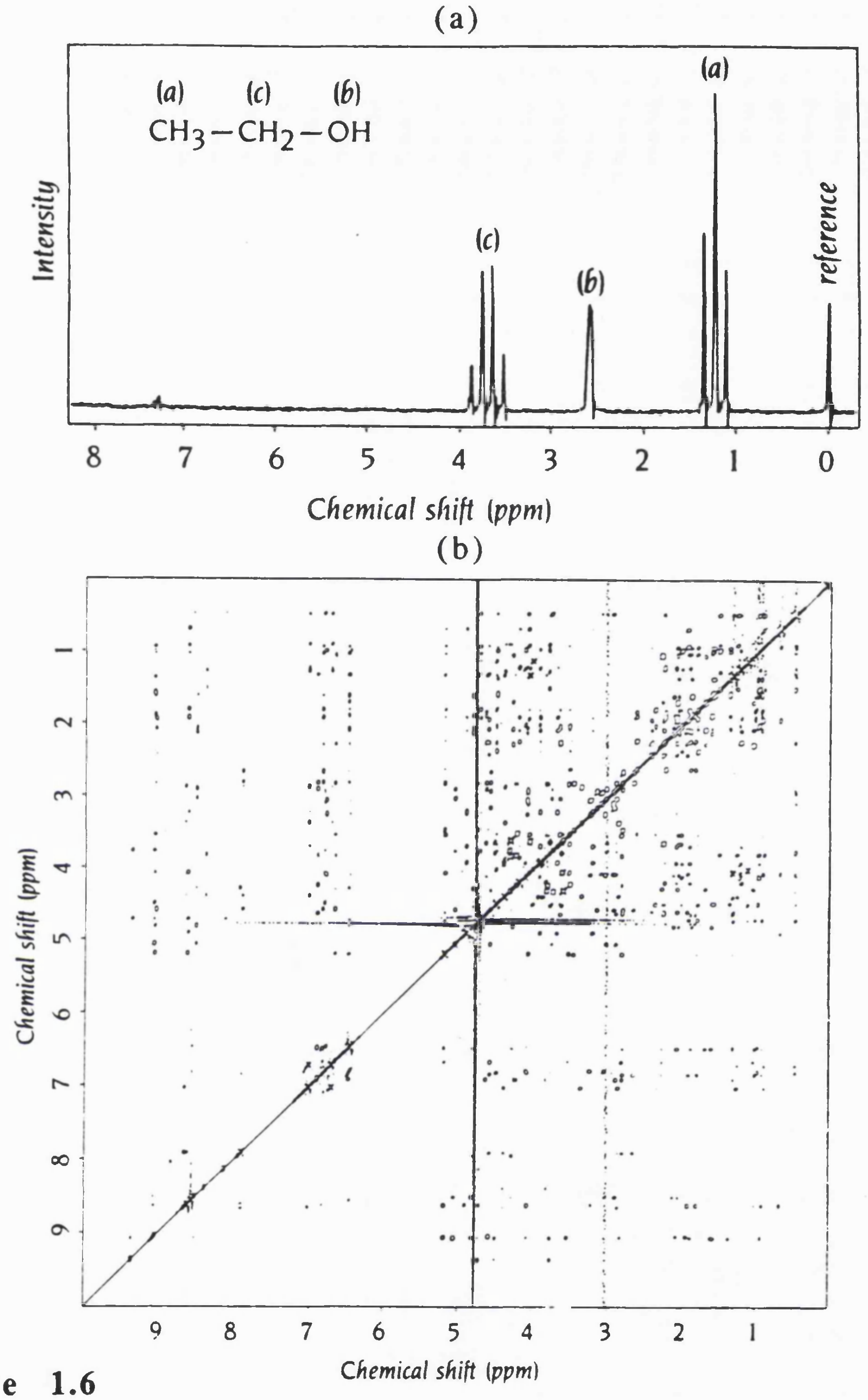


Figure 1.6

(a) A 1-D ^{1}H-NMR spectrum for ethanol. The chemical shift for each hydrogen atom in the labelled groups is clearly visible. The signal from the CH_3 group hydrogens is split into three peaks and the signal from the CH_2 group hydrogens is split into four peaks, due to experimental conditions.

(b) A 2-D NOE NMR spectrum of the C-terminal domain of cellulase. The off-diagonal peaks represent interactions between hydrogen atoms that are separated by less than 5Å in space.

Both diagrams are taken from Branden & Tooze (1991)

Hence NOESY spectra give information on which residues are close together in the folded protein, despite being remote in sequence.

To solve a protein structure by N.M.R. the 2-D spectra must be correctly interpreted to give a series of distance constraints between atoms. If sufficient constraints are obtained the relative positions of all the atoms can be calculated. In real cases certain areas of the protein structure are insufficiently constrained, this leads to multiple possibilities for the solution and hence a family of possible structures are obtained. The quality of an N.M.R. structure cannot be assessed by the same measures as a crystallographic structure. Instead the average root mean square deviation of each structure in the solution family from a representative structure is calculated. The representative structure is the average for all the possible solutions found, this structure has no physical meaning and is not likely to be a member of the solution family.

Comparisons between independently solved X-ray and N.M.R. structures for the same protein have been made (Billeter, *et al.*, 1989). Such studies show a large degree of agreement of between the different methods, particularly in the interior of the protein. At the exterior of the protein, where crystal structure tend to have large B-values and N.M.R. structures have few constraints, differences in side-chain orientation are found.

1.4. Modelling by Homology

The existence of secondary and tertiary structure motifs, and the tendency for three-dimensional structure to be more highly conserved than primary structure (Bajaj and Blundell, 1987) implies that homologous proteins (ones with similar sequences) should have similar three dimensional structures. Knowledge of the full structure of one protein could therefore be used to suggest a structure of an homologous protein. This idea was first exploited by Browne *et al.* (1969) to create a model for bovine α-lactalbumin from the structure of hen egg white lysozyme. This first model was created by hand, later attempts were able to utilise computer graphics.

The first step in any such modelling is to search for similarities between the sequence of the protein of unknown structure and those of the known structures. Early methods used sequence alignment algorithms, such as those of Needleman & Wunsch (1970) and Waterman, (1983), to detect regions of sequence similarity and report the statistical significance of such regions. The structure of the protein with the closest homology was then used as a template. Insertions, deletions and replacement of amino acids were then carried out using a computer graphics program, such as FRODO (Jones, 1978). Energy minimization was then carried out to produce a structure that was well packed. This method was carried out on molecules such as insulin-like growth factors, serine proteinases, immunoglobulins and others. The success of several model building studies was reviewed by Ripka (1986).

Later methods used the tertiary structure of a series of homologous proteins to produce more accurate models (Blundell, *et al.*, 1987; Greer, 1990). These methods make a structural alignment of all the proteins in an homologous family. This alignment matches residues which are in the same position in the structure. The family structural aligment is then examined for structurally conserved regions (SCRs). These regions generally form secondary structure elements, essential framework residues and any active sites in the protein. The SCRs are characteristic of the protein family and are assumed to have the same structure. Separating the SCRs are variable regions (VRs), generally surface residues, whose structure is not conserved between family members. The sequence of an unknown protein, homologous to a particular family, can be fitted on to the SCRs of that family. To model the VRs sequence alignments are made with the corresponding fragments of each homologous protein, and the best match is chosen as a template. If no suitable sections are found then they are generated using *ab initio* computer methods (Bruccoleri and Karplus, 1987; Moult and James, 1986). Greer (1990) applied this method to mammalian serine proteinases and analysed its success. He found the method very reliable, except in cases when sequences violated the expected sequence template in the structurally conserved region, or when homologous variable regions could not be found.

Insertions and deletions often happen in variable regions of the protein. If no sequence match can be found between such a region, often a loop, and any of the corresponding regions in the homologous proteins then a short segment of the protein cannot be modelled. This means that the main-chain of the modelled structure breaks off at one point in space and resumes again a known distance away. One way to assign this segment is to search all the known protein structures for sections that have the correct number of residues and span the correct distance. Jones & Thirup (1986) searched 37 highly refined protein structures for segments that match the spatial requirements of a loop in retinol binding protein. When a comparison was made with the known loop structure it was found that all of the best 20 matches had the correct conformation.

1.5. Protein-Protein Interactions

1.5.1 Covalent Bonds.

The atoms in a protein are held together by covalent bonds. A covalent bond involves the sharing of one or more pairs of electrons between two atoms. For a particular atom type, only electrons in certain electron shells are available for covalent bonding. This means that the covalent bonds of particular atoms occur at particular positions. The resulting bond lengths and bond angles are largely fixed for a specific atomic species. Rotation around the axis of a covalent single bond is allowed, and this gives the protein its flexibility. Cystine is the only amino acid that can make covalent bonds to residues that not adjacent to it in sequence. Two cystine residues that are close in space, but remote in sequence, can be reduced and form a disulphide bridge.

Non-covalent bonding takes place between secondary and tertiary structure elements, stabilising the 3-D structure of a protein. Also, non-covalent bonds play a large part in stabilising protein-protein interactions.

1.5.2 The Leonard-Jones Potential.

The two forces acting on neutral atoms are often, for convenience, combined and considered as a single force (Figure 1.7).

$$F_{neutral} = F_{attract} + F_{repulse}$$

The repulsive force acts when the electron shells of the two interacting atoms begin to overlap. This distorts the shells, moving the electrons into orbits closer to the nucleus. Short radius orbits have higher kinetic energies and hence work must be done to achieve this distortion. As the atoms come closer together the

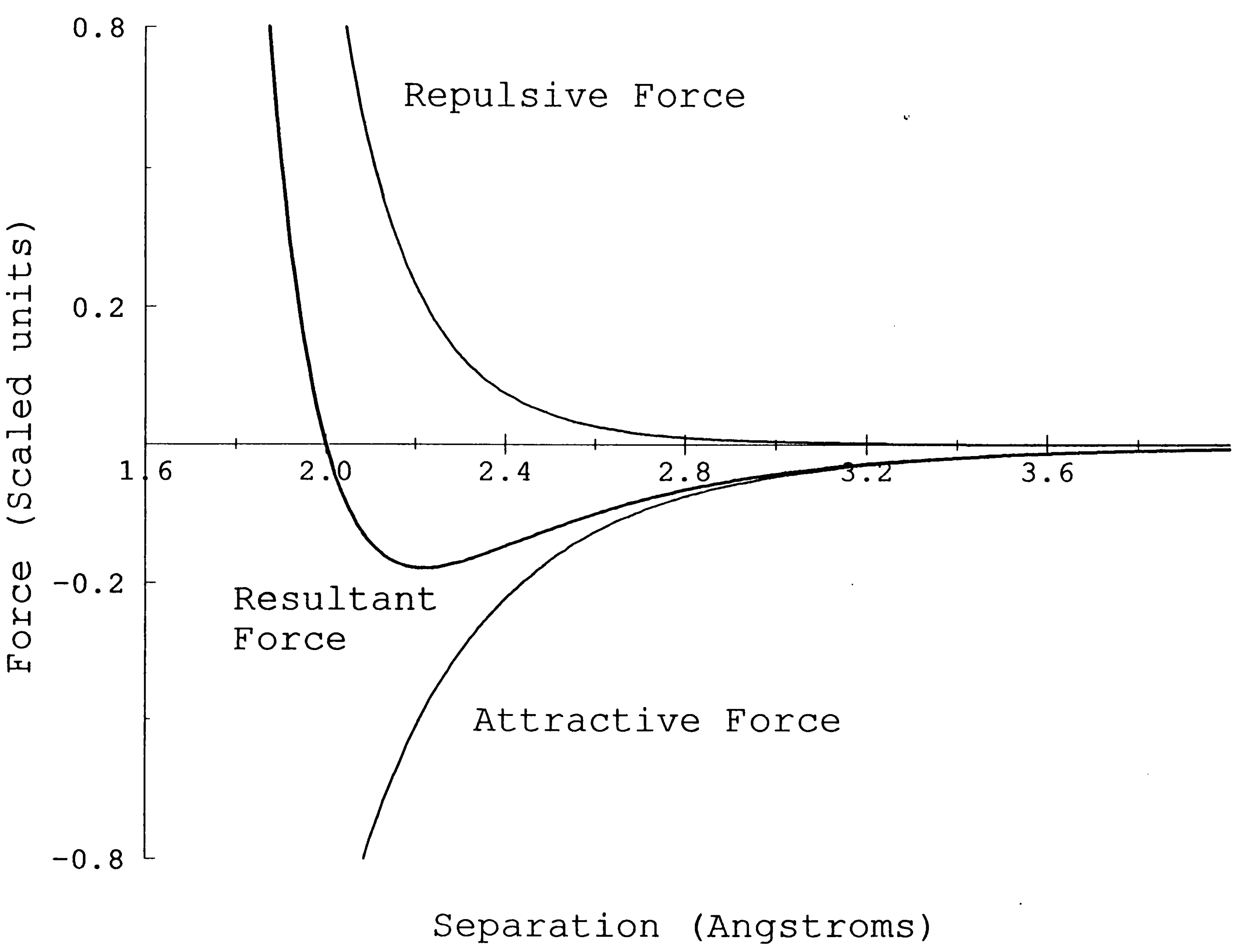


Figure 1.7

The total force acting between two neutral atoms split into its constituent parts, an attractive force and a repulsive force.

distortion increases, eventually trying to push outer electrons into inner orbits (which are forbidden by the Pauli exclusion principle), and the repulsive force becomes so large it dominates all other forces acting, as shown in Figure 1.7. This repulsive force is generally modelled using an inverse power of the separation

$$F_{\text{repulse}}(r) = -\frac{A}{r^{n}} \quad ,$$

where r is the separation, A and n are constants; typically $n=11$.

The attractive force, known as the van der Waals attraction, is due to rapidly fluctuating dipole moments produced by the atoms. Although any atom is electrically neutral overall the random motion of electrons produces slight asymmetries in the charge distribution, leading to a small dipole moment. This dipole induces asymmetrical charge distributions in other atoms, leading to a dipole-dipole attraction. This attraction can be represented by

$$F_{\text{attract}}(r) = \frac{B}{r^{7}} \quad .$$

Integrating the total force gives the potential energy, the amount of energy needed to take an atom from infinity to a distance r from a neutral atom. In the case of perfectly neutral atoms, e.g. the inert gases, this potential is called the Leonard-Jones potential.

$$V(r) = \varepsilon\left[\left(\frac{a_o}{r}\right)^{12} - 2\left(\frac{a_o}{r}\right)^{6}\right] \quad ,$$

where r is the actual separation, a_0 is the equilibrium separation for the two atoms and ε is the binding energy at this separation.

1.5.3 Hydrogen Bonding.

If two atoms are covalently bonded then they share a pair of electrons. If the two atoms are identical then the electrons are shared equally and the bond is said to be nonpolar. However, if the bond is between different atoms then the electrons can be drawn closer to one atom than to the other. This ability to attract electrons is called electronegativity. This effect is strongest when hydrogen (which has a low electronegativity) bonds to a highly electronegative atom, such as oxygen or nitrogen. In such cases the hydrogen acquires a partial positive charge, balanced by a partial negative charge on the other atom. Partially charged hydrogens are electrostatically attracted to species with negative charges, such as the partial charge on atoms with lone pair electrons. This attraction leads to a hydrogen bond, which is relatively weak (typically around 5-20 kcal/mol *in vacuo* (Fersht, 1987)). Hydrogen bonds are particularly important in protein systems because N-H and O-H bonds are common, and hence many hydrogen bonds can be formed.

At the protein surface hydrogen bond donors and acceptors are readily satisfied by water, (or other aqueous solvents). In the protein core, however, the majority of the solvent has been forced out and such bonds are lost. It would be energetically unfavourable for these bonds to remain unsatisfied, and hence the matching of hydrogen bond donors and acceptors is a major driving force in the energetics of protein folding. Hydrogen bonding in globular proteins was extensively analysed and reviewed by Baker & Hubbard (1984).

1.5.4 Electrostatic Potential.

In the extreme case the electronegativity of two bonded atoms can be so large that the electrons involved in the bond become permanently associated with one of the atoms. The atom that gains the electron(s) becomes negatively charged (an anion) and the other atom gains an equal but opposite positive charge (a cation). The atoms are said to be ionically bonded. Ionic bonds result from the attraction of positive charges to negative charges. This attraction is not directional. Hence, unlike covalent bonds, ionic bonds have no orientational preference.

Many ions are important in biological systems, for example potassium (K^+), calcium (Ca^{2+}) and chloride (Cl^-) ions are often found bound to proteins. Such ions are often used to carry messages, such as nerve impulses, across membrane boundaries. When unbound such ions are surrounded by a cage of ordered water molecules which shield the free charge.

The electrostatic potential between two atoms with charges q_1 and q_2, separated by distance r, is given by

$$V_e(r) = -\frac{q_1 q_2}{4\pi \varepsilon_0 \varepsilon_r r} ,$$

with ε_0, the dielectric constant of free space, and ε_r the relative dielectric constant of the medium. ε_r represents the shielding effect of the medium in which the charges are embedded. This constant is 1 *in vacuo*, by definition, but the value in water, and other solutions, cannot be precisely defined. Generally the dielectric constant for a protein is estimated to be around 3, whereas that for water is 80. This large difference means that ionic interactions at the water surface contribute less to the protein stability than those buried in the interior. Fersht (1972) calculated the free energy of formation for a buried ion-pair to be roughly -0.8 kcal/mol whilst the corresponding figure for surface pairs is approximately -0.3 kcal/mol.

Barlow & Thornton (1983) found that approximately one third of charged residues in proteins are involved in salt bridges (i.e. ion pairs). Three quarters of these interactions stabilise tertiary structure, rather than secondary structure. In general, buried salt bridges are not conserved. In some proteins salt bridges have very specific functions to perform, and in such cases interactions are conserved.

1.5.5 The Hydrophobic Effect.

At room temperature there is a large entropic gain in burying non-polar residues, this is called the hydrophobic effect. The size of the effect shows a strong correlation with the non-polar surface area buried by the protein during the folding process. Chothia (1974) evaluated the contribution to be 24 cal/Å^2. This suggests that the hydrophobic effect is the dominant favourable force in the folding process (Dill, 1990).

When packed against non-polar residues the conformation of the solvent is restricted by the need to form hydrogen bonds. Burial of such non-polar residues means that hydrogen bonds can be formed with the polar atoms of the protein and hence a larger range of conformations can be adopted by the solvent. It is this increase in solvent entropy that drives a folding protein to bury its non-polar residues.

Many attempts have been made to measure or calculate the hydrophobicities of the amino acids (Chothia, 1976; Fauchere and Pliska, 1983; Janin, 1979; Miller, *et al.*, 1987b; Rose, *et al.*, 1985;

Wolfenden, *et al.*, 1981). Two categories of method have been used, experimental measurement and calculation from known structures.

In their experimental work, Wolfenden *et al.* (1981) measured the equilibrium distribution of amino acid side chains between their dilute aqueous solutions and the vapour phase. Measurements were taken over a range of solute concentrations and the results were corrected to pH 7. A broad range of free energies of transfer was obtained, ranging from +2.4 kcal/mol for glycine to -20.0 kcal/mol for arginine. No value was reported for proline, due to its unique main-chain/side-chain structure, and the value for arginine was derived in a different manner to the others, due to the low concentration of the vapour phase material. Another experimental method (Fauchere and Pliska, 1983) attempted to correct these problems by making measurements of a different system. Instead of measuring the concentrations of the solvent/vapour phase equilibrium; Fauchere & Pliska measured the concentrations in an octanol/water system. In this experiment, octanol represented the hydrophobic core of the protein, and hence free energies of transfer for each residue from the core to the solvent accessible surface were measured.

Theoretical calculations of a hydrophobicity scale involve examining the ratio of solvent-accessible residues to buried residues in known protein structures. This procedure has been carried out several times, with slightly different definitions of buried and increasing numbers of known structures. One of the most recent studies by Miller *et al.* (1987a) used 46 high quality crystal structures. They calculated the accessible surface area (ASA) for each one of these proteins and compared it with the ASA of the extended peptide, estimated from a calculated glycine-X-glycine tripeptide.

Those residues whose ASA was less than 5% of the extended ASA were deemed to be buried. This cutoff was chosen because virtually no residues were completely buried. The free energy of transfer from the protein core to the solvent could then be calculated for each residue type.

$$\Delta G_{transfer} = -RT\ln\left(\frac{N_{exposed}}{\sum N_{exposed}} \Big/ \frac{N_{buried}}{\sum N_{buried}} \right),$$

where T is the temperature (in Kelvin), R is the universal gas constant (2.02 cal/K/mol), $N_{exposed}$ is the number of solvent exposed residues of that type and N_{buried} is the number of core residues for that type. The ratio is normalised using the total number of exposed and the total number of buried residues, this corrects for the variation in the ratio of Nexposed and Nburied with protein size. This measure gives transfer free energies between -2.0 kcal/mol for lysine and 0.74 kcal/mol for isoleucine.

Despite the wide range of methods used to derive different hydrophobicity scales there is broad agreement between them. Miller *et al.* (1987a) measured correlation coefficients of 0.97 between their work and that of Janin (1979), 0.89 with that of Chothia (1976), 0.85 with that of Wolfenden *et al.* (1981) and 0.87 with the scale of Fauchere & Pliska (1983). These correlation coefficients measure only the agreement in relative positioning of the amino acid residues, the magnitudes of the scales vary between ranges of 20 kcal/mol for water/vapour transfers and a range of less than 3 kcal/mol for calculated surface/interior transfers. Comparison

Residue	Milleret *al.* (1987) [a]	Wolfenden *et al.*(1981)[b]	Janin (1979)[c]	Fauchere & Pliska (1983) [d]
Ala	0.20	1.94	0.3	0.42
Arg	-1.34	-19.92	-1.4	-1.37
Asn	-0.69	-9.68	-0.5	-0.82
Asp	-0.72	-10.95	-0.6	-1.05
Cys	0.67	-1.24	0.9	1.34
Gln	-0.74	-9.38	-0.7	-0.30
Glu	-1.09	-10.20	-0.7	-0.87
Gly	0.06	2.39	0.3	0.00
His	0.04	-10.27	-0.1	0.18
Ile	0.74	2.15	0.7	2.46
Leu	0.65	2.28	0.5	2.32
Lys	-2.00	-9.52	-1.8	-1.35
Met	0.71	-1.48	0.4	1.68
Phe	0.67	-0.76	0.5	2.44
Pro	-0.44		-0.3	0.98
Ser	-0.34	-5.06	-0.1	-0.05
Thr	-0.26	-4.88	-0.2	0.35
Trp	0.45	-5.88	0.3	3.07
Tyr	-0.22	-6.11	-0.4	1.31
Val	0.61	1.99	0.6	1.66

All values are quoted in kcal/mol.

Table 1.2

[a] Miller *et al.* calculated the surface/interior transfer free energy from crystallographic protein structures.

[b] Wolfenden *et al.* experimentally determined the vapour/water transfer free energy (hydration potential).

[c] Janin calculated the surface/interior transfer free energy from crystallographic protein structures.

[d] Fauchere & Pliska obtained a hydrophobicity constant for each side-chain, relative to glycine, in a octanol/water system. This hydrophobicity scale was converted into a free energy scale by Eisenberg & McLachlan (1986). The free energy figures are given here.

of the scales (Table 1.2) shows that, despite the broad agreement, some residues exhibit significantly different hydrophobicities in different methods. It is unclear whether these outliers are caused by experimental errors or whether the different systems used (water/vapour, water/octanol etc.) cause genuinely different hydrophobicities.

1.6. Molecular Recognition

The bonding forces and other effects described above apply equally well to both inter- and intra-protein interactions. The main aim of this thesis is to investigate and predict molecular recognition. Important aspects of three important protein/protein recognition systems will now be reviewed in detail, antibody/antigen recognition, proteinase/inhibitor recognition and helix/helix packing.

1.7. Antibody Systems

1.7.1 The Humoral Defence System.

Antibodies are a crucial part of the defence mechanism of the body. They exist in large numbers within the body and perform the vital task of recognising foreign molecules. The humoral defence mechanism is triggered by the synthesis of an antibody which binds specifically to a given invading molecule. The study of the humoral defence system has been greatly aided by the ability to produce large amounts of pure antibodies. It has long been known that a certain form of cancer, multiple myeloma, causes the uncontrolled proliferation of antibody producing cells. A single antibody producing cell can be cloned many times by this process and each resultant cell produces the same antibody. Thus tumours of this type can secrete large amounts of a single antibody. By fusing myeloma cells with a cell that is known to produce a particular antibody hybridoma cells are created. These cells, and their progeny, produce large amounts of the required antibody. These are called monoclonal antibodies because they originate from a single parent cell (Milstein, 1980).

1.7.2 Antibody Structure.

Antibodies are made up from four separate chains, two identical heavy chains, and two identical light chains. The light chains each have roughly 220 residues, and there are two basic types, a λ chain and a κ chain (Kabat, 1978). Different classes of antibody exist in the body, each with a particular role. Many exist in multimeric forms. The different antibody classes have heavy chains with different sizes or connectivities (Figure 1.8a).

Each antibody chain can be divided into repeats of the same general type of domain. This domain, called the immunoglobulin domain, is roughly 110 residues long. The light chain comprises two such domains, the heavy chain has either four or five domains in the monomeric form. There is a large amount of sequence homology between different types of domain (up to 40% identity), which suggest that there was once a single, ancestral immunoglobulin domain from which all the present day variations are descended.

Analysis of the variability of antibody sequences shows that three regions of unusually high sequence variation occur in the amino terminal domain of both the light and the heavy chain. These are called the hypervariable or complementarity determining regions (CDRs). The domains which contain these regions are called variable domains, whereas all the others are called constant domains.

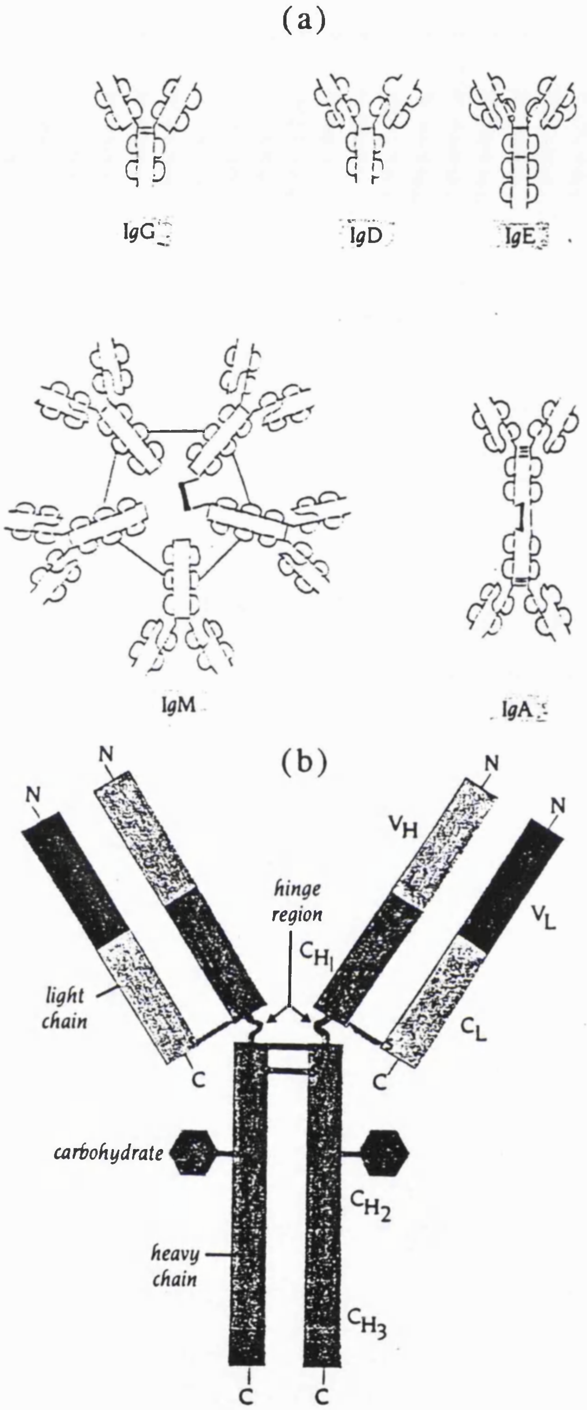


Figure 1.8

(a) Five classes of antibody structure. The classses are distinguished by the number of constituent immunoglobulins domains and by the disulphide linkage between the domains.

(b) The full structure of an IgG antibody. Each unique domain is labelled and disulphide linkages are shown.

Both diagrams from Branden and Tooze (1991)

Antibody molecules are made up from multiple repeats of the immunoglobulin domain. Antibodies have two identical antibody binding fragments (Fabs) connected, through the heavy chain, to a single constant fragment (Fc) (Figure 1.8b). The constant fragment is specific to the class of antibody being formed, and contains the characteristic number of immunoglobulin domains. Each Fab is constructed from four immunoglobulin domains. Two of these domains form the heavy chain, which is linked to the Fc fragment, and two form the light chain. The antibody binding region is formed by the end domains of the light and heavy chains. These domains contain the hypervariable regions and are called the variable light (V_L) and variable heavy domains (V_H).

The immunoglobulin domain is a barrel structure made by the packing of two β sheets (Figure 1.9). The constant domains have three strands in one sheet and four in the other. The variable domains have two extra strands to make a sheet of five strands.

The antigen binding domain is made up from six loops, three from the V_H domain and three from the V_L. These loops correspond to the complementarity determining regions of the sequence. For convenience these loops are termed L1,L2,L3 (for the light chain loops) and H1,H2,H3 (the heavy chain loops). Figure 1.10 illustrates this notation.

1.7.3 Antibody Diversity

Antibodies are able to bind to molecules as small as single metal ions and as large as viruses. They are able to recognise a particular antigen with amazing specificity. The key to this diversity is the ability of the body to produce millions of antibodies, each with a different binding region.

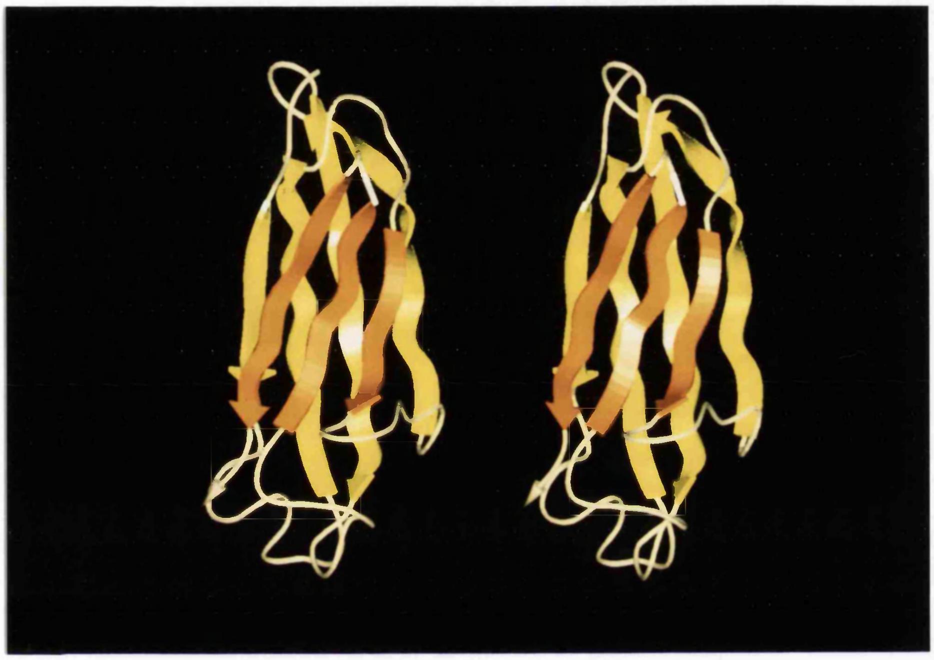


Figure 1.9

A stereo diagram of an immunoglobulin constant domain. Individual β-strands are shown as arrows. A four stranded β-sheet (yellow) is packed against a three stranded sheet (orange).

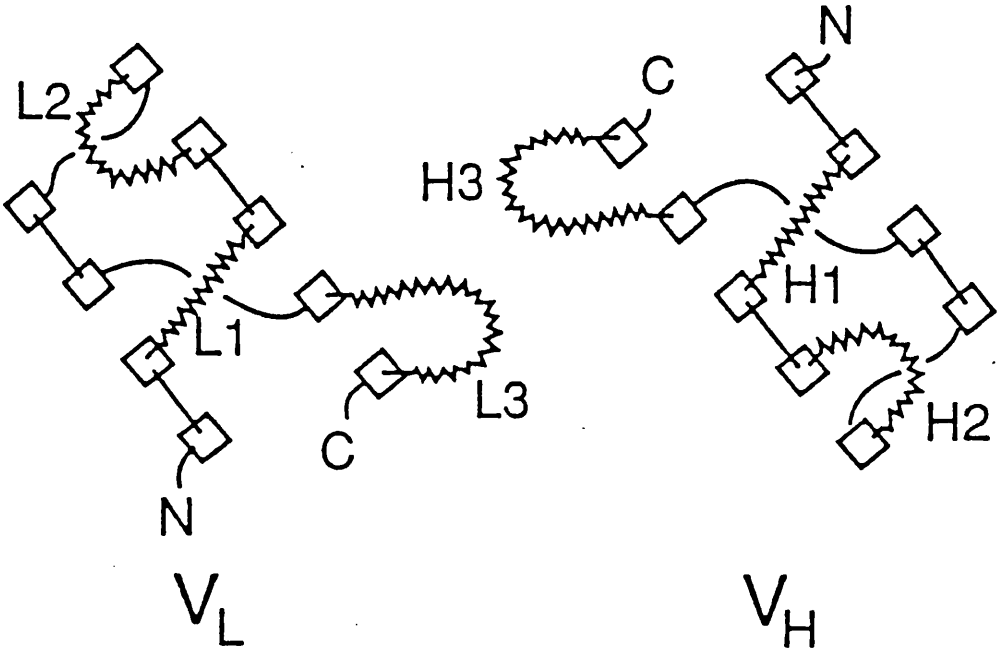


Figure 1.10

A plan view of the antibody binding region. Each of the hypervariable loops is shown as a sawtooth line. The sequence position of each of the loops corresponds well to the CDRs. Diagram from Tramontano & Lesk (1992).

The genetic code for antibodies is contained in over one thousand different sections of DNA, clustered into three gene pools (Leder, 1982). One pool codes for all the possible immunoglobulin heavy chains, one for the κ light chain and one for the λ light chain. Each gene pool is a library of possible fragments. The heavy chain library is split into three sections, approximately 1000 V sections, between 10 and 15 D sections and 4 J sections. The complete heavy chain is constructed by randomly choosing one DNA fragment from each section and joining them together to form a single, continuous exon. At the two junctions, V-D and D-J, considerable sequence diversity is added by errors in the joining process. It is these junction regions that contain the DNA code for the antibody binding loop, hence the loops have a much greater diversity than the rest of the chain. The two light chain libraries are split into two sections, V and J. A single light chain is manufactured by joining a V and a J fragment from either the κ or λ the library. There is no known distinction between the two light chain libraries, different species have differing proportions of κ chains to λ chains.

A complete antibody is constructed by joining the heavy and light chains together and adding the constant domains that are relevant to the antibody family being produced.

1.7.4 Antibody-Antigen Complexes

The structures of four antibody-antigen complexes have been solved crystallographically. These are three monoclonal-hen egg white lysozyme complexes: D1.3 (Amit *et al.*, 1986), HyHEL-5 (Sheriff *et al.*, 1987) and HyHEL-10 (Padlan *et al.*, 1989) and one for a monoclonal with neuraminadase (Colman *et al.*, 1987). These four complexes show that antibody/antigen recognition is broadly similar to other associations of folded protein chains, such as of proteinase/inhibitors and subunit/subunit. Analysis of the three antibody/lysozyme systems reveals several features. The monoclonals bind to different regions of lysozyme with only a slight overlap being observed between D1.3 and HyHEL-10 (Davies & Padlan, 1990). Figure 1.11 shows all three antibody molecules binding to the relevant sections of lysozyme. Despite the different locations, the interface regions are remarkably similar. The solvent accessible surface area lost during the docking process is approximately 1,500Å^2 in each complex (Davies & Padlan, 1990), roughly half from each species. Virtually all solvent is removed from this interface region: two water molecules remain in the HyHEL-5 structure, one in D1.3, and no water molecules are observed in the interface of the HyHEL-10 complex. This lack of water leads to a high degree of steric complementarity between the respective surfaces. HyHEL-5 is particularly rich in electrostatic complementarity with three salt bridges spanning the interface.

1.8. Antibody Modelling

Analysis of the known antibody sequences has shown that the hypervariable loops often have a similar size and that certain residues in the loops, and the surrounding framework, are strongly conserved (Kabat, 1978). These conserved residues are thought to be important in determining the conformation of the hypervariable regions (Padlan, 1977).

Chothia *et al.* (1986) predicted the structure of the antibody D1.3 from its sequence by comparison with the known crystallographic structures of five other antibodies. The framework

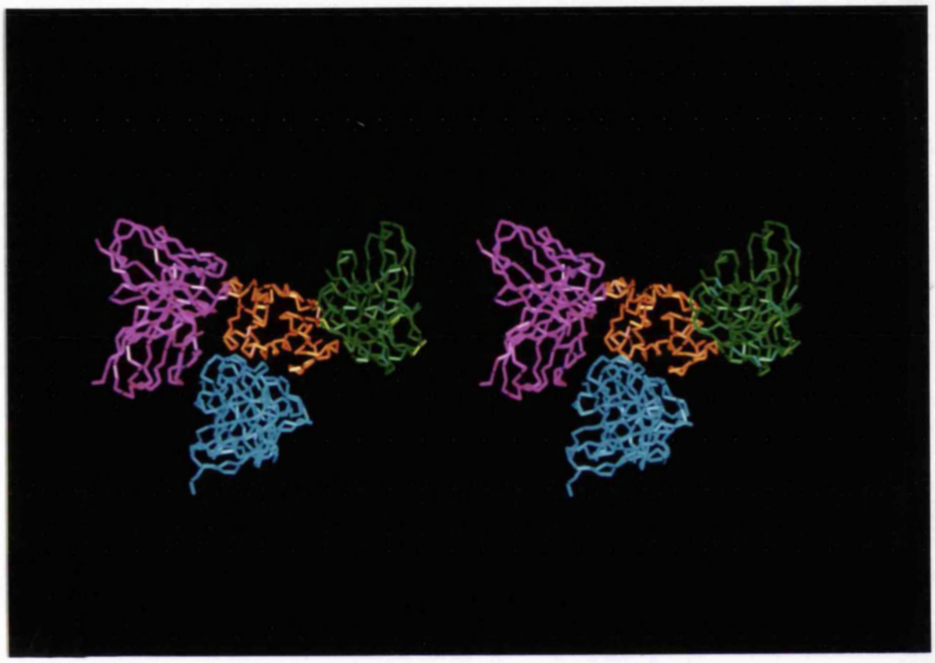


Figure 1.11

A stereo diagram of the Cα traces of the three crystallographically determined antibody/lysozyme complexes. The antibodies are shown bound to the correct section of lysozyme (shown in orange). HyHEL-10 is shown in cyan, HyHEL-5 in green and D1.3 in magenta.

regions of all five antibodies were in the same conformation, and D1.3 was assumed to follow this pattern also. They identified hypervariable residues of D1.3 that determined the loop conformation, either by an ability to adopt unusual conformations (such as glycine and proline) or by being in unique bonding or packing situations. Table 1.3 shows the homologies between D1.3 and the antibodies chosen as a template structure for each loop. The D1.3 structure was then predicted by putting each hypervariable loop into the conformation predicted by homology and fitting these loops onto a framework structure.

The structure of D1.3 was later solved crystallographically (Amit, *et al.*, 1986; Amit, *et al.*, 1985) and a comparison with the model was made by Chothia *et al.* (1986). The model was found to agree closely with the refined crystal structure, and proved to be more accurate than a preliminary crystallographic structure. The carbon alpha trace of the model differed from the crystallographic one by 1.5Å r.m.s. The side-chain positions showed a 2.2Å r.m.s. deviation overall (Figure 1.12).

Subsequent analysis revealed that five of the six loops apparently have a limited repertoire of possible main chain conformations (Chothia and Lesk, 1987), called canonical structures. Chothia *et al.* (1989) described the possibilities for loops L1,L2,L3,H1 and H2 (Figure 1.10). The loop H3 was found to have a much greater spread of sequence size and composition, and a limited canonical set could not be found.

Martin *et al.* (1989) combined both database search and *ab initio* calculation in a single algorithm. Instead of using a database of loops derived solely from antibody binding fragments they used loops from the entire protein databank. The algorithm also used the

Loop	Antibody	Sequence														
L1	D1.3	Ser	Gly	Asn	Ile	His	Asn							Tyr		
	REI	Ser	Gln	Asp	Ile	Ile	Lys							Tyr		
	McPC603	Ser	Gln	Ser	Leu	Leu	Asn	Ser	Gly	Asn	Gln	Lys	Asn	Phe		
L2	D1.3	Tyr	Thr	Thr												
	REI	Glu	Ala	Ser												
	McPC603	Gly	Ala	Ser												
L3	D1.3	Phe	Trp	Ser	Thr	Pro	Arg									
	REI	Tyr	Gln	Ser	Leu	Pro	Tyr									
	McPC603	Asp	His	Ser	Tyr	Pro	Leu									
H1	D1.3	Gly	Phe	Ser	Leu	Thr	Gly	Tyr								
	McPC603	Gly	Phe	Thr	Phe	Ser	Asp	Phe								
	NEW	Gly	Thr	Ser	Phe	Asp	Asp	Tyr								
	KOL	Gly	Phe	Ile	Phe	Ser	Ser	Tyr								
H2	D1.3	Gly	Asp			Gly										
	McPc603	Asn	Gly	Lys	Asn	Lys										
	NEW	Tyr	His			Gly										
	KOL	Asp	Asp		Gly	Ser										
H3	D1.3	Arg	Asp										Tyr	Arg	Leu	Asp
	McPC603	Tyr	Tyr	Gly	Ser							Thr	Trp	Tyr	Phe	Asp
	NEW	Leu	Ile	Ala								Gly	Cys	Ile	Asp	
	KOL	Gly	Gly	His	Gly	Phe	Cys	Ser	Ser	Ala	Ser	Cys	Phe	Gly	Pro	Asp

Table 1.3: The sequence homology between D1.3 antibody and each of the chosen template antibodies. Reproduced from Chothia *et al.* (1986)

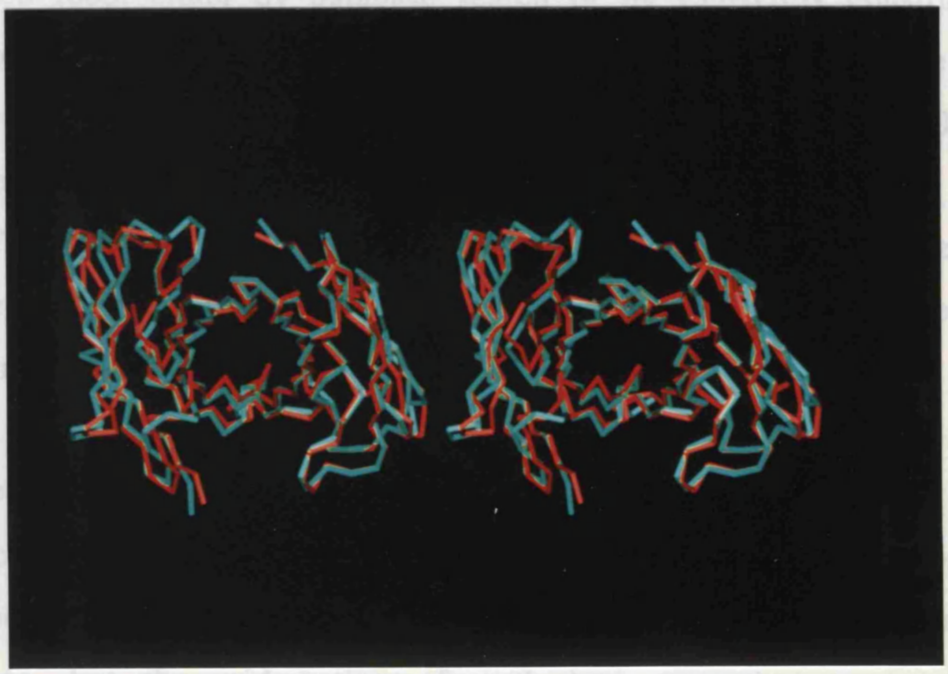


Figure 1.12

A stereo diagram comparing the antibody combining regions of the predicted D1.3 antibody (in red) and the crystallographically determined structure (in cyan). Cα traces are shown.

conformational search program CONGEN (Bruccoleri and Karplus, 1987) in conjunction with a modified version of the GROMOS (Åqvist, *et al.*, 1985) energy potential to construct some main-chain segements and to generate side-chain conformations.

The method relied on generating many different possibilities for each loop, either by database search or by CONGEN construction. The energy of each possibility was then calculated and the five most favourable conformations saved. The conformation with the lowest hydrophobic exposed area was then selected from these five.

The possible backbone conformations were generated in different ways, depending upon the sequence length of the loop. The conformation of loops of length less than six were generated solely using CONGEN. Conformations of loops of length six or seven were recalled from the loop database. The database was searched using distance constraints from the N and C terminal carbon alpha atoms of the loop. Loops of lengths over seven were also recalled from the database, but the mid-section of each loop was then reconstructed using CONGEN. Once possible backbone conformations had been found CONGEN was used to reconstruct side-chain orientations for each possibility, before energy calculations were performed.

This procedure was used to model the hypervariable loops of both HyHEL-5 and Gloop2. In both cases the algorithm produced accurate models. The HyHEL-5 model had a backbone r.m.s. deviation of 1.1Å from the crystallographic structure and an all atom r.m.s. of 1.9Å. The Gloop2 structure was compared with unpublished co-ordinates. A backbone r.m.s. deviation of 1.0Å and an all atom r.m.s. of 2.2Å were observed. The predictions of this method proved equally reliable for each of the six loops, including the H3 loop which Chothia *et al.* (1987) found hard to model.

1.9. Enzyme-Inhibitor Systems

One group of enzymes, the serine proteinases, has been widely studied and several crystallographic structures have been solved. Proteinases perform a wide variety of functions, from extracellular proteinases that bacteria use to break down surrounding proteins to uses such as food digestion, the immune response and blood clotting control in higher organisms. Although proteinase activity can be found in single domains of large proteins, the crystal structures solved have been of small, independent proteinases, often complexed to a relevant inhibitor. Proteinase inhibitors are used *in vivo* to regulate the activity of the proteinase. This is achieved by binding to the protease and blocking the active site, thus destroying the specificity to the substrate transition state. There are four different families of proteinases known. These are the serine, cysteine, aspartic and metallo proteinases, named by the most prominent functional group in the active site.

X-ray crystallography has been used to elucidate the structure of nine serine proteinase/inhibitor complexes. The structures of many uncomplexed proteinases and several uncomplexed inhibitors have also been determined. Many studies have been carried out on these structures (Hubbard, *et al.*, 1991; James, *et al.*, 1980; Laskowski and Kato, 1980) and much is known about the mechanism of these enzymes. The serine proteinases cleave a peptide bond of a polypeptide substrate, forming two individual, correctly terminated peptides by the addition of water. Different sub-families of serine proteinases cleave in different places. For example, trypsins cleave adjacent to a lysine or an arginine; chymotrypsins cleave adjacent to large hydrophobic residues and elastases cleave adjacent to small residues. The main functional section of all serine proteinases is the catalytic triad formed by a serine, a histidine and an aspartic acid residue. These residues are remote in sequence but come together in a specific geometry during folding. Trypsin and subtilisin have different three dimensional structures, and yet the catalytic triad still conforms to the same geometry suggesting that the same catalytic mechanism has evolved independently (convergent evolution). Like all enzymes, the serine proteinases bind to the transition state of the reaction they catalyze. The purpose of the catalytic triad is to encourage the formation of a short-lived tetrahedral reaction intermediate, which is stabilized by a covalent bond to the serine and a hydrogen bond to the oxy-anion hole, which is another conserved feature of the serine proteinase family. The specificity of a serine proteinase is governed by a series of loops adjacent to the catalytic triad. The three residues of the polypeptide substrate extending from the cleaved bond towards the N-terminus are conventionally named P1,P2 and P3, and those extending towards the C-terminus P1',P2' and P3'. The coresponding residues of the enzyme are named S1,S2,S3 and S1',S2',S3'.

1.10. α-Helix Packing

Several models of α-helix/α-helix packing have been proposed (Chothia, *et al.*, 1981; Crick, 1953; Efimov, 1979; Richmond and Richards, 1978). All these models use steric constraints to analyse possible modes of α-helix/α-helix packing. The most important parameter to be considered is the angle between the axes of the α-helices The distance of separation depends to a large extent on the exact side-chains present at the docking site.

Crick (1953) proposed the first simplified model of α-helix packing. Residues were represented by 'knobs' which formed regular clusters on the α-helix surface, leaving identifiable holes. Crick (1953) then used steric arguments to propose ways of fitting the knobs on one α-helix into the holes of the other, to form a stable association. In this way he predicted the main packing configuration to have a 20° angle between the two α-helices.

Chothia *et al.* (1977) improved upon the 'knobs into holes' model of Crick (1953) by describing the α-helix surface in terms of ridges and grooves. Figure 1.13a shows an α-helix which has been slit parallel to its axis and flattened out. Each residue is represented by a single sphere of fixed radius. This representation shows the ridges and grooves of an α-helix surface. Taking i to be the central residue and n to be an integer, the ridges are made up from the residues $i \pm 4n$, $i \pm 3n$ and $i \pm n$ (Figure 1.13b), these ridges will simply be called the 4 ridge, the 3 ridge and the 1 ridge respectively. For each of the ridges there is a corresponding groove. The majority of observed α-helix/α-helix packings correspond to a ridge from one α-helix packing into a groove from the other α-helix. Since α-helix/α-helix docking is symmetrical (ignoring differences in side-chain size), this implies that the corresponding groove from the first α-helix is filled by the corresponding ridge of the second α–helix. Figure 1.13c shows the relative positions of the α-helix residues, looking down the helix axis. For a packing that centres upon residue i the 4 ridge is the most important, since it is the most prominent on the α-helix surface. The 3 ridge is less prominent and the 1 ridge is the least important, sterically. These relative importances are borne out in observed structures. Examining 50 helix/helix packings, Chothia *et al.* (1981) found that 25 corresponded to the 4 ridge of one α-helix packing into

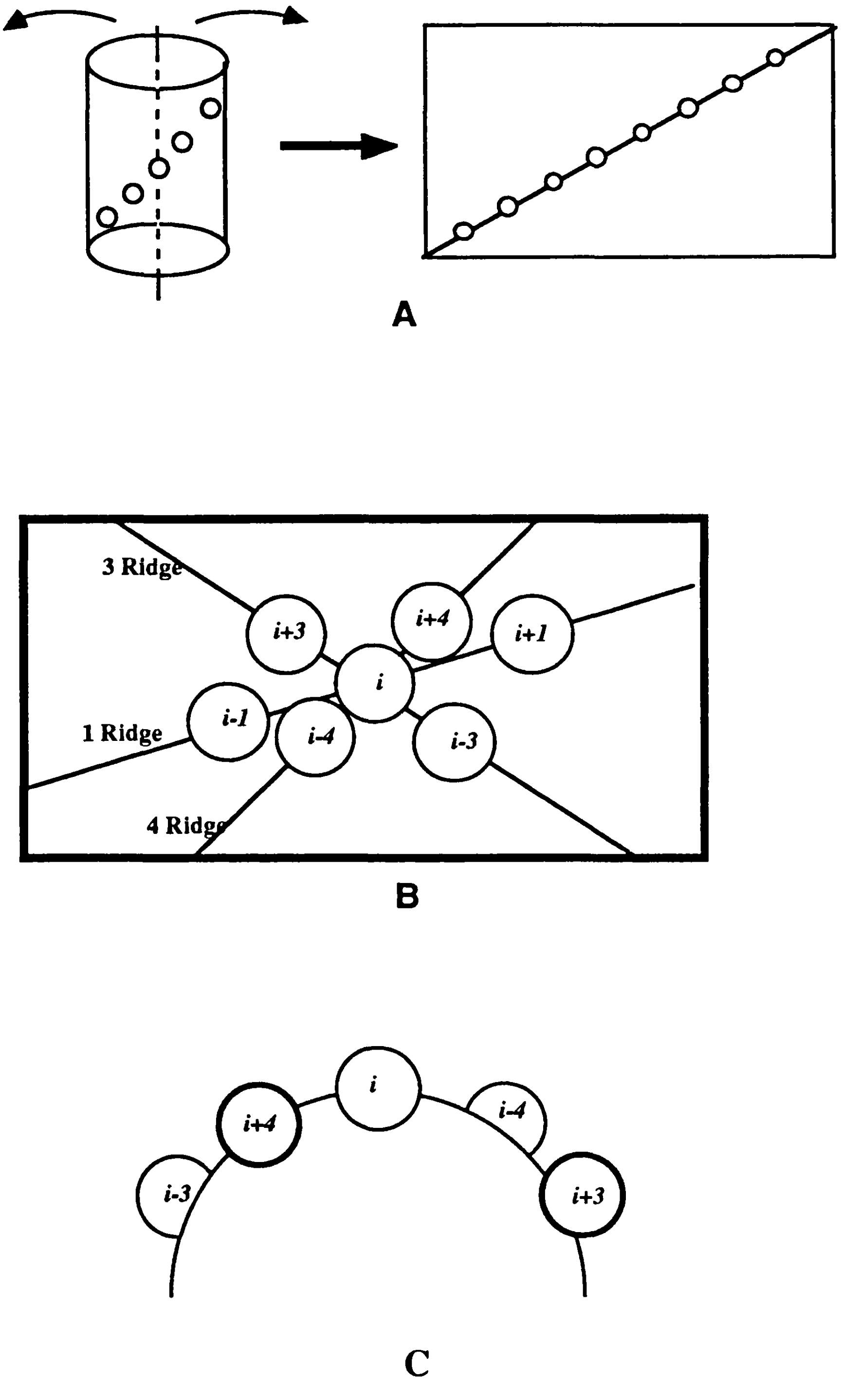


Figure 1.13

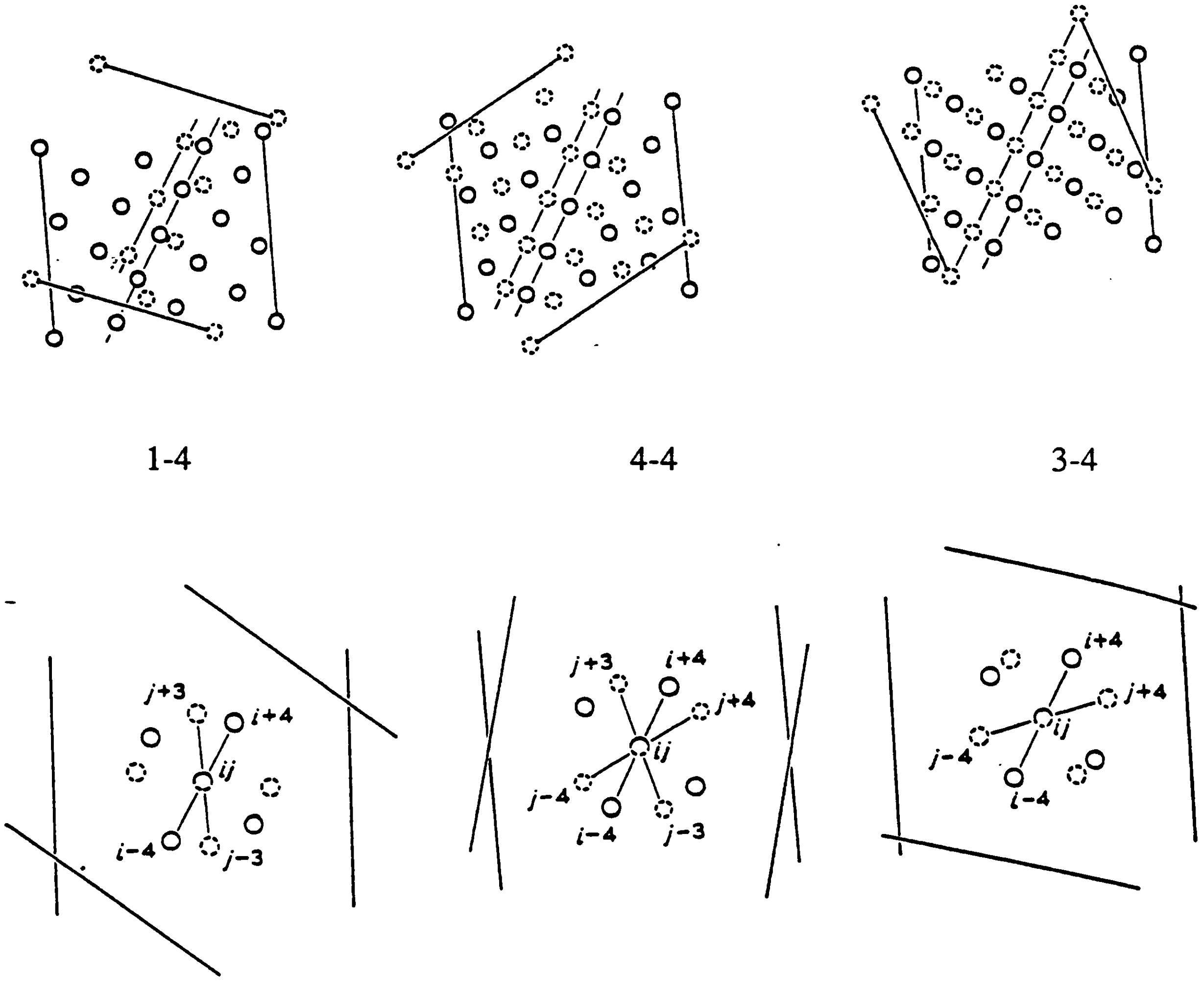
a. The flattened helix representation.

b. The ridges present on a α-helix. The central residue is labelled *i*, other residues are numbered relative to this residue.

c. The relative importance of helix ridges. The i±4 ridge is most prominent, followed by the i±3 ridge. (Adapted from Chothia, 1984) the 4 groove of the other. This was termed a 4-4 packing. Eight 3-4 packings, with the 4 ridges packing into the 3 grooves, were observed, along with four 1-4 packings and one 3-3 packing. Figure 1.14 illustrates all these packings using a 'flattened α-helix' model. The angle between the two α-helices can be calculated for each mode of association, using simple geometry and assuming regular α-helices. These angles, shown in Table 1.4, proved to be sensitive to changes in the pitch of the α-helix and the observed angles varied from those quoted in Table 1.4 (which were calculated assuming standard α-helix geometry) by as much as 15°.

The central residues of an α-helix/α-helix packing are important. In particular, very small residues in these positions can break up a ridge. This discontinuity makes it possible for a ridge from one α-helix to cross-over one from another α-helix. Chothia *et al.* (1981) found seven α-helix/α-helix packings where the 4 ridges of the α-helices crossed over each other (4x4 packing). Three 3x4 packings were also found. The central residue in each case was either alanine, glycine or serine.

Richmond & Richards (1978) examined known α-helix/α-helix packings from the proteins deoxymyoglobin, carp muscle parvalbumin, thermolysin and concanavalin. They measured solvent exclusion; bringing together the two α-helices along the contacting normal. Surface area began to be excluded when the α-helices were still 6Å from their final position. In myoglobin between 120 and 260 $Å^2$ of contact surface area is excluded from the solvent for each pair of interacting helices, this corresponds to a hydrophobic effect of between twenty and sixty kcal (Chothia, 1974). The α-helix/α-helix packing angles were found to be in agreement with those predicted by Chothia *et al.* (1977).



1-4 4-4 3-4

3x4 3x4 and 4x4 4x4

Figure 1.14

Six common helix/helix packings, shown in the flattened helix representation. Solid spheres represent residues of the lower (face up) helix, dashed spheres represent residues of the upper (face down) helix. The top row shows three ridges-into-grooves packings, the bottom row shows three crossed ridge packings. Taken from Chothia *et al.*,1981

Classification	Packing Angle
1-4	-105°
3-4	23°
4-4	-52°
3-3	-117°
3x3	+55°
3x3 & 4x4	-15°
4x4	-105°

Table 1.4

The ideal packing angle for each of the commonly observed helix/helix packings. (data from Chothia *et al.*, 1981).

1.11. The Generic Docking Problem

Given the structures of two molecules which are known to bind to each other, it is possible to search for plausible modes of association. The aim of the docking problem is to discriminate between the correct, biologically observed orientation and other reasonable orientations. The docking problem for two molecules involves six degrees of freedom (Figure 1.15). In Figure 1.15 and, where appropriate, throughout this thesis the larger of the two molecules is called the host molecule and the smaller the guest. To find the best way for two molecules to interact an algorithm must have a strategy for exploring these degrees of freedom and a discrimination function that evaluates how favourable a particular docking orientation is.

The most direct strategy for exploring the degrees of freedom is to evaluate the discrimination function for as many, regularly spaced docking orientations as possible. This is called a global search. Other methods, such as simulated annealing (Kirkpatrick, *et al.*, 1983) start with some random orientation and, by making small changes to the orientation, attempt to minimise the discrimination function. This is equivalent to following a path in the six dimensional search space. The advantage of such methods is that they evaluate fewer orientations, and hence are faster. A global search, however, is guaranteed to find the global minimum of the discrimination function, whilst other search methods can get 'trapped' in local minima. The accuracy and specificity of a global search requires an adequate coverage of search space and a good discrimination function.

One of the greatest difficulties faced by any docking algorithm is allowing for, or explicitly treating, induced fit. Two rigid bodies

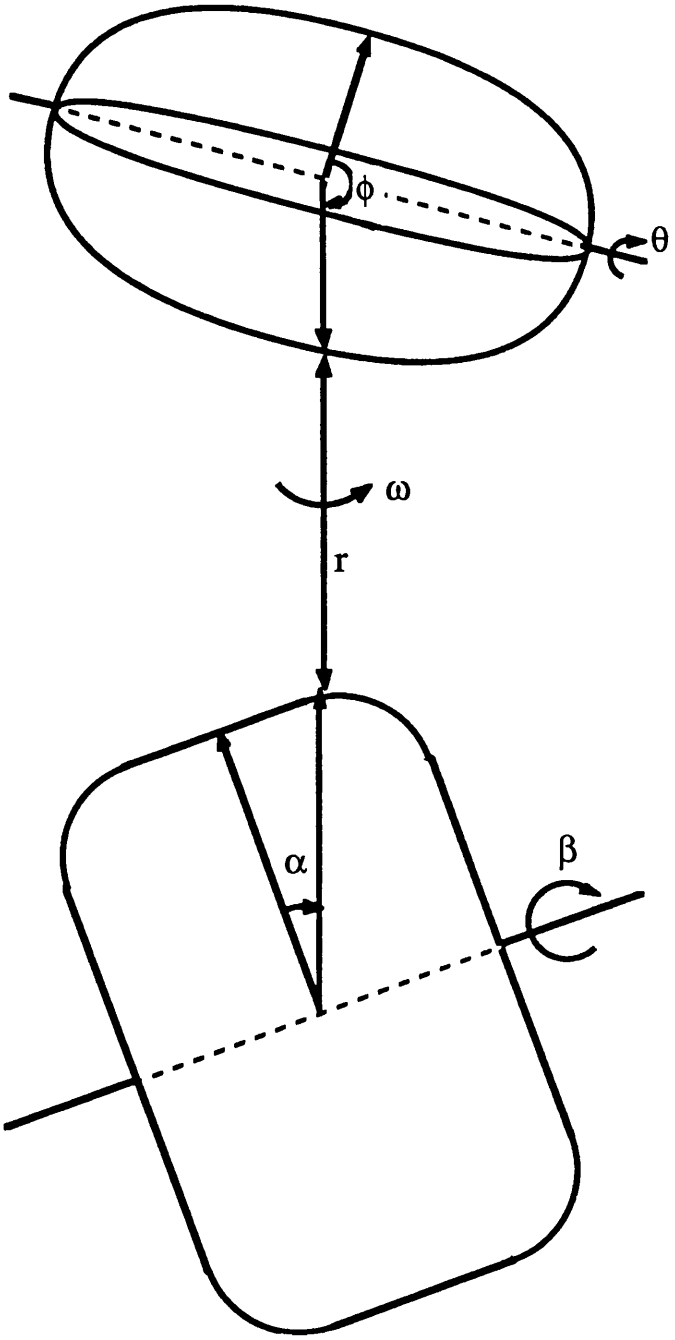


Figure 1.15

The degrees of freedom examined during the docking process. The guest molecule is shown uppermost with the host molecule below.. The rotational degrees of freedom θ and ϕ bring any section of the guest molecule into this docking area. The angles α and β are the equivalents for the host molecule. The perpendicular distance between the two centroids, r, can then be varied. A rotation ω perpendicular to line connecting the centroids completes the six degrees of freedom.

have six degrees of freedom during the docking process. In reality proteins, however, have many hundreds of degrees of freedom. These extra ones are due to movements of surface side-chains, the main-chain and shifts in the tertiary and quaternary structure of the protein. Whilst docking the side-chains of the two proteins begin to interact. This can lead to shifts in position which enhance the final complementarity between the interacting surfaces. A dichotomy exists between pure induced-fit docking, which require no initial steric or charge complementarity, and rigid-body models which allow no side-chain movements. Many observed protein docking appear to correspond much more closely to rigid body docking than to pure induced fit. Evidence for this comes from examination of the structure of hen egg white lysozyme solved in its free form and when complexed to various antibodies.

1.12. Previous Docking Methods

Previously many different approaches have been tried to solve the molecular docking problem. Several methods that dock small ligands to proteins have been reported (e.g. Goodsell and Olson, 1990; Kuntz, *et al.*, 1982), all of which use rigid body methods to achieve native-like conformations. It is, however, not clear how these algorithms would perform on larger systems, and the extent to which conformational changes can be included.

Protein/protein docking has also been explored (e.g. Goodsell and Olson, 1990; Wodak and Janin, 1978; Yue, 1990). Zielenkiewicz & Rabczenko (1984) examined the auto-association of insulin by mapping certain surface properties, such as hydrophilic/hydrophobic behaviour and potential hydrogen bonds, onto a two dimensional grid. They then examined such surfaces from the dimer forming region and then from the hexamer forming region. By varying the Euler angles of the system, Zielenkiewicz & Rabczenko (1984) showed that the known modes of association corresponded to orientations that produced large numbers of coincident properties at the surfaces. However, due to lack of computer time, they were unable to perform a global search and could not take account of any steric forces.

Warwicker (1989) used a finite difference potential to explore globally six dimensional space. This process took up to 100 hours on a VAX 8800 to produce estimations of binding energy for many millions of different orientations. The systems HyHEL-5/lysozyme, cytochrome c peroxidase/cytochrome c, and trypsin/bovine pancreatic trypsin inhibitor (BPTI) were examined. In the cases of HyHEL-5/lysozyme and trypsin/BPTI, the reduced stereochemical and electrostatic model identified the native conformation near a local minimum. However, when energy surfaces were plotted for a particular surface of one molecule whilst rotating the second, it could be seen that other minima were present.

Jiang & Kim (1991) used a surface cube representation of the protein and, by using a soft docking potential, implicitly allowed for side-chain movement. This method was applied to binding of small molecule (NADPH and methotrexate) to dihydrofolate reductase. It was found that the top ranked solution varied from the native conformation by only a two or three degrees of rotation in each axis. When applied to larger systems such as the binding of lysozyme to the variable fragment of HyHEL-5 antibody, near native solutions were eighth and twelfth in the list. The best solution (the twelfth) differed from the native by Euler rotations of (7°, 4°, 7°). No simulations are reported for either the HyHEL-10 or D1.3 complexes.

Cherfils *et al.* (1991) used a simulated annealing approach with a simplified protein model to explore the docking of antibody-antigen and proteinase-inhibitor systems. Each residue was represented by a single sphere and a simplified potential used to evaluate the energy of the complex. The method began with a randomly chosen set of docking parameters. These parameters were then randomly perturbed and the energy difference calculated. The perturbation was accepted or discarded according to a probability formed from this energy difference and the 'temperature' of the system. Continual perturbations were made and the 'temperature' slowly decreased, thus freezing the complex in an energy minimum.

Using the relevant complexed form of lysozyme Cherfils *et al.* (1991) found near-native conformations for both HyHEL-5 and HyHEL-10, but not for D1.3. These orientations occurred first and third (respectively) in a list of fewer than ten possibilities. Using the uncomplexed form of lysozyme a native solution for HyHEL-5 is again found to be at the global minimum. However no such solution was found for either HyHEL-10 or D1.3.

Shoichet & Kuntz (1991) used an extension of earlier small molecule work, (Kuntz, *et al.*, 1982), to examine proteinases/inhibitor docking. The surface of each molecule was described by a set of spheres of varying sizes. For the receptor molecule (the proteinase), these spheres filled the empty volume around the receptor site, generating a negative image of the site. For the ligand (the inhibitor), the spheres were placed within the molecular surface, hence a positive representation of the ligand was produced. These sphere sets were then used to search for docking orientations that did not overlap the two molecules but allowed good contact. Orientations that satisfied these criteria were then evaluated by more rigorous methods, using an all atom representation. The methods used included electrostatics, buried surface area, free energies of solvation and molecular mechanics. In particular minimisation was carried out using the AMBER package (Weiner, *et al.*, 1984). The method was applied to three proteinase-inhibitor systems, using both bound and unbound forms of the structures. In every system conformations were found with r.m.s. deviations from the crystallographic structure of less than 1Å. Also in every system, the lowest energy conformation was found to be within 5Å r.m.s deviation of the crystallographic structure. Shoichet & Kuntz (1991) report that a number of false positive structures were found. These structures had energies similar to the native-like solution but did not bind in the same manner, often involving completely different sections of the protein surface. More false positive structures were generated when attempting to dock the unbound crystal structures than when using the bound forms.

Graph theory (Deo, 1974) has also been used to explore the docking problem. Kasinos *et al.* (1992) constructed graphs which described the surface atoms of the two proteins to be docked. For the larger (host) molecule the graph described the distances between polar surface atoms. For the smaller (guest) molecule Kasinos *et al.* (1992) constructed a set of points which defined possible interaction sites for each atom, based upon the binding properties of each atom type. Kasinos *et al.* (1992) then used a variation (Subbarao and Haneef, 1991) on the Ullman subgraph isomorphism algorithm (Ullman, 1976). This algorithm finds the largest, common section between two graphs. In this application the maximum subgraph represents the best interactions between the two molecules. To allow for small inconsistencies between the two graphs the distances matching was carried out with a ±3Å tolerance.

This method was applied to several systems. In all but one case, a single mode of interaction was produced. The complex between D1.3 and bound lysozyme was reconstructed with a 3.1Å r.m.s. deviation from the crystal structure. The complex between trypsinogen and the bound form of BPTI was reconstructed to 1.6Å r.m.s. and the interaction between the two chains of insulin was predicted with 0.2Å r.m.s. deviation. In predicting the dimerisation of insulin the method produced two possible modes of binding, one of which was correct to 2.1Å r.m.s. and one of which was eliminated by the authors due to poor fitting between the monomers. The method was also applied to several small molecule/protein and protein/DNA systems with equal success.

1.12.1 Summary

Table 1.5 summarises the approach of four methods which were applied to either proteinase/inhibitor or antibody/lysozyme systems. Table 1.6 shows the results obtained by three of these methods when applied to the antibody/lysozyme docking problem. It is difficult to compare the accuracy of these methods since different measures are used by the authors themselves to estimate their success, and different authors applied their algorithms to different systems. Despite this it is clear that each of the algorithms produces a solution with about 3Å r.m.s. deviation of the known structure for at least one antibody complex. Only Cherfils *et al.* (1991) apply their algorithm to all three crystallographic antibody/lysozyme systems, finding solutions in all but one case. No attempt has been made to

Authors	Method	Coverage	Comments
Kasinos *et al.*	Graph Theory	All of guest part of host	This method generally produces a single orientation. Considerable success is reported for reconstructing known complexes. No uncomplexed protein / protein problems are tackled. All of the guest molecule was searched for similarity to a directed graph of the binding site of the host.
Cherfils *et al.*	Simulated Annealing	All of guest part of host	After simulated annealing the best structures were refined to improve accuracy. Only solutions which involved an area within 30° of the host binding site were allowed.
Jiang & Kim	Soft docking of surface cubes	All of guest All of host	The docking used included both steric and electrostatic terms. The search was carried out in 2 stages, a full search using a coarse grid and a more specific search using a high density grid. By altering the size of the grid small molecule / small molecule bindings were also examined.
Shiocet & Kuntz	Steric match of reduced surface representation	All of guest All of host	The algorithm used always reported an orientation extremely close to known complex, however no figures for the number of possibilties or the relative positioning of the native-like solution were given. The authors explore the utility of a number of filters to reduce the number of alternatives. These filters include buried surface area, detailed energy calculations and solvation free energy. None of these were found to be entirely satisfactory.

Table 1.5

A summary of key facts about 4 recent docking methods, including a brief outline of the method, the scope of the search and the systems to which the method was applied.

	HyHEL-5(c)
	&
	Lysozyme(c)
		HyHEL-5(c)
		&
		Lysozyme(f)
			D1.3(c)
			&
			Lysozyme(c)
				HyHEL-10(c)
				&
				Lysozyme(c)

Kasinos *et al*			3.1Å r.m.s 1/1	
Cherfils *et al*	Native 1/7	Native 1/6	No solution found	Native 3/9
Jiang & Kim		(7°,-4°,7°) (0,0,0)Å 12/15		

Table 1.6

The success of three algorithms that have been applied to antibody/lysozyme systems. Each entry consists of a measure of accuracy and a measure of specificity. Kasinos quote their accuracy as an r.m.s. deviation of all atoms; Jiang & Kim give a set of Euler rotations (°) and a translation vector (Å); Cherfils *et al.* state only that the solution is native-like, which they define as within 7° of the orientation of the X-ray structure. The specificity measure is the number of possible orientations resulting from the procedure (second number) and the rank of the native-like solution within this list (first number), e.g. Jiang & Kim find their most native-like structure ranked twelfth in an ordered list of 15 possible orientations.

The structure of unbound lysozyme has been determined, as well as bound to each of three antibodies. Docking algorithms either use the appropriate complexed form of lysozyme, termed lysozyme(c), or the free form, lysozyme(f).

dock the uncomplexed form of lysozyme to either the D1.3 or HyHEL-10 antibody.

Several authors have applied their algorithms to proteinase/inhibitor systems (Table 1.7). Despite the large number of solved structures of this type most authors tackle the trypsin/BPTI system, since each molecule of this system has been solved in the uncomplexed form, as well as the complexed form. Cherfils *et al.* (1991) do not find any solutions for this system, despite their success with the antibody/lysozyme systems. Shiocet & Kuntz (1991) always find a solution within 0.8Å r.m.s. deviation of the known orientation, but do not quote how many solutions are found, nor the position of the best structure within this list.

	Trypsin(c)
	BPTI(c)
		Trypsin(c)
		BPTI(f)
			Trypsin(f)
			BPTI(f)
				Chymotrypsin(c)
				Ovomucoid(c)
					Chymotrypsin(f)
					Ovomucoid(f)
						Trypsinogen(c)
						BPTI(c)

Kasinos *et al.*						1.6Å r.m.s. 1/1
Cherfils *et al.*	No solution found	No solution found				
Jiang & Kim	(-20°,4°,3°) (0,0,0)Å 1/2		(2°,10°,2°) (2,2,4)Å 6/9			
Shiocet & Kuntz	0.3Å r.m.s		0.5Å r.m.s	0.7Å r.m.s	0.8Å r.m.s.	

Table 1.7

The success of four algorithms that have been applied to enzyme/inhibtor systems. As in Table 1.6 rach entry gives a measure of accuracy, when available and a measure of specificity. A wide variety of enzyme/inhibitor structures are known, often the structures of the uncomplexed components have also been solved. Each molecule is denoted (c), indicating that the complexed form was used in the docking, or (f), indicating the use of the unbound (free) form. Accuracy is given as either a r.m.s. deviation or as a set of translations and rotations. The specificity measure is given in the form; rank of native-like structure / number of possible structures. Shiocet & Kuntz give no indication of the number of orientations they find, nor the position of the native-like structure within this list.

1.13. Overview of this thesis.

This thesis details the design, implementation, testing and applications of a novel protein/protein docking algorithm. This algorithm is based on the steric matching of complementary surfaces, with electrostatic interactions used as a subsequent filter. Explicit information about which residues are involved in the association is incorporated into the algorithm, and the relevance of this is discussed. The algorithm is implemented on a parallel architecture computer, which enables a large number of possible configurations to be assessed in a reasonable period of time.

A key feature of the algorithm is that it is not critically dependent on precise atomic positions. This allows a predicted antibody structure to be used as a docking target for the first time. The level of accuracy of a model can never be as high as for a well resolved, refined crystallographic structure and hence a soft potential method is crucial to success. The algorithm is also used to investigate α-helix/α-helix packing modes, and used in a modelling situation to predict the binding of epidermal growth factor (EGF) and a model of epidermal growth factor binding protein (EGFBP).

Chapter 2 of the thesis describes the algorithm in detail, and its implementation. Chapter 3 reports and evaluates the results obtained for several antibody/antigen systems, including the docking of hen egg white lysozyme with a model of the D1.3 antibody. Chapter 4 describes the application of the algorithm to the EGF/EGFBP system and Chapter 5 discusses the use of the algorithm in investigating the association of α-helices. Finally, Chapter 6 discusses the general conclusions derived from this study and suggests possible lines for further research.

Chapter 2

Method

2.1. Synopsis

This chapter describes the design and implementation of a novel docking algorithm. The main purpose of this algorithm is to enable the use of uncomplexed and modelled protein structures in the docking process. This is achieved by using a soft, steric matching process. A full description of the aims of the algorithm is given and the relevance of these aims to the field of protein modelling are discussed. The use of purely steric matching is justified by an examination of interface complementarity for all three known antibody/lysozyme crystal structures. A detailed, step-by-step description of the algorithm used is then given. The algorithm was tested at several stages. The results of these test are discussed in the final section of the chapter, and the subsequent changes made to the algorithm are highlighted. Later chapters discuss the application of the algorithm to several systems of interest.

2.2. Aims of the Method

To be of practical use, any docking algorithm must be able to accommodate structural changes caused by induced fit. Algorithms which can only reconstruct complexes that have previously been determined experimentally are only useful for investigating the docking process. Explicit inclusion of induced fitting seems impossible with present day computing power, as it would require a search of the many hundreds of degrees of freedom due to side-chain rotation, main-chain distortion and bulk changes to the tertiary and quaternary structure. An implicit inclusion therefore seems necessary. Such a method does not try to predict side-chain movement, but instead allows a certain degree of steric and electrostatic mismatch between the host and guest molecules. The level of mismatch allowed must be carefully considered, too much will destroy the specificity of the algorithm, too little may mean the true solution is missed.

The number of experimentally determined protein structures is still quite low. Protein modelling attempts to overcome this limitation by constructing plausible structures for a protein from the structures of homologous proteins. This approach has proved effective for a wide variety of proteins, including immunoglobulins and serine proteinases. Inevitably, the models proposed are not as accurate as well determined experimental structures. Despite this inaccuracy models are often useful in a wide variety of applications, such as suggesting mutations to make the protein more active, less active or to change stability. Models are often used to provide information on the active site of a protein, proposing possible mechanisms of action and thereby suggesting possible ligands. There is, therefore, a need for a docking method which can use models as a starting point. The algorithm presented in this thesis takes soft docking a stage further and attempts to predict possible modes of binding for a mixture of complexed, uncomplexed and modelled structures.

2.3. Evaluation of Interface Complementarity

A major component of the docking algorithm to be presented will be the use of a soft potential to search for areas of steric complementarity between the molecular surfaces of host and guest proteins. This search will only be effective if the lock-and-key binding model is largely correct. The level of complementarity at protein/protein interfaces can be judged by examining the three known antibody/lysozyme complexes (Table 2.1).

The nature of the antibody/antigen interface has been described (Mian, *et al.*, 1991; Padlan, 1990; William, *et al.*, 1990). In order to test further the steric complementarity in this region a volume calculation was carried out on each of the three antibody/lysozyme complexes. The method used was that of Gellatly & Finney (1982). A hypothetical solvent shell was placed around the complex. The shell completely surrounded the complex but did not seep into the antibody/lysozyme interface. The volume inside this shell was then divided between the atoms in the complex, in proportion to their van der Waals radii. This involved finding nearest neighbours to each atom and placing a plane between them, perpendicular to the line passing directly though the atom centres and with distances between the plane and the atoms in proportion to their van der Waals radii. Many such planes were created around each atom until a closed polyhedron was formed. The volume of this polyhedron was taken to be the volume occupied by the atom.

Structure	Resolution (Å)	PDB file	Author
HyHEL-10	3.0	PDB3HFM	Padlan *et al.* (1989)
HyHEL-5	2.5	PDB2HFL	Sheriff *et al.* (1987)
D1.3	2.8	obtained from Dr S. Phillips	Amit *et al.* (1986)
Lysozyme	2.0	PDB6LYZ	Diamond *et al.* (1974)

Table 2.1

Protein structure data for the three antibody/lysozyme complexes and for the unbound form of lysozyme. The structures have resolutions between 2Å and 3Å, sufficient to correctly assign main-chain and side-chain structure. Three of the structures are deposited in the Brookhaven Protein DataBank (Bernstein *et al.*, 1977), entry codes are given.

Once volumes have been assigned to each atom, it was possible to find the mean volume occupied by each atom type. For each atom type an expected volume was calculated by averaging the volume results for several proteins, not including the antibody/lysozyme complexes. Examining the differences between the volumes seen at the interface and those seen on average allowed an estimation of the packing density relative to the protein core

The results of the volume calculations are summarised in Table 2.2. A volume calculation was performed for each atom within the interface ($V_{observed}$). The volume calculation was also carried out on unrelated proteins, and an average volume occupied for each atom type, was calculated (V_{expect}). This predicted volume was then compared to the expected volume for an atom of that type. Taking a ratio, for each atom, of calculated volume to expected average volume gives a packing density.

$$\begin{array}{l}\text{packing density}\\ \text{for atom } i \text{ of type } t\end{array} = \frac{V^{i}_{\text{observed}}}{V^{t}_{\text{expect}}}$$

This number is 1.0 for an atom that occupies the standard amount of volume. Atoms that are too tightly packed occupy too little volume and therefore have a packing density less then 1.0, whereas atoms which are too loosely packed have a packing density greater than 1.0. The packing densities shown are averages of a group of atoms. The global packing density is the average over all atoms, the interface packing density is the average for the interface atoms. The standard deviation of the global packing densities is also shown. The differences between the global and interface packing densities are very much smaller than the standard deviation for all structures.

Complex	Global Packing Density	Interface Packing Density	Standard Deviation
HyHEL-10	1.01	1.00	0.16
HyHEL-5	0.99	1.02	0.18
D1.3	1.01	1.00	0.18

Table 2.2

The results of the volume calculations on the antibody/lysozyme complexes. A global packing density, an interface packing density and the global standard deviation of packing density is given for each of the complexes. The packing densities are measured relative to a set of sixty eight well resolved protein structures.

This means that the interface region is as tightly packed as the protein core. Thus, the interface region must exhibit the same degree of steric complementarity as the protein core. These results are in agreement with those of Chothia & Janin (1975) who examined protein/protein interface regions and found them to be close packed.

2.4. Overview of the Method

Figure 2.1 is a flow chart of the approach. The entire surface of the guest molecule and the binding region of the host are represented as a series of slices of the protein surface. Each slice details the van der Waals surface of the protein over a 32Å x 32Å region. The surfaces of the proteins are smoothed and represented as a contour map. A soft potential, based loosely on the Leonard-Jones potential, is used to check each possible pair of maps for surface complementarity. This comparison process is computationally intensive and so a parallel architecture computer is used. The matching process produces a large number of possible orientations, which must be clustered to produce a manageable dataset. After clustering several hundred docking orientations remain. Each one of these orientations represents a sterically plausible way of docking the host and guest molecules which is significantly different from the other orientations in the set. To reduce further the number of docking orientations constraints are applied. The main docking procedure concentrates solely upon steric information and so the constraints used are based on other information that may be available. These are electrostatic complementarity, epitope information, and a single imprecise distance constraint.

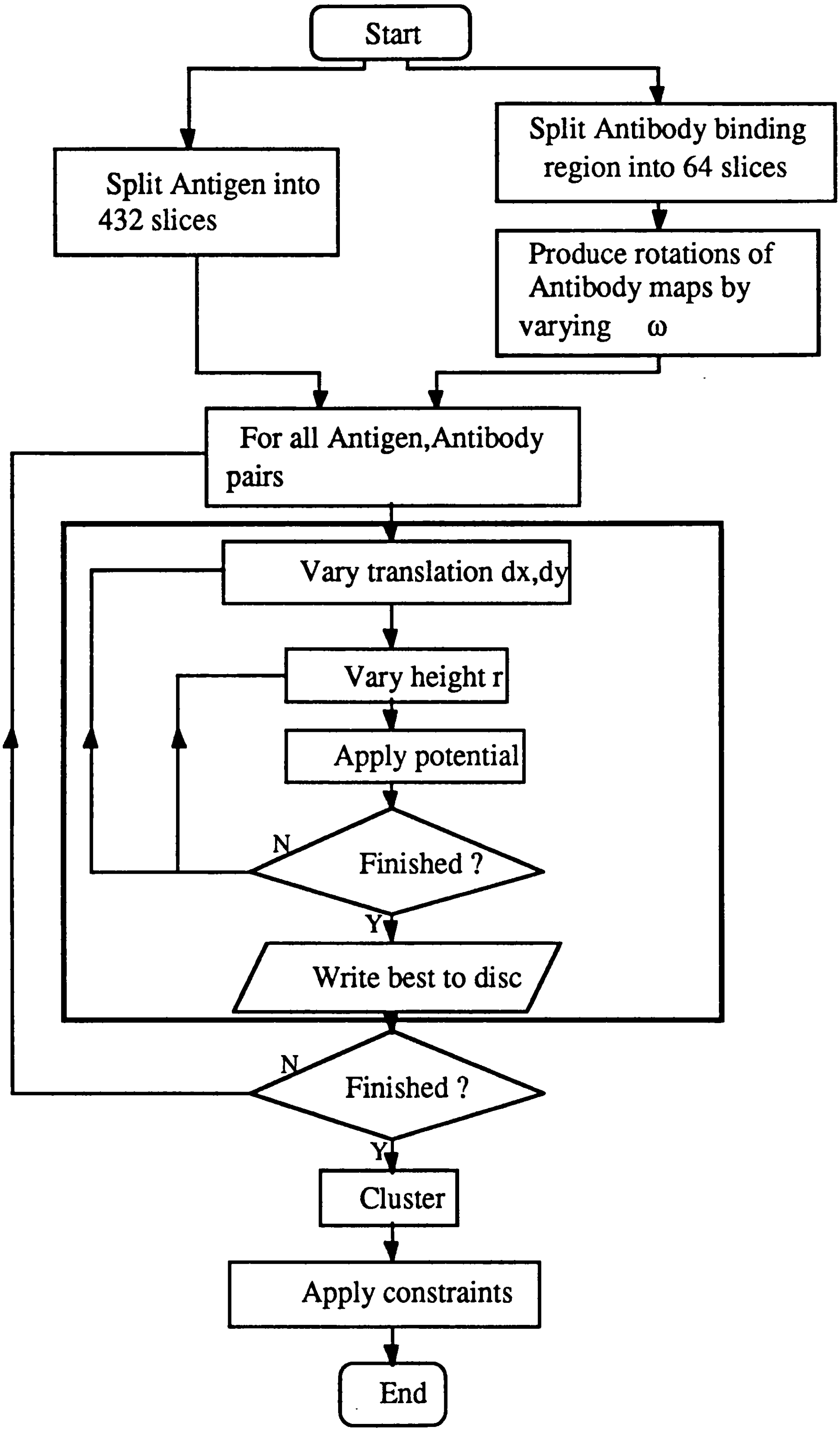

Start
Split Antigen into 432 slices
Split Antibody binding region into 64 slices
Produce rotations of Antibody maps by varying ω
For all Antigen,Antibody pairs
Vary translation dx,dy
Vary height r
Apply potential
N
Finished ?
Y
Write best to disc
N
Finished ?
Y
Cluster
Apply constraints
End

Figure 2.1

A flowchart of the whole DAPMatch algorithm. The section enclosed in a heavy-lined box is the computationally intensive section, and was implemented on the DAP computer.

The main docking procedure and the soft potential used, were developed and tested on the HyHEL-10 system. The constraints were benchmarked using both the HyHEL-10 and D1.3 systems. Oncesuitable parameters for the potential and subsequent constraints had been found the method was applied to HyHEL-5, the D1.3 model and the enzyme/inhibitor systems without any changes. This approach ensured that the method was not optimised to produce the correct results.

2.5. Implementation

Simulations of the docking problem are often computer intensive. The more orientations examined, and the more complete the treatment of the discrimination function, the more computer time required. To avoid this problem a massively parallel architecture machine, the DAP (AMT Ltd, Reading RG6 1AZ), was used. The DAP (*D*istributed *A*rray of *P*rocessors), is a 64x64 grid of closely coupled, simple processors (Figure 2.2). Each processor carries out the same instruction, but on different data. The DAP is therefore a single instruction, multiple data stream (SIMD) machine. The DAP is connected by a fast input/output (I/O) bus to a conventional architecture computer, called the host. This computer controls the execution of programs by the DAP and performs all the necessary file I/O for both computers. The DAP is programmed in a highly machine-specific form of FORTRAN which allows parallel instructions to be written in an understandable and compact form. DAPFORTRAN has the full set of FORTRAN mathematical routines, and many more which are useful in the parallel context of the machine, but unfortunately lacks any high level I/O functions.

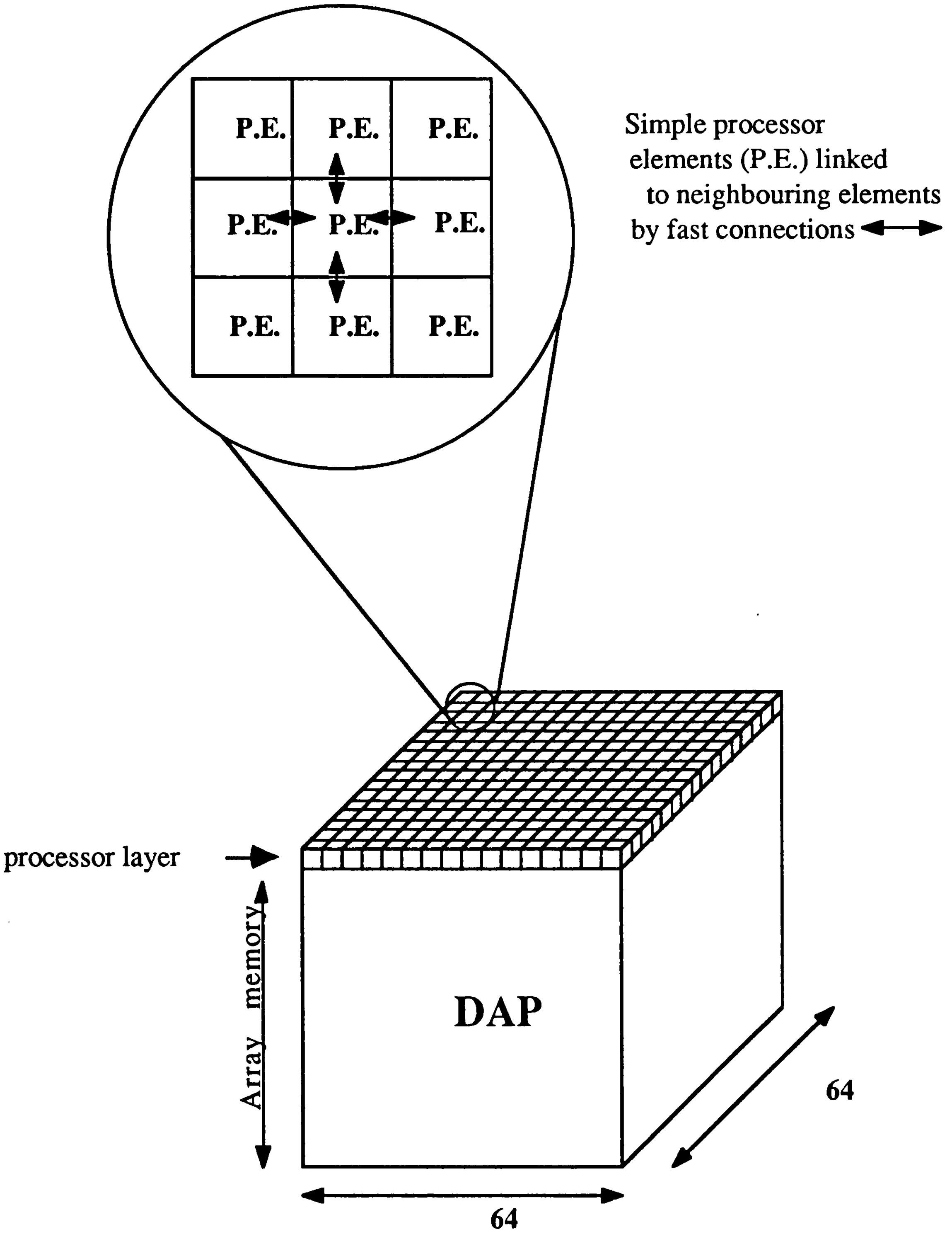

P.E.
P.E.
P.E.
P.E.
P.E.
P.E.
P.E.
P.E.
P.E.
Simple processor elements (P.E.) linked to neighbouring elements by fast connections ←→
processor layer
Array memory
DAP
64
64

Figure 2.2

The parallel architecture of the DAP computer, a 64x64 grid of simple processor elements (P.E.s). Each processor has fast connections to its four nearest neighbours (as shown in the inset). All instructions are carried out at the processor layer, the required data is brought to this layer from the array memory.

The DAP is best suited to problems which can readily be expressed in a parallel algorithm, which involve mainly integer arithmetic and which entail little I/O. It had already been decided to investigate the extent to which the steric matching of surfaces could be used to solve the docking problem. The simplest way of mapping a protein surface onto the DAP architecture was to take a planar slice through the protein, divide the plane into a 64x64 grid, and at each grid point find the height of the protein surface above the plane (Figure 2.3). For convenience these heights were taken to be integers in the range 0 to 63. Since the surface slices taken were 64 elements square it was convenient to map a single element onto each DAP processor. This allowed the energy summation to be carried out simultaneously for each of the 4096 elements. This simple mapping of the problem onto the DAP architecture allowed large speed improvements. A search could be completed within 2 days on the DAP, whereas it would have taken a Sun SPARC II around 100 days.

Whilst the DAPMatch steric search program was being developed the DAP at the ICRF was connected to a SUN 3. The SUN 3 was relatively slow and the clustering and constraint algorithms could not be implemented on it. These processes were also quite I/O intensive and so it was not convenient to implement them on the DAP. Instead the pre-processing and post-processing were implemented on a SUN SPARC II. This meant that large intermediate files had to be generated and passed between the computer systems. The DAP is now hosted by a SUN SPARC II. This allows the DAPMatch program to be more highly integrated, reducing the need for intermediate files and making the package easier to use.

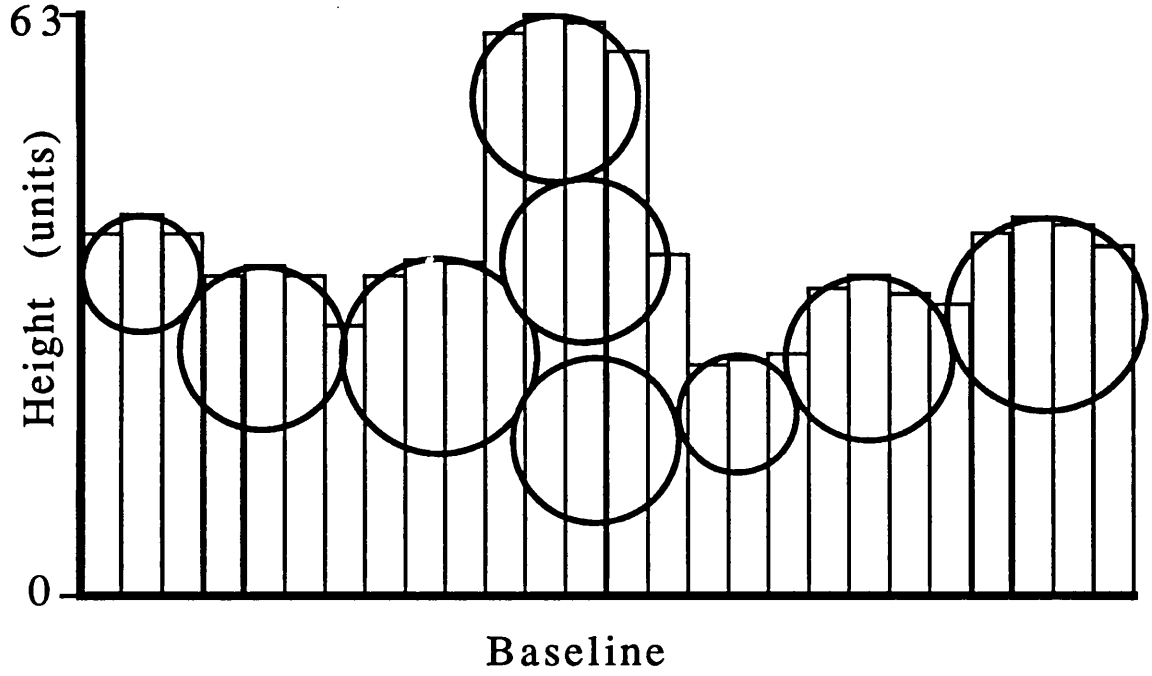


Figure 2.3

A two dimensional example of the DAPMatch representation of a protein surface. The van der Waals spheres of a hypothetical section of protein are shown. The surface is divided into 0.5Å strips, and a representative height for each strip is found. These heights are then scaled in the range 0 to 63, in 0.25Å steps. The highest surface element is always 63. The baseline shown corresponds to height=0 units, 16Å below the highest point.

2.6. Division of Sphere

To perform a global search of the guest molecule it is necessary to take many slices of the surface in as even a manner as possible. It would be difficult to account for the precise curvature of each protein. Instead the assumption was made that, for a globular protein, an equal division of a sphere would suffice.

The method used produced an N sided object by regular tessellation of an icosahedron. An icosahedron has 20 triangular faces. These faces were split into 10 adjacent pairs and each pair was tessellated identically. For convenience an icosahedron with unit side length was used.

The first step of the tessellation procedure was to choose two vectors, directed along the sides of one of the triangles, and with lengths $\frac{h}{h^2+hk+k^2}$ and $\frac{k}{h^2+hk+k^2}$. One vertex was chosen and translated using these vectors to create a new set of points (see Figure 2.4). Points which fell outside the triangle pair or which duplicated other points were discarded. This translation procedure was repeated for each set of new points until no unique points which fell within the triangle pair were found. This procedure resulted in a regularly tessellated icosahedron. The number of points, P, produced by this method depends on the integers h and k. Large values for these integers produce small vectors and hence a greater density of points. The number of points is given by the formula

$$P = 2 + 10(h^2 + hk + k^2) \quad .$$

When $h \neq k$ the tessellation produced is not symmetric and a handedness is introduced. This handedness does not prejudice the regular spacing of the points and hence is not important for this application. Finally, the spherical polar co-ordinates of each point

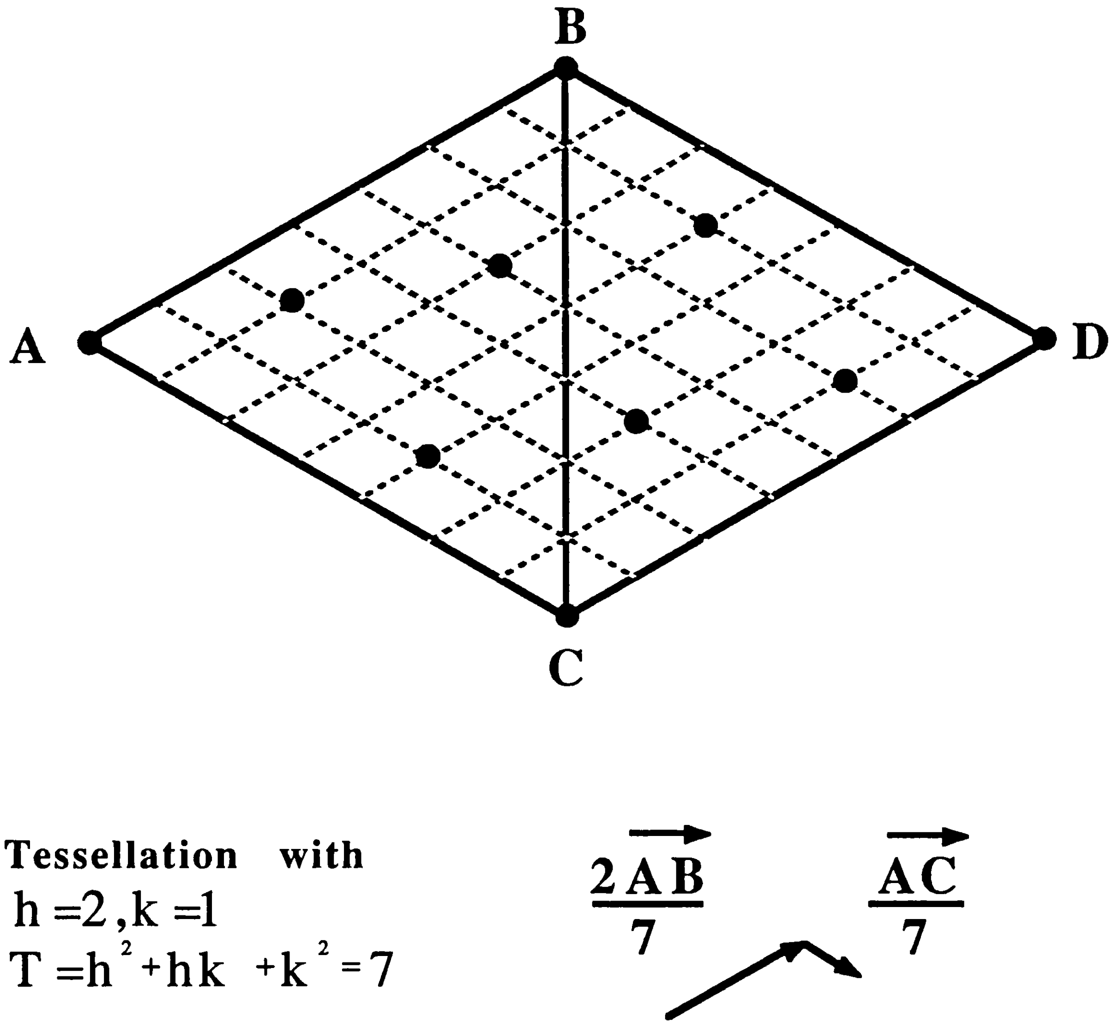

B
A
D
C
Tessellation with
$h=2, k=1$
$T=h^2+hk+k^2=7$
$\frac{2\overrightarrow{AB}}{7}$
$\frac{\overrightarrow{AC}}{7}$

Figure 2.4

An example tessellation. The icosahedron is split into ten pairs of triangular faces. Each pair is treated identically, the diagram shows one such pair. The tessellation shown is for h=2,k=1; this tessellation, carried out on the full icosahedron, would result in a semi-regular object with 72 vertices.

were found. The (θ, ϕ) co-ordinates could then be used to subdivide a sphere of any given radius (Figure 2.5). Projecting the points from the icosahedron to a sphere produced slight distortions in the spacing between points, this effect was minor, particularly with high densities of points.

2.7. Surface Slices

The regular object produced by tessellation must now be mapped onto each protein and used to produce an unbiased sample of the molecular surface. The centres of mass of the molecule to be sliced and the tessellated icosahedron were superimposed. The co-ordinates of the tessellated icosahedron were then projected onto the protein surface. The ensemble was then rotated so that each tessellation point in turn lay uppermost on the Z-axis. A slice of the protein surface was taken, centred upon this point. This slice was a 64x64 array of surface heights taken at $\frac{1}{2}$Å intervals, hence a 32Å x 32Å area was represented. The van der Waals surface for the protein was then calculated using a standard set of radii (Table 2.3). The greatest surface height found for each surface element was then found.

Once the height of each surface element, z'(i,j), was known the data was reduced to a set of integers in the range 0 to 63, intervals of $\frac{1}{4}$Å were taken and the scale arranged so that the highest surface point was denoted as 63 and anything further than 16Å below this point was denoted as 0. This method of quantizing the surface into discrete elements minimised the impact of small variations in conformation.

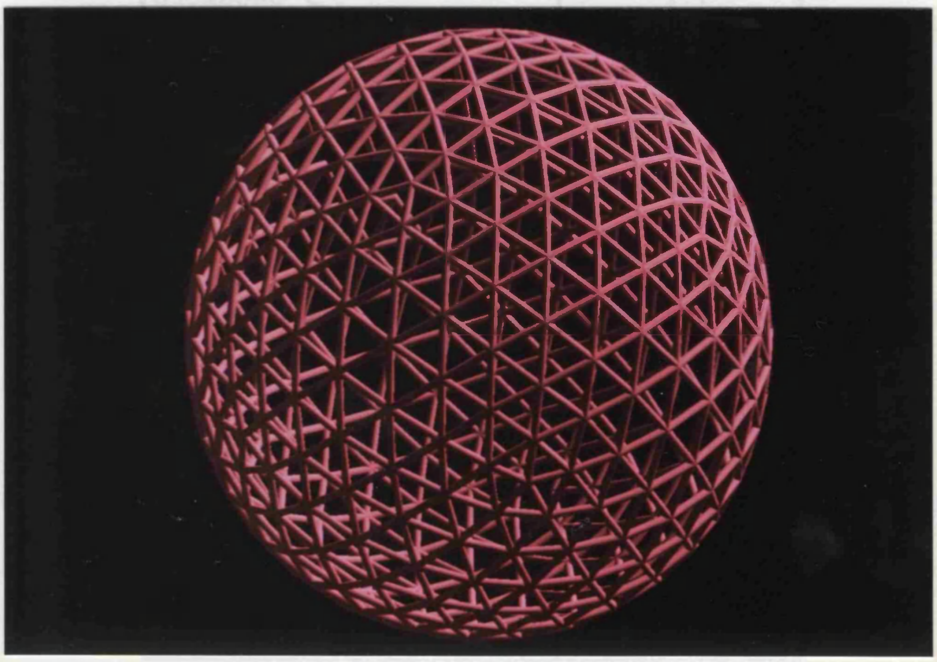


Figure 2.5

The 372 vertices of a tessellated icosahedron projected onto a sphere. The object was created by the tessellation corresponding to h=4, k=3.

Group Type	van der Waals radius (Å)
backbone N	1.7
backbone Cα	2.0
backbone Cα (proline)	1.8
backbone C	1.7
backbone O	1.4
sidechain CH CH2 CH3	2.0
N	1.7
NH	1.8
NH2	2.0
SH	1.5
S	1.8
hydroxyl oxygen	1.6
carboxylic oxygen	1.6
amide carbon	1.7
carboxylic carbon	1.7

Table 2.3

The van der Waals radii used for each atom type. These radii are taken from Gellately & Finney (1982).

To reduce further these effects, the maps were smoothed by adding in height components from the nearest neighbour and diagonal elements, so that for a particular element i,j

$$z(i,j) = \frac{z'(i,j)}{2} + \frac{z'(i-1,j) + z'(i+1,j) + z'(i,j-1) + z'(i,j+1)}{12} + \frac{z'(i-1,j-1) + z'(i+1,j-1) + z'(\mathrm{i}-1,j+1) + z'(i+1,j+1)}{24}$$

The fractional addition ensured that the range of heights remained the same. Finally, to enable a one-to-one correspondence between surface elements of host and guest maps, the surface of the antigen had to be inverted in the x-axis (notionally turning the surface upside down).

This procedure was carried out for both the host and guest molecules. The number of slices taken of the host molecule was generally restricted to a known receptor region (see the results Chapters for further details). The resulting data files were stored in DAP format using low level I/O functions so that they could be quickly retrieved when required.

2.8. Global search

A global search of the docking problem involves exploring six degrees of freedom (Figure 2.6). Four rotational degrees of freedom were accounted for by taking slices of the host and guest molecules. An additional degree of freedom was the perpendicular distance between the slices, r. The slices were held at a distance such that their surfaces just touched, the potential was calculated, and then

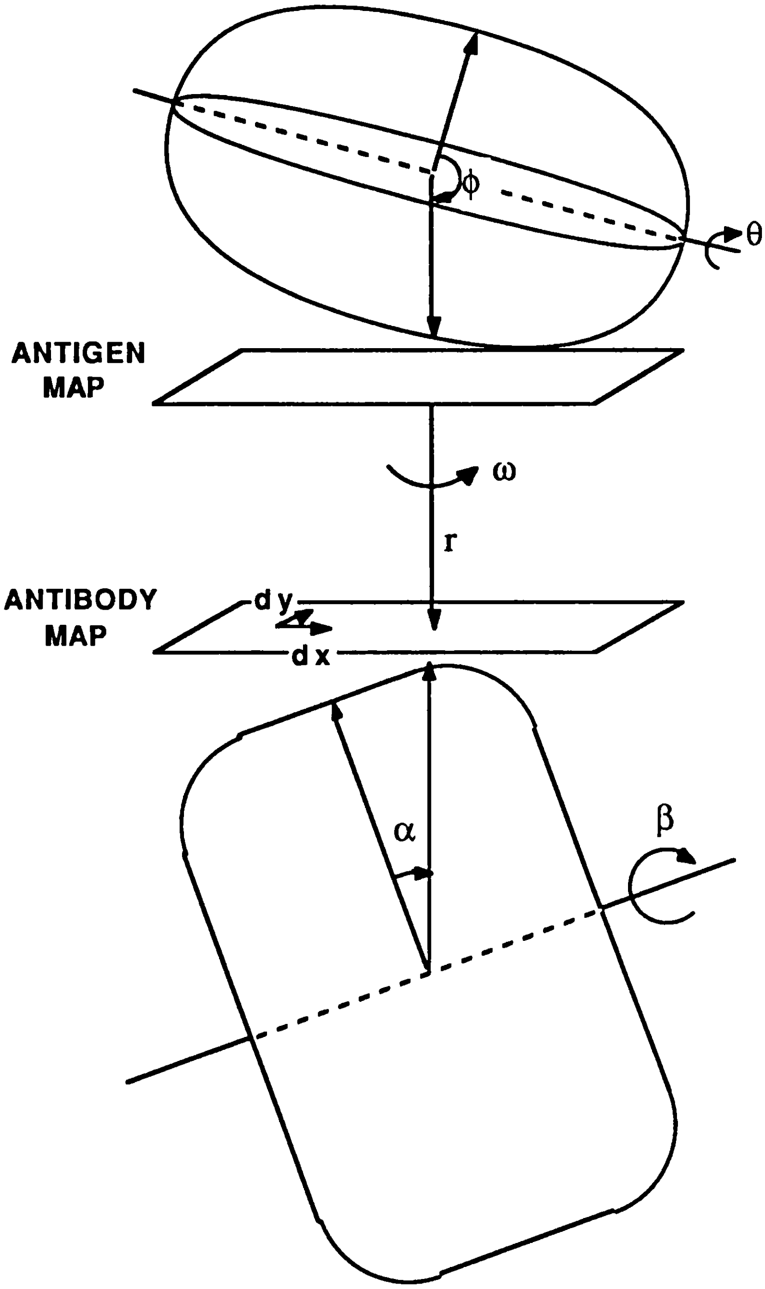


Figure 2.6

The degrees of freedom examined during the docking process (c.f. Figure 1.15). The guest molecule is shown uppermost with the host molecule below. The antibody and antigen maps show the notional docking area. The rotational degrees of freedom θ and φ bring any section of the guest molecule into this docking area. The angles α and β are the equivalents for the host molecule. The perpendicular distance between the two molecular surfaces, **r**, can then be varied. A rotation ω perpendicular to the plane of the surfaces completes the classic six degrees of freedom. Two perpendicular shifts in the plane of the docking area, **dx** and **dy**, were used to increase coverage of the search space without significantly increasing computational costs.

they were then moved together in $\frac{1}{2}$ Å steps until there two surface elements overlapped by 5Å (Figure 2.7). This level of overlap represents an unacceptable clash between the docking surfaces. Precise atomic detail was lost during the coding and smoothing process. This meant that the rotation in the plane of the slice, ω, the final degree of rotational freedom, could not be directly carried out on the slice. Instead, the protein structure was rotated in 8° steps and 45 maps slices taken for each host surface segment.

This gives six degrees of freedom. However, in the DAP architecture the translation of the surfaces could be done very quickly, since neighbouring processor are directly connected. It was therefore decided to make small shifts in the plane of the slices, dx,dy. This increased the coverage of each molecule without significantly decreasing the speed of the algorithm.

Only the binding region of the host molecule was used in the study, hence only 64 maps were taken instead of the global 432. The number of orientations considered, allowing for all degrees of freedom, was

$$N_{\text{orientations}} = \begin{array}{rll} & 432 \text{ guest molecule maps} & (\theta,\phi) \\ \text{x} & 64 \text{ host molecule maps} & (\alpha,\beta) \\ \text{x} & 45 \text{ rotational} & (\omega) \\ \text{x} & 5 \text{ internal dx shifts} & (dx) \\ \text{x} & 5 \text{ internal dy shifts} & (dy) \\ \text{x} & 10 \text{ height parameters} & (r) \end{array}$$

$$= 311{,}040{,}000 \quad .$$

The program DAPMatch was used to evaluate a simple steric potential for every possible combination of slices. The edges of a slice often contained no protein, and hence had a zero height. These

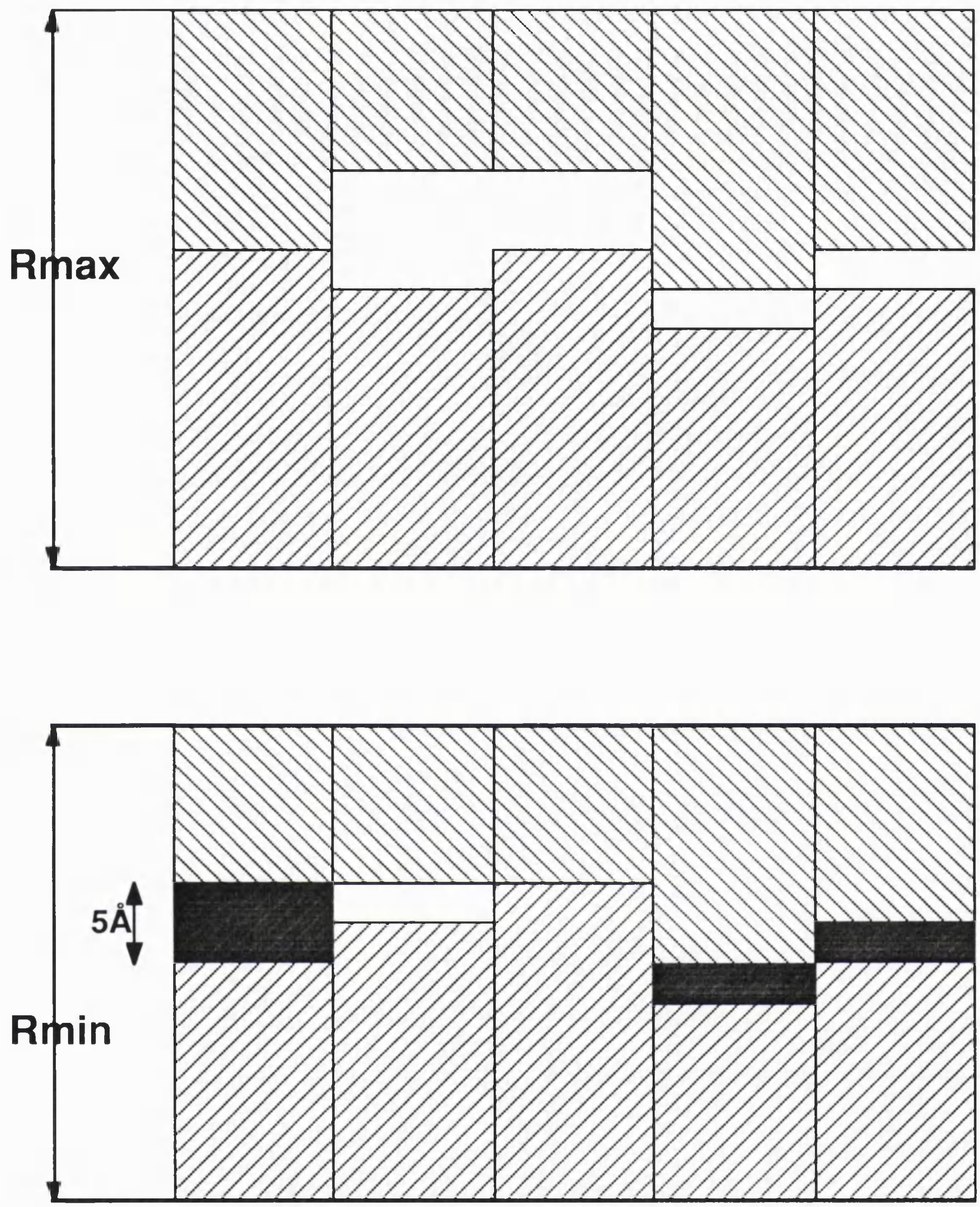


Figure 2.7

The range of the DAPMatch height search. The search starts with the surface slices just touching (R_{max}). The slices are then moved together in 0.5Å steps until two elements of the surfaces overlap by more than 5Å (R_{min}); the search is then terminated.

sections of the map were masked out of the steric calculation. A simple measure of the overlap between two maps was taken by counting the number of non-zero elements of one slice that were paired with non-zero elements of the second slice. If the overlap between the two maps was at least 2,000 elements then the best orientation was written to disc. By insisting on an overlap of 500Å^2 (2000 elements) a large number of docking possibilities, which were infeasible due to the small amount of contact between the surfaces, could be quickly discarded.

2.9. Steric Match

A major aim of the approach is to accommodate structural changes caused by induced fit and structural inaccuracies due to crystallographic and modelling errors. A softer type of Leonard-Jones potential, $V_{soft}(x)$, is used.

$$V_{soft}(x) = \begin{cases} 256x^4 & x < 0\text{Å} \\ 32x^2 & 0\text{Å} \leq x \leq 4\text{Å} \\ 512 & x > 4\text{Å} \end{cases} .$$

The displacement x is a measure of the spacing between two surface elements. Hence, x is negative when the hard-sphere van der Waals surfaces overlap, x is zero when the surfaces are just touching and x becomes increasingly positive as the surfaces separate.

This potential, $V_{soft}(x)$, follows roughly the same form as the Leonard-Jones potential (Figure 2.8) but is modified in two ways:

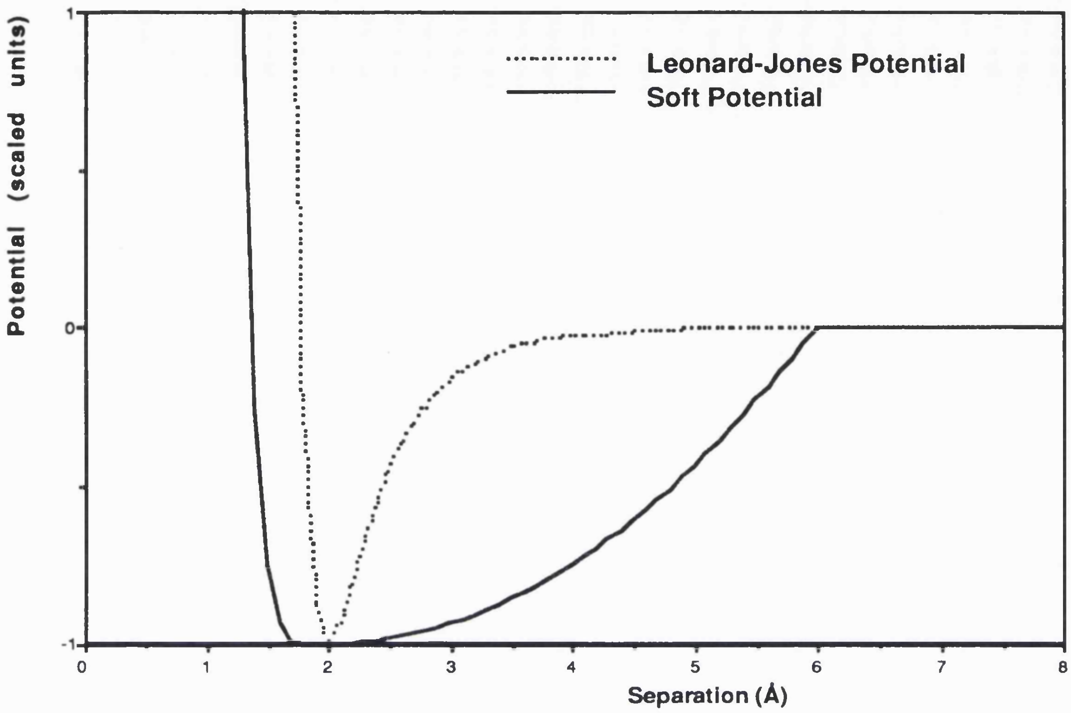


Figure 2.8

The soft docking potential used compared with a typical Leonard-Jones potential. The potential minimum of the soft potential is much wider than that of the Leonard-Jones potential. As the separation increases the soft potential tends to zero more slowly than the Leonard-Jones, this results in a preference for additional space between atoms.

i. The energetic potential for overlapping surface elements is much reduced, it is assumed that some degree of overlap can be accommodated by side-chain rearrangement.

ii. The soft potential has a much wider minimum than the standard Leonard-Jones potential. The standard potential has a narrow minimum well, with a width of around 0.5Å. This size is smaller than the amount of shift expected due to induced fit, and is roughly equivalent to the expected error in the large crystallographic structures such as the antibody/lysozyme complexes. The wider minimum of the soft potential provides tolerance of these shifts, thus allowing a steric match to be found even if areas of unfavourable overlap between the surfaces do exist.

At the point x=4Å a cut off is applied and the soft potential becomes zero. At this point the gradient of the potential is discontinuous; this is not, however, a problem since the soft potential will only be used to calculate the steric complementarity of static structures.

The soft potential is summed over all surface elements where protein data was contained in both maps. Each map contains areas where no protein data is found, particularly at the edge of the map. Such areas contribute nothing to the summed potential, hence the magnitude of the potential depends on the extent of overlap between the two maps. To correct for this effect a factor dependent on the number of overlapping elements ($N_{overlap}$) is subtracted from the final, summed potential.

$$V_{\text{total}} \quad = \quad \sum V_{\text{soft}}(x) \quad - \ 100N_{overlap}$$

Without this factor pairs of maps that had a large number of overlapping elements would have unfavourable (large and positive) potentials, simply because more elements contribute to the potential. The $N_{overlap}$ term is designed to over-correct for this tendency. This makes map pairs with a large overlap more favourable than those with little overlap. Thus the algorithm favours orientations which bury a large amount of surface area on docking.

2.10. Clustering

After the DAPMatch run, a file containing more than 200,000 possible docking orientations was obtained, each having a summed energy score that was a measure of the association potential. This file was reduced to a more manageable size by a simple clustering algorithm. The following approach was used to remove all orientations that were similar to an orientation with a better (i.e. lower) docking potential. The file was ordered upon the energy score and the most favourable orientation chosen. Then, one proceeds down the list and removes each similar orientation. One way of measuring the similarity between two guest molecule orientations, with angles ($\phi 1$, $\theta 1$) and ($\phi 2$, $\theta 2$), is to project these angles onto a unit sphere and measure the distance between them.

For each orientation

$$\begin{aligned} x &= \sin(\phi)\cos(\theta) \\ y &= \sin(\phi)\sin(\theta) \\ z &= \cos(\phi) \quad . \end{aligned}$$

The similarity is taken to be the square of the distance between them.

$$
\begin{aligned}
S_{\text{guest}}(\phi,\psi) = \; & (\ \sin(\phi 1)\cos(\theta 1) - \sin(\phi 2)\cos(\theta 2)\ )^2 \\
& + (\ \sin(\phi 1)\sin(\theta 1) - \sin(\phi 2)\sin(\theta 2)\ )^2 \\
& + (\ \cos(\phi 1) - \cos(\phi 2)\ )^2
\end{aligned}
$$

The similarity is evaluated analogously for the host maps, with angles $(\alpha 1, \beta 1)$ and $(\alpha 2, \beta 2)$. There is also the angle ω, describing the relative planar rotation, which cannot be easily included in this scheme. The solution was to add in an extra term,

$$
S_{\text{host}}(\alpha,\beta,\omega) = \quad S_{\text{guest}}(\alpha,\beta) + 2\left|\ \sin\left(\frac{\omega 1 - \omega 2}{2}\right)\right| .
$$

These similarity scores were calculated for both the host and guest maps and compared against a set tolerance level (0.45 in each case), if both scores were lower than this then the orientations were taken to be similar and the less energetically favourable was discarded. Once all orientations similar to the top one had been discarded the process was repeated for the orientation now second in the list, and so on till the last orientation was reached. In this way the number of orientations was reduced to less than 2,000 , but an even coverage of the host and guest surface was maintained.

The measure of similarity used was easy to calculate and sufficient for the purpose of this clustering algorithm. The more discriminating the DAPMatch program is, the less important the exact details of the clustering become. If the steric search produces one near-native orientation with a favourable score, surrounded by less favourable orientations, the results from the clustering algorithm will always contain this solution. A more mathematically rigorous difference function could be used.

2.11. Electrostatics

The next step was to eliminate orientations that had unfavourable electrostatic interactions. An exact energy model could not be used since side-chain position is often changed during the docking process. A slight χ angle rotation, for instance, can make a large change in position of a polar or charged group at the end of a residue. A simplified model, similar to that of Levitt (1976), was used to overcome this problem (Figure 2.9). Each residue was represented by a single sphere, of fixed radius (Table 2.4). The radii shown are those of Levitt (1976). These values were multiplied by a factor of 1.5 to decrease the positional dependence of the model. For glycine, this sphere was placed at the carbon alpha (Cα) position, in all other residues it was placed at the centre of co-ordinate of the side-chain atoms. The radius for the glycine sphere takes into account this difference, and so is larger than the alanine radius. A simple residue-residue interaction energy was devised (Table 2.5a), assigning each possible interaction as favourable (-1), neutral (0) or unfavourable (+1), based upon the known properties of the residues involved (Table 2.5b).

The simple sphere model was used to find possible host-guest interactions for each remaining orientation. The model was created and host spheres that overlapped guest spheres were found, these residues were considered likely to interact in the complex (Figure 2.9). Summing the energies over all interactions found gave an total energy term indicative of electrostatic complementarity. Configurations that were not favourable overall (i.e., summed potential ≥ 0) were rejected.

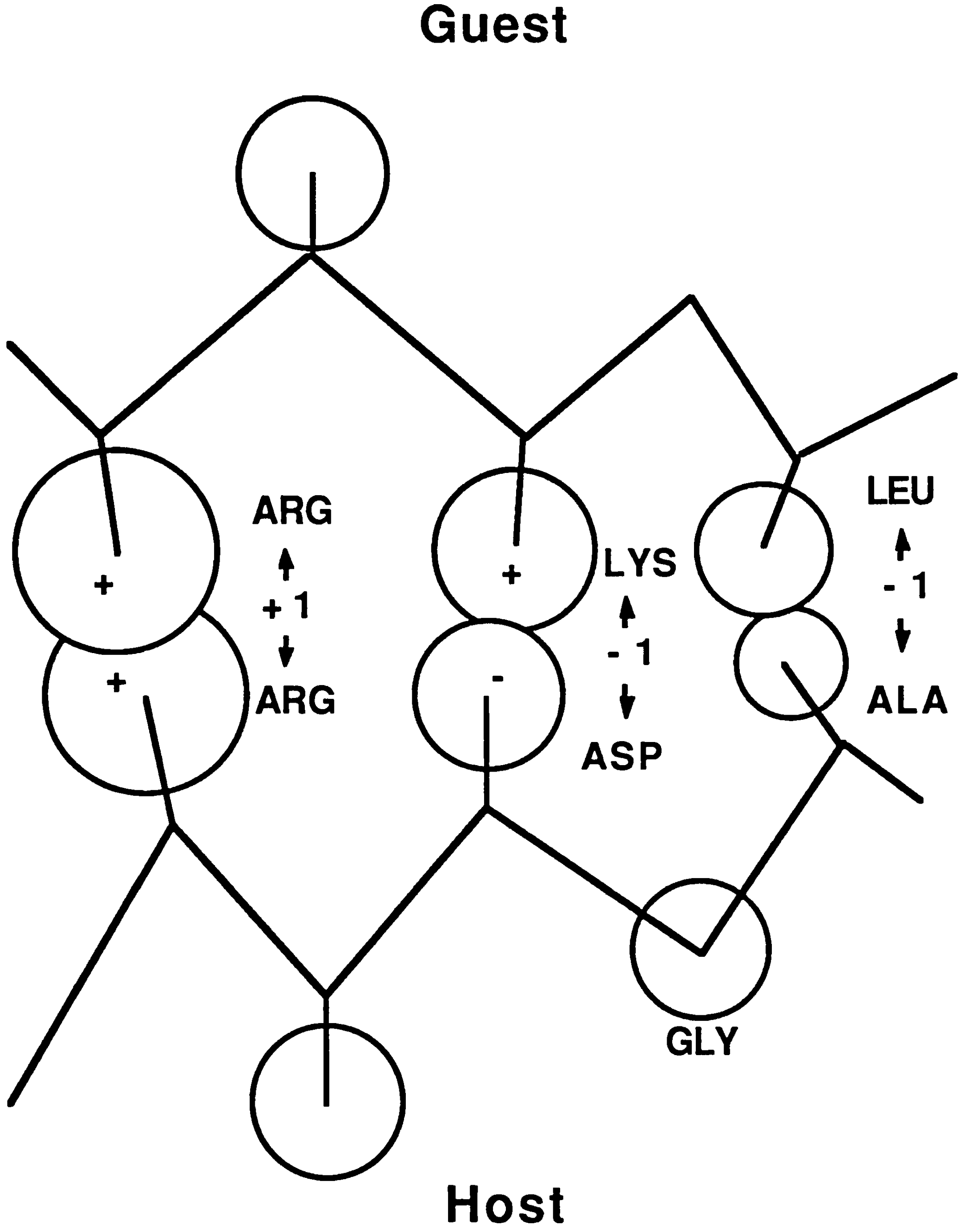


Figure 2.9
The simplified electrostatic model. The Cα chains of both the host and guest are shown. Each residue is represented as a sphere of specific radius (Table 2.4) centred at the centroid of the sidechain atoms. Glycine is represented by a sphere centred on the Cα position. The spheres of interacting residues overlap. A simple potential (Table 2.5a) is summed for all residue pairs which interact across the host/guest interface.

Residue	Radius of Gyration (Å)
Ala	0.8
Arg	2.4
Asn	1.5
Asp	1.4
Cys	1.2
Gln	1.8
Glu	1.8
Gly	1.1
His	1.8
Ile	1.6
Leu	1.5
Lys	2.1
Met	1.8
Phe	1.9
Pro	1.3
Ser	1.1
Thr	1.2
Trp	2.2
Tyr	2.1
Val	1.3

Table 2.4

The electrostatic model parameters. Each residue is represented by a single sphere, with the given radius (see also Figure 2.9).

	Polar	Hydrophobic	Positive	Negative
Polar	- 1	0	0	0
Hydrophobic	0	- 1	+1	+1
Positive	0	+1	+1	- 1
Negative	0	+1	- 1	+1

Table 2.5a

A simple electrostatic potential. Favourable interactions, e.g. a positive charge/negative charge interaction, score -1. Neutral interactions score 0 and unfavourable interactions, e.g. a hydrophobic/positive interaction, score +1.

Hydrophobic	Ala, Cys, Gly, Ile, Leu, Phe, Pro, Val
Polar	Asn, Gln, Met, Ser, Thr, Trp, Tyr
Positive	Arg, His, Lys
Negative	Asp, Glu

Table 2.5b

Amino acid properties. Each amino acids is classified according to its major characteristic.

2.12. Epitope and Distance Constraints

After clustering and electrostatic evaluation, several possible associations still remained. The next step was to explore the extent to which information about the surface residues involved in a contacts across the interface of a particular complex could further reduce the possible orientations. The simplest assumption that could be made was that the epitope of the guest molecule was known. This effectively reduces the number of valid guest slices by insisting that certain residues must be present in each slice. Only slices that contained at least 90% of the epitope residues were considered.

A more limiting constraint was the requirement that a single interaction exist across the interface. Thus it was required that a single residue of the guest molecule was within interaction range of a single residue of the host. A precise constraint could not be applied due to uncertainty in side-chain position. Instead a model similar to the electrostatic model was used. The Levitt parameters (Table 2.4) were multiplied by the factor 3.5, and a check was made that the two residues were within the appropriate interaction distance. This, very loose, distance constraint allowed for large shifts in side-chain position.

2.13. Testing

The algorithm presented in this chapter was used to explore both antibody/antigen and the EGF/EGFBP systems, and with small changes to explore helix/helix packing. Many of the parameters involved in the steric search were chosen empirically as the result of testing with the HyHEL-10 complex. The first parameter to be chosen was the scale of the grid used in representing the surface. The dimension of the grid, 64x64, was chosen to correspond to the size of the DAP. Each element was taken to represent a 0.5Å square of the surface. This meant an entire slice represented a 32x32 Å section of the protein. The value of 0.5Å was convenient since it was comparable to the scale of the errors in a protein structure. The resulting 32x32Å map covers 1024Å^2 of the surface at a time. Typically an antibody/antigen complex buries 1500Å^2 of protein surface, roughly 750Å^2 from each molecule. Hence a single surface slice could encompass the whole of the relevant epitope of a molecule. Cross sections of globular proteins are not square but ellipsoidal and so in practise small portions of the surface were chopped off at the edge of the grid. Nevertheless, trials with the epitope region of HyHEL-10 suggested that larger element sizes (such as 0.75Å and 1.0Å) lost too much detail of the protein surface whilst adding little to the amount of surface covered.

Trials using the HyHEL-10 complex involved taking the binding region map from the antibody and matching it with six areas of the antigen slices. One of these antigen slices was known to bind to the antibody. The other five were non-overlapping, randomly chosen slices. The antigen slices were rotated in 5° steps, perpendicular to the interface (angle ω in Diagram 2.6) and compared with the antibody slice at each step. The steric potential for each rotation was examined to see if the correct (binding slice with ω=0°) solution had the most favourable potential value. The effect of altering the steric matching function, the grid size and the amount of smoothing performed on both maps was explored in this way.

Particular care was taken to produce the best steric matching function. The general form of the potentials tried was

$$V_{soft}(x) \quad = \quad \begin{cases} Ax^{n} & x < 0\text{Å} \\ Bx^{m} & 0\text{Å} \leq x \leq d\text{Å} \\ C & x > d\text{Å} \end{cases} \quad .$$

The most important parameters to be chosen were *n* and *m* , which determine the curvature of the potential when atoms clash and when atoms become widely spaced respectively. The Leonard-Jones potential rises sharply as the atoms approach each other from the equilibrium point. To model this *n* was chosen to be 4, this created a potential that effectively disallowed atomic overlap of several angstroms yet was tolerant of overlaps less than this. It was found that values of *n* less than 4 allowed too much mismatch between the surfaces, and discrimination of steric complementarity was lost. Values of *n* greater than 4 were not sufficiently tolerant and only produced favourable interaction values for surfaces which had a near perfect match. Such potentials would not have tolerated the level of inaccuracy created by using uncomplexed molecules or models. The value of *m* was chosen to be 2. This produced a potential which favoured close packing between the surfaces but which was tolerant of cavities between them. The values chosen for *A*, *B* and *C* produced a potential with a wider minimum than that seen for a typical Leonard-Jones potential (Diagram 2.8). This potential, when used to match two surfaces, allowed areas of the surfaces to mismatch by several angstroms without rejecting them completely. A value of 4Å for *d*, the distance beyond which the surfaces ceased to interact, was found to be acceptable. The values taken for *d* and *C* effectively determine the depth of the potential well.

Differing levels of smoothing were also tried. Smoothing was necessary to remove small scale detail from the surface slices. These details would be highly likely to change during the binding process and so could not be used in a steric complementarity search. Higher levels of smoothing than the one eventually chosen were found to remove too much detail from the slices, leaving only gentle, large scale undulations of the surface which were not specific enough to allow the correct orientation to be found.

The electrostatic model was evaluated using preliminary results produced for both the HyHEL-10 and D1.3 complexes. Several sets residue-residue interaction energies were tried, and compared with the simple set (Table 2.5a,b). The most rigorous set of energies evaluated were those of QPACK, derived by Gregoret & Cohen (1990) by observing the residue/residue contacts present in a set of well defined protein structures. A model similar to the to the simple interaction energies but which took account of hydrogen bonding was also tried. It was found that increasing the complexity of the electrostatic model produced, at best, no significant increase in accuracy. The model which explicitly took account of hydrogen bonding performed badly. This was due to the large number of hydrogen bonds potentially available at the protein surface, all of which had to be counted as possible bonds due to the high mobility of surface side-chains. The simple favourable/neutral/unfavourable model was therefore used.

Chapter 3

Antibody/Antigen Results

3.1. Synopsis

This chapter discusses the results obtained when the DAPMatch soft docking algorithm was applied to antibody/lysozyme complexes. The difficulties of assessing the accuracy of a particular docking are discussed in detail and a set of random trials have been carried out to help determine the significance of the results. To demonstrate the effectiveness of the soft docking algorithm the unbound crystal structure of lysozyme has been used throughout. Exactly the same method is followed in each application of the DAPMatch algorithm. This demonstrates the general applicability of the method to antibody systems. A model structure of the D1.3 antibody has been used as the host molecule in a docking search and the relevance of the result obtained is discussed in relation to recent advances in antibody modelling. A selection of binding orientations that had favourable steric scores but do not correspond to the crystallographic structure will be examined to assess whether there are any systematic differences between non-native and native solutions.

3.2. Introduction

The DAPMatch program was developed for and tested on the antibody/lysozyme systems. These are an obvious choice for investigation by a docking algorithm because:

i. Three antibody/lysozyme structures have been solved crystallographically to an adequate resolution. A docking algorithm can be applied to all three. If it is successful in each case then it is likely that the algorithm is generally applicable to antibody/protein systems, rather than specific to a single complex.

ii. The structure of uncomplexed lysozyme has also been determined crystallographically. This enables realistic docking simulations where induced fit has not produced perfect steric complementarity.

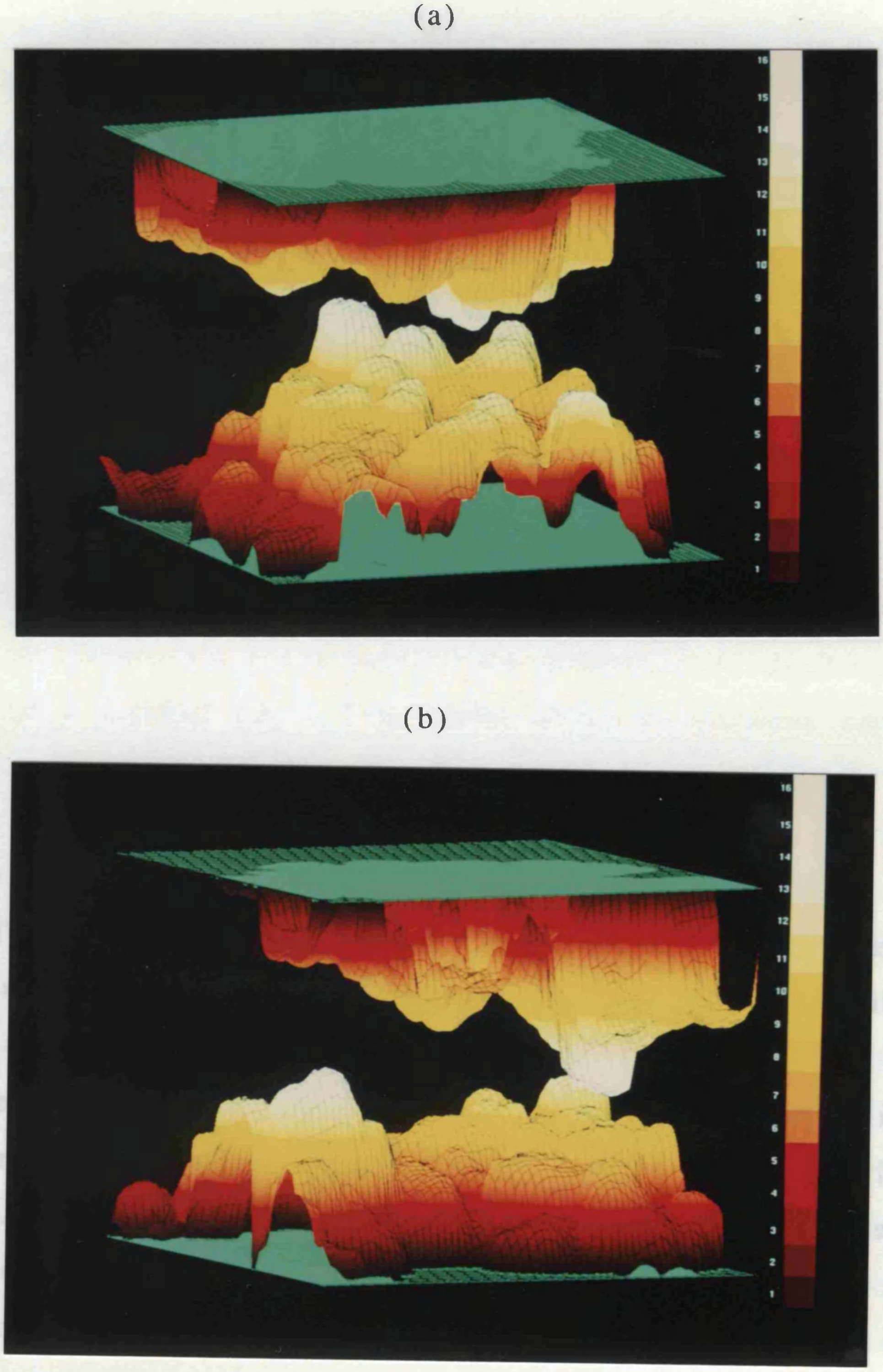
iii. Much progress has been made on predicting the structure of an antibody from its sequence (Section 1.8). A docking algorithm which could dock an antigen onto a predicted crystal structure would be useful in areas such as antibody engineering experiments.

iv. The general properties, in particular the large surface area of contact, of the antibody/lysozyme complexes are similar to many other protein/protein docking problems e.g. subunit association. An algorithm which was designed and tested on antibody/antigen systems would be readily applicable to these similar problems.

Author and resolution data for each of the antibody/lysozyme complexes is given in Table 2.1. The structures for both HyHEL-5 and HyHEL-10 have been deposited in the protein databank (Bernstein, *et al.*, 1977). The structure of the D1.3 complex was provided by Dr S.

Phillips. Each of the crystal structures has a resolution in the range 2.5Å-3.0Å. This level of resolution implies a well determined electron density map leading to a structure with generally reliable side-chain positions. As stated in Section 1.7.4, each complex involves the burial of roughly 1500Å^2 of solvent accessible surface area and the loss of virtually all water from this interface region. Figure 3.1 shows the interface regions of the HyHEL-10, HyHEL-5 and D1.3 complexes. The lysozyme slices are shown uppermost with the corresponding antibody slice below. Each slice is colour coded on height and the 0.5Å grid lines are marked. In the crystallographic complex these surfaces touch; in Figure 3.1 they have been separated to aid viewing. Also, the element heights of each slice have been exaggerated, by a factor of 2, to make the degree of steric complementarity more obvious. The D1.3 interface shows a convex region of lysozyme occupying the concave depression formed by the antibody. This general situation is also seen in the HyHEL-5 complex. HyHEL-10, however, has a broad, flat interface region. Despite the differences in overall interface shape all the complexes show many specific areas of antibody/antigen surface complementarity. This clear steric complementarity between the surface slices of each interface region justifies the approach taken by the DAPMatch algorithm.

The crystal structure of unbound hen egg white lysozyme was solved by Blake *et al.* (1965) with a resolution of 2.0Å. The structure was later refined (Diamond, 1974) and deposited in the protein databank. The structure of the D1.3 antibody was predicted by Chothia *et al.* (1986), prior to its being determined crystallographically (Section 1.8). The coordinates for this structure were supplied by Dr A. Lesk.


(a)
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
(b)
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1

(c)

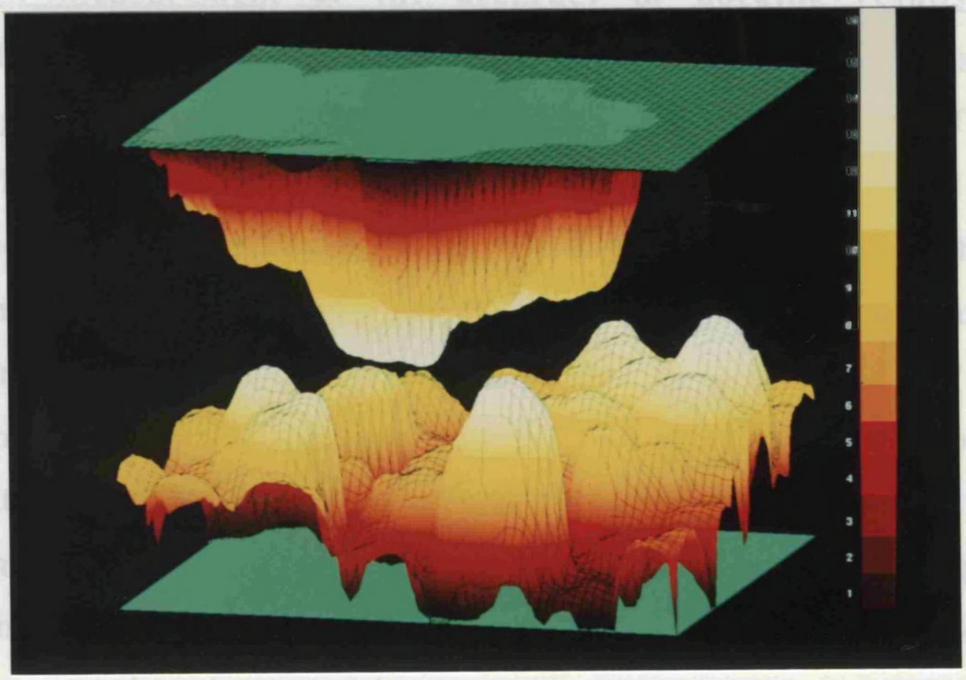


Figure 3.1

The interfaces regions of the three antibody/lysozyme complexes, HyHEL-10 (a), D1.3 (b) and HyHEL-5 (c). The antibody surface is shown below the lysozyme surface. The surfaces have been separated and scaled to aid viewing. The height scale (far right) is in angstroms. The interface of HyHEL-10 is broad and flat. The surfaces of D1.3 and HyHEL-5 show a greater degree of large scale complementarity, with the convex lysozyme epitopes fitting into concave depressions formed by the antibody binding regions. All three interfaces contain areas of specific complementarity between antibody and antigen.

3.3. Measurement of the accuracy of docking results

The result of the docking algorithm is a number of possible docking orientations. A fair assessment must be made of the similarities and differences between each docking orientation and the true structure. The antibody structure for each docking is fixed, and so superimposing these structures enables a one-to-one comparison between corresponding lysozyme atoms. Since the structure of free lysozyme has been used in the docking procedure the side-chain orientations differ from the complexed form. Even if the two lysozyme molecules being compared are in exactly the same spatial position and orientation, the difference in position of the side-chain atoms means that the r.m.s. between the molecules is non-zero. This is called a residual r.m.s. This makes direct comparisons between similar orientations difficult. One possible measure of similarity that largely avoids this problem, is a r.m.s. deviation between Cα atoms of the docked and true structures. This measure is reasonable because the protein backbone is largely unchanged during the docking process (Section 1.11).

By replacing free lysozyme with the complexed form in the predicted complex it is possible to determine rigorously the translations and Euler rotations necessary to superimpose the calculated lysozyme position on to the actual position. To superpose two structures requires six degrees of freedom, hence six numbers are produced. This measure is more descriptive of the orientational difference between the two structures, e.g. it is possible to judge how much of the difference is due to bulk translation and how much is due to rotation. Unfortunately it is difficult to compare many sets of orientations using this measure as it is not clear whether, for example, a molecule which differs from the target molecule by

(-1,3,-2)Å translation and (0°,3°,-5°) is closer to the target than one which differs by (0,0,-1)Å translation and (10°,-5°,8°) rotation.

One major problem with both the r.m.s. deviation and the superposition measure is that both methods are global and therefore large changes remote from the interface region produce unfavourable figures, even if the interface region is reproduced reasonably. Since the DAPMatch algorithm examines only the interface region, better measures of similarity would use only interface atoms. For this reason a r.m.s. of Cα atoms in the interface is calculated. The final measure of a docking solution is whether it forms the residue/residue contacts observed in the crystallographic complex. An useful docking algorithm must produced orientations which form the majority of the correct contacts across the host/guest interface. To assess whether the antibody/lysozyme results produced by the DAPMatch algorithm meet this criterion tables will be presented which show the residue/residue contact distances across the interface for both the crystallographically observed complex and the predicted complex.

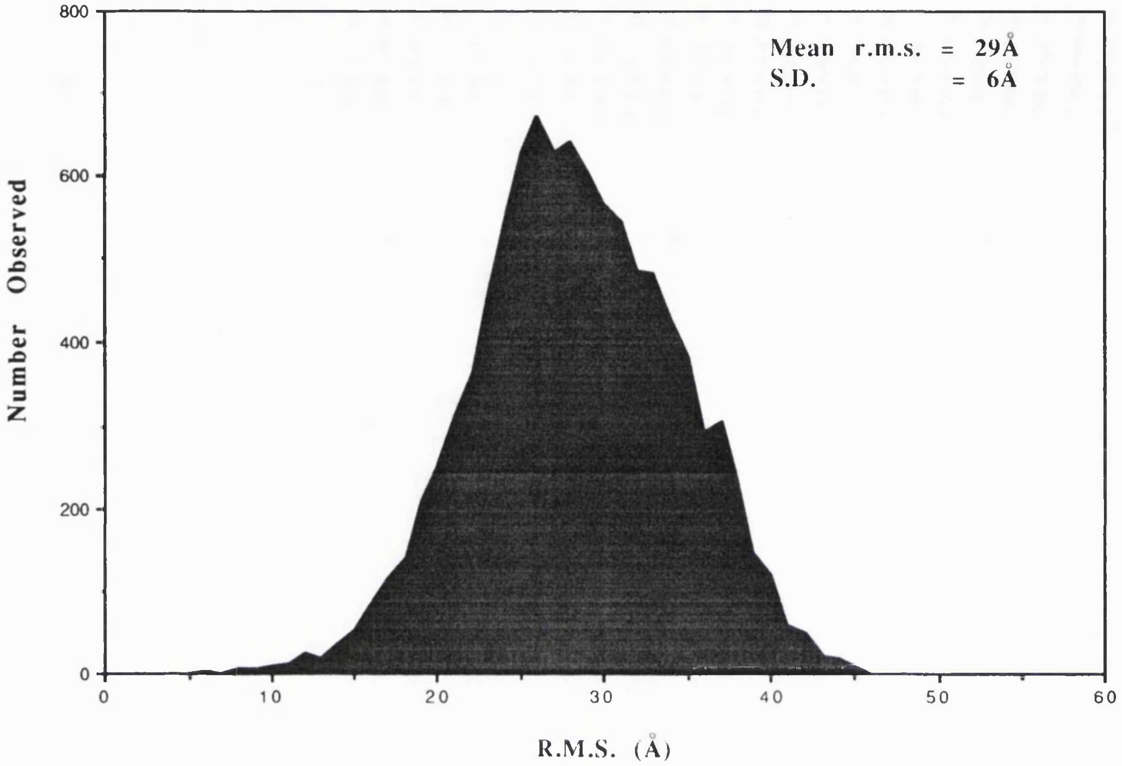
The measure of success most often quoted is the r.m.s. deviation between predicted and observed guest molecules. Before presenting results of this type it is necessary to carry out random trials using the known antibody/lysozyme structures to discover the significance that can be attributed to a given r.m.s. deviation.

3.4. Histograms.

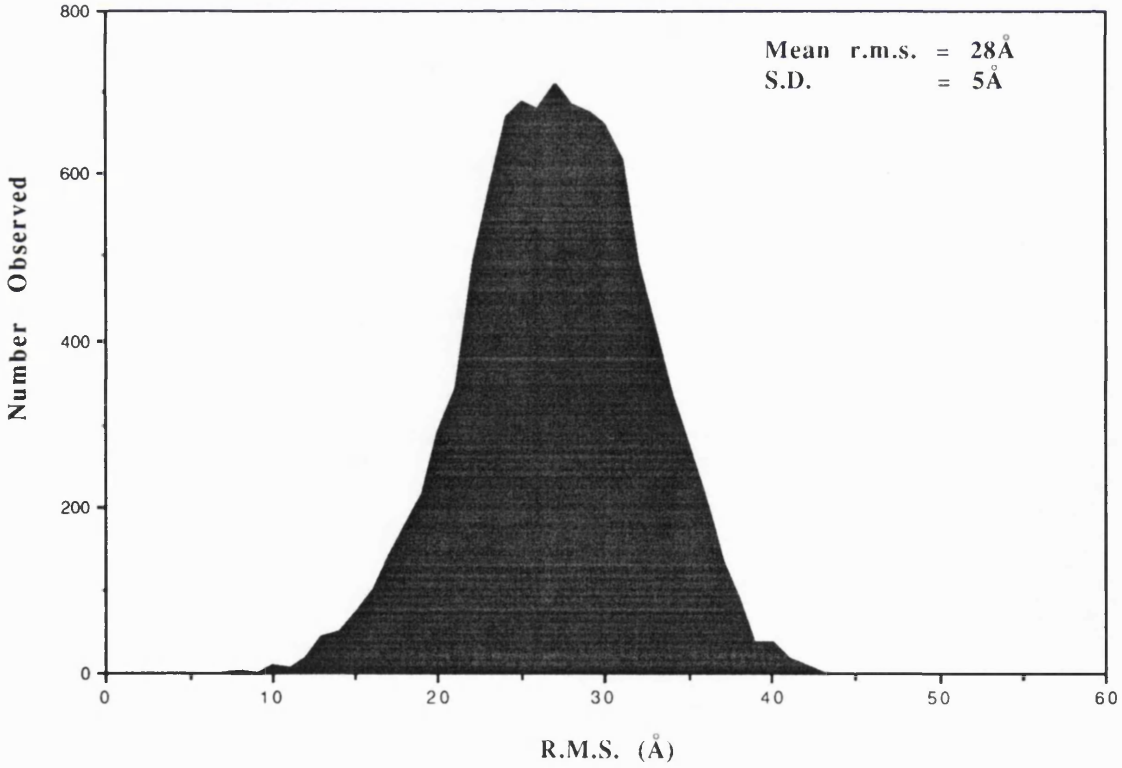
To test the significance of the r.m.s. deviations obtained a set of random configurations were calculated. By choosing suitable ranges for each docking parameter it was ensured that the random configurations covered exactly the same area of the search space as those generated by the DAPMatch program. Ten thousand random configurations were obtained in this way and the r.m.s deviations from the true complex were measured for the three interface regions. Figure 3.2 shows the distributions obtained, along with the mean r.m.s and the standard deviation from this mean. This approach is a simple measure of the significance of the deviations to be quoted later. The interface distributions vary slightly because the distance from the interface to the centroid varies; this distance is crucial to the r.m.s. change due to rotation. The interface of HyHEL-10 is closer to the centroid than average and so the r.m.s. deviations seen are lower than for the other complexes. The random trials produced a mean interface r.m.s. deviation of 29Å and a standard deviation of 6Å. Four trials had a deviation lower than 7Å, none of these were below 6Å. The best random configuration, when measured against the D1.3 interface, had a r.m.s of 5.5Å. This orientation was one of only four with deviations less than 8Å. The highly curved interface of HyHEL-5 produces a very broad spread of random r.m.s. deviations. The standard deviation rises to 8Å and one random configuration has a r.m.s. deviation of 4.5Å. This is the only deviation below 5Å.

The random trials for each complex show that 10,000 orientations do not produce significant numbers of orientations with r.m.s. deviations lower than 10Å. The HyHEL-5 random trial is the only one to produce a r.m.s. deviation below 5Å. The docking method presented results in fewer than 2,000 orientations after steric matching. Hence interface deviations lower than 5Å will be considered significant.

(a) HyHEL-10 Random R.M.S. Distribution



(b) D1.3 Random R.M.S. Distribution



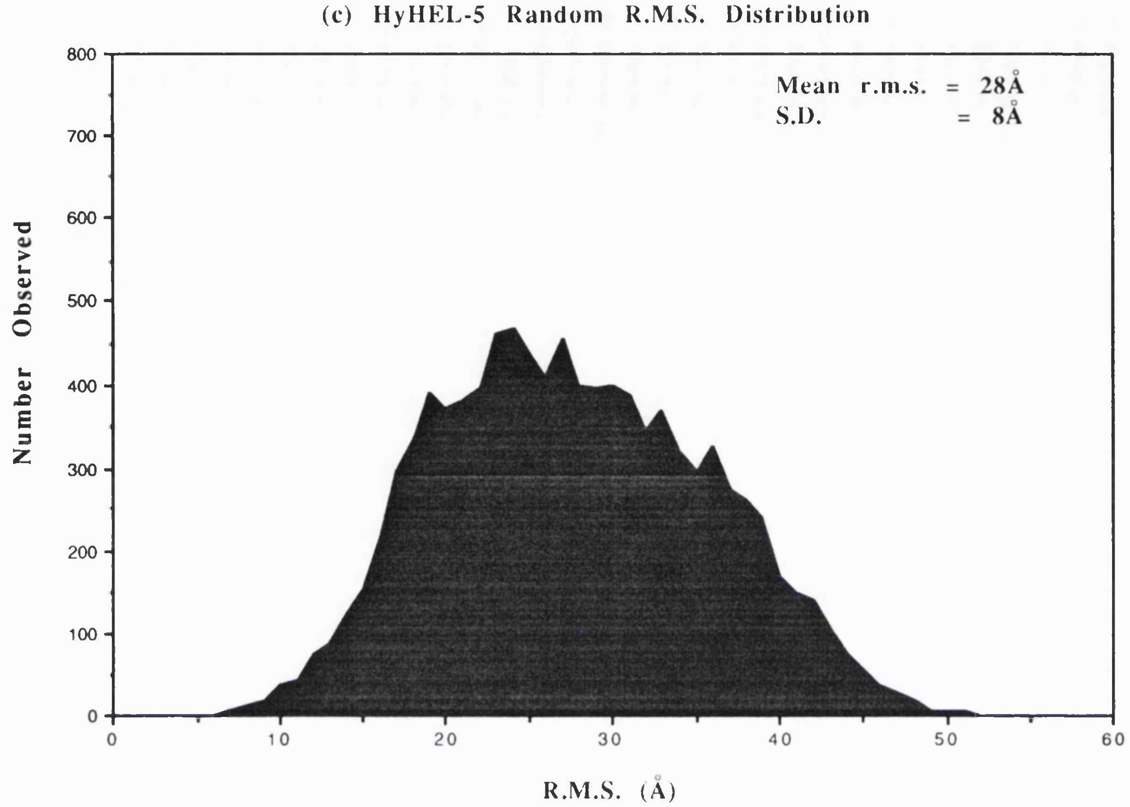


Figure 3.2

The distribution of r.m.s. deviations in the interface region of each of the three complexes, HyHEL-10 (a), D1.3 (b) and HyHEL-5 (c). Ten thousand random orientations were generated and their r.m.s. deviations from the true complex structure calculated. The mean deviation, and the standard deviation from this mean, are shown. The HyHEL-5 distribution is notable for the broad range of r.m.s. deviations it produces. Each distribution is contained mainly within the 10Å < r.m.s < 50Å range with very few orientations having deviations below 10Å.

3.5. Filtering.

Table 3.1 shows the results of docking native lysozyme to the four antibodies (three structures and a model) being considered. The table shows the number of orientations and the position of the best orientation at various stages of the procedure. Table 3.2 shows several measures of how well the best fit achieved by the algorithm corresponds to the crystallographic complex.

Of the 310 million orientations considered less than one thousandth are written to disc; the rest have insufficient contact areas or differ by only a slight translation or planar rotation from a better solution. The clustering process reduces the number of orientations from more than 200,000 to less than 2,000 without significantly altering the distribution of scores or r.m.s. deviations. When the electrostatic cut-off is applied a further 200-400 orientations are removed. These orientations have electrostatic clashes that would be impossible to accommodate in a structure.

Epitope information (Table 3.3) was used to reduce the number of antigen slices from the initial 432 to around 70. Each slice was required to contain at least 90% of the epitope residues. This left 58 valid slices for D1.3, 59 for HyHEL-10, and 86 slices for HyHEL-5 (see Table 3.1). The choice of epitope residues was not critical to the results produced; epitope information could be varied, by adding or removing a few residues, without significantly altering the resulting number of epitope slices. Each orientation remaining after the clustering process was checked to see if it involved an epitope slice, those that did not were discarded. This procedure resulted in around 300 orientations that involved the antigen epitope, were electrostatically feasible and were sterically favourable.

Procedure	HyHEL10	HyHEL5	D1.3	D1.3 Model
Store good fits	232,509	207,360	229,756	229,380
Cluster	1,587	1,791	1,627	1,837
Electrostatic Cutoff[1]	178 / 1,126	1,183 / 1,608	387 / 1,205	1,014 / 1,116
Antigen Maps[2]	$\frac{59}{432}$	$\frac{86}{432}$	$\frac{58}{423}$	$\frac{58}{432}$
Restrict to Antigen Epitope[1]	31 / 224	233 / 346	72 / 225	106 / 184
Single Distance Constraint[1]	4 / 21	31/ 43	5 / 25	9 / 15
Epitope and Distance Constraints[1]	3 / 18	30 / 40	5 / 25	9 / 15
R.M.S[3]. Overall	3.4	7.5	1.7	11.4
R.M.S.[3] in Interface Region	2.2	3.5	1.9	4.8

Table 3.1

The number of orientations remaining, and the position of the most native-like structure, at key stages of the docking algorithm.

[1] Results presented in the form n/m show that the best structure occurred at position n in a list of m.

[2] The number of antigen slices that contain the epitope, from an initial number of 432.

[3] All R.M.S deviations are calculated using the Cα atoms only

Structure	Translation (Å)	Rotation (°)	R.M.S. Whole	R.M.S. Interface
HyHEL-10	(-1,-2,-2)	(2,-5,1)	3.4	2.2
HyHEL-5	(0,-6,0)	(3,16,-19)	7.5	3.4
D1.3	(0,0,0)	(-7,6,0)	1.7	1.9
D1.3Model	(8,-1,1)	(-30,-19,-22)	11.4	4.7

Table 3.2

The agreement between the resultant structure of the docking algorithm and the true complex structure. Three measures are given, a set of translation and rotation vectors, a r.m.s. deviation of Cα atoms between the entire lysozyme molecules (native and predicted), and a r.m.s. deviation of Cα atoms in the interface region.

Complex		Interface Residues[1]
HyHEL-10	Antibody	30L 31L 32L 50L 53L 91L 92L 96L 30H 31H 32H 33H 50H 52H 53H 54H 56H 58H 98H
	Lysozyme	15 16 20 21 63 73 75 89 93 96 97 98 100 101 102
D1.3	Antibody	30L 32L 49L 50L 91L 92L 93L 30H 31H 32H 52H 53H 54H 99H 100H 101H 102H
	Lysozyme	18 19 22 23 24 25 27 116 117118 119 120 121 124 129
HyHEL-5	Antibody	31L 32L 50L 91L 92L 93L 95L 33H 35H 47H 50H 55H 57H 58H 59H 95H 97H
	Lysozyme	41 43 44 45 46 47 48 49 53 67 68 69 70 84

Table 3.3

Definition of the epitopes and interface region for each complex.

[1] Data taken from Davies & Padlan (1990)

Finally, the data were then filtered using a single, loose distance constraint. Residues that interacted strongly across the interface, e.g. forming a salt bridge, were chosen as distance constraints. This type of information might be available in realistic applications, where the complex structure is not known. The only salt bridge across the interface of the HyHEL-10 complex is the Asp 32H/Lys 97 pair; hence this interaction was chosen as the constraint. HyHEL-5 has 3 strong electrostatic interactions. Glu 50H is involved in two of these, and so the Arg 45/Glu 50H pair was chosen to be the HyHEL-5 constraint. D1.3 has no salt bridges. However, the Asp 100H residue makes a large, negative contribution to the change in the Gibbs free energy on forming the complex (Novotny, *et al.*, 1989), hence the chosen constraint was between Ser 24 and Asp 100H. The constraint was applied to both the epitope-restricted and global orientations. The margin of error allowed in the bond distance was so large that the distance constraint was not always strong enough to eliminate all non-epitope solutions. This shows that the constraint used demanded only that the residues involved were both in the interface region and that the possibility of an interaction existed. A hard distance constraint that demanded a standard bonding distance could not be used since the precise position of each residues side-chain would be too important. This type of constraint would be intolerant of changes due to induced fit, particularly side-chain rotation.

3.6. Individual Results

In each simulation a structure was found that corresponded to the crystallographic solution. The results of the random trials (Section 3.4) are used to assess the significance of the results.

3.6.1 HyHEL-10

After all constraints had been applied the best solution for HyHEL-10 was found third in a list of eighteen (Table 3.1). This configuration strongly reproduces the central helix position of lysozyme which is crucial to binding. Figure 3.3 shows the calculated lysozyme position (in green) docked with the antibody (in blue) and with the complexed lysozyme (in yellow) superimposed as a reference. Figure 3.3a shows a view looking down the helical axis. The r.m.s. deviation of Cα atoms in the interface region is 2.2Å. This figure is broadly in line with the degree of error inherent in the algorithm and the data. The global measures show an overall translation of 3Å combined with Euler rotations of (2°, -5°, 1°). The view chosen shows that the translation and rotation have partially cancelled out in the interface region. Table 3.4 compares the Cα-Cα distances in the predicted structure to those in the actual structure for all contacting residues. The table shows the difference between these two distances to be typically less than 1Å. The predicted interaction distance is predominantly larger than the actual one, showing that the algorithm tends to leave extra space between residues rather than packing too tightly, this feature further allows for side-chain movement during the docking process.

3.6.2 D1.3

The best solution for D1.3 was fifth in a list of 25. The calculated configuration has the correct global translation and so differs by only rotations from the crystallographic form. The interface and global r.m.s. deviations are 1.9Å and 1.7Å respectively. Figure 3.4 shows a comparison of the calculated complex with the correct one, a close correspondence between the Cα chains can be

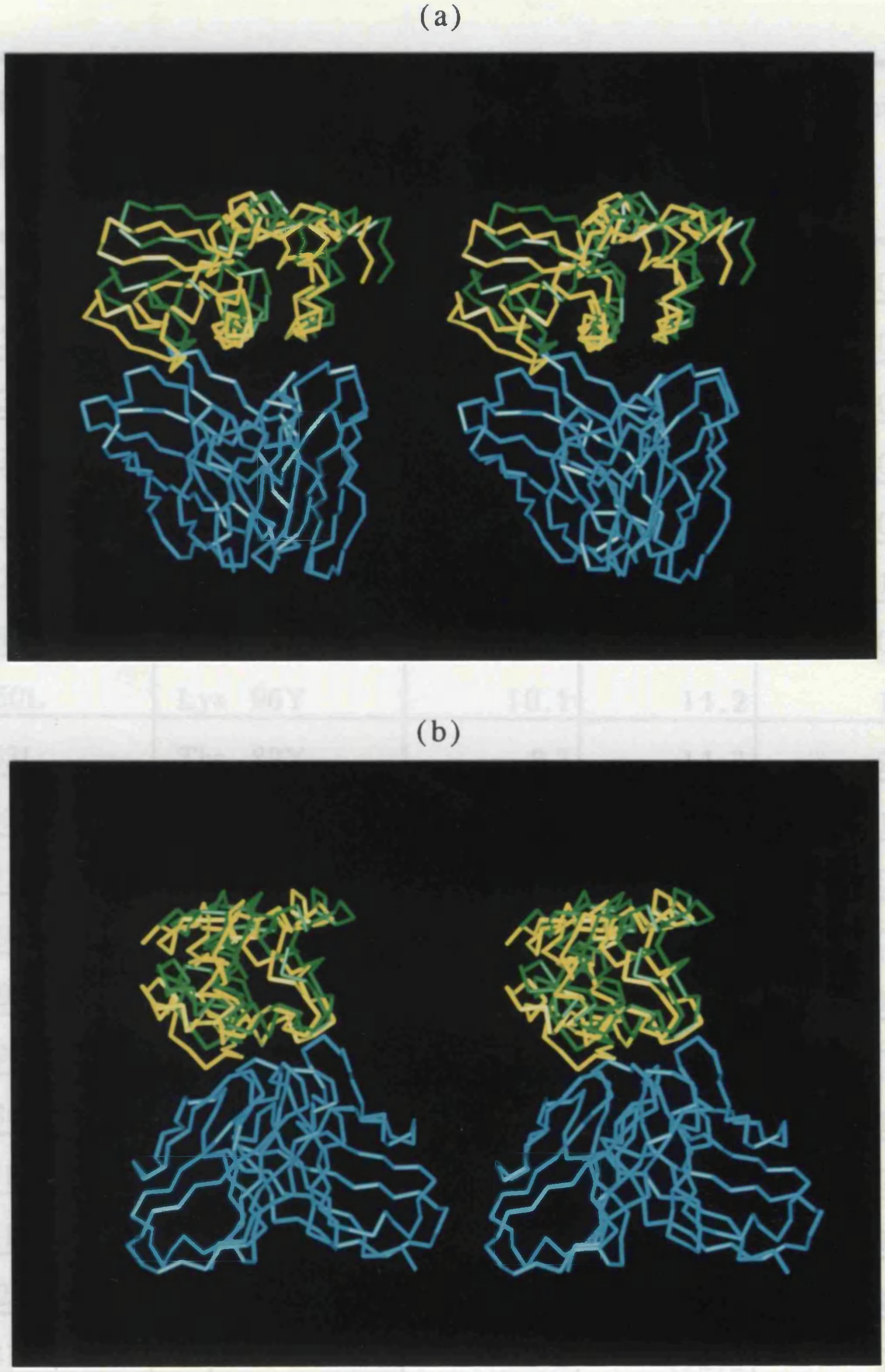


Figure 3.3

Two stereo views comparing the predicted structure of HyHEL-10 to the known crystallographic structure. The Cα traces of HyHEL-10 antibody (cyan) and the complexed form of lysozyme (yellow) are shown; the green Cα-trace shows the lysozyme orientation resulting from the steric matching procedure.

HyHEL-10 Residue	Antigen Residue	Actual Distance (Å)	Predicted Distance (Å)	Difference (Å)
Gly 30L	Gly 16Y	4.7	4.1	-0.6
Asn 31L	His 15Y	6.4	6.5	0.1
Asn 31L	Gly 16Y	4.8	5.5	0.7
Asn 31L	Lys 96Y	11.3	11.6	0.3
Asn 32L	Gly 16Y	6.6	7.4	0.8
Asn 32L	Tyr 20Y	8.7	9.7	1.0
Tyr 50L	Asn 93Y	8.4	8.7	0.3
Tyr 50L	Lys 96Y	10.1	11.2	1.1
Gln 53L	Thr 89Y	9.3	11.3	2.0
Gln 53L	Asn 93Y	10.3	11.3	1.0
Ser 91L	Tyr 20Y	8.8	9.5	0.7
Asn 92L	Tyr 20Y	5.7	6.0	0.3
Asn 92L	Arg 21Y	5.5	6.7	1.2
Tyr 96L	Arg 21Y	10.5	11.4	0.9
Thr 30H	Arg 73Y	7.9	9.2	1.3
Ser 31H	Arg 73Y	7.8	8.2	0.4
Ser 31H	Leu 75Y	7.0	8.3	1.3
Asp 32H *	Lys 97Y *	10.3	11.5	1.2
Tyr 33H	Trp 63Y	12.2	12.8	0.4
Tyr 33H	Lys 97Y	9.1	10.1	1.0
Tyr 33H	Ile 98Y	9.7	11.6	1.9
Tyr 33H	Ser 100Y	9.2	11.0	1.8
Tyr 33H	Asp 101Y	8.6	8.7	-0.1
Tyr 50H	Arg 21Y	15.0	15.7	0.7

Tyr 50H	Ser 100Y	10.3	11.3	1.0
Ser 52H	Asp 101Y	5.5	6.1	0.6
Tyr 53H	Trp 63Y	11.4	11.5	0.1
Tyr 53H	Leu 75Y	9.0	9.4	0.4
Tyr 53H	Asp 101Y	7.0	7.7	0.7
Ser 54H	Asp 101Y	7.8	8.9	1.1
Ser 56H	Asp 101Y	5.8	7.0	1.2
Ser 56H	Gly 102Y	5.8	5.8	0.0
Tyr 58H	Arg 21Y	12.6	12.7	0.1
Tyr 58H	Ser 100Y	9.7	9.8	0.1
Tyr 58H	Gly 102Y	8.7	8.8	0.1
Trp 98H	Arg 21Y	14.5	15.9	1.4
Trp 98H	Lys 97Y	8.4	9.7	1.3
Trp 98H	Ser 100Y	9.5	10.9	1.4

Table 3.4

Residue/residue interaction distances for the crystallographic and predicted HyHEL-10 complex. * denotes the residue pair used as a distance constraint.

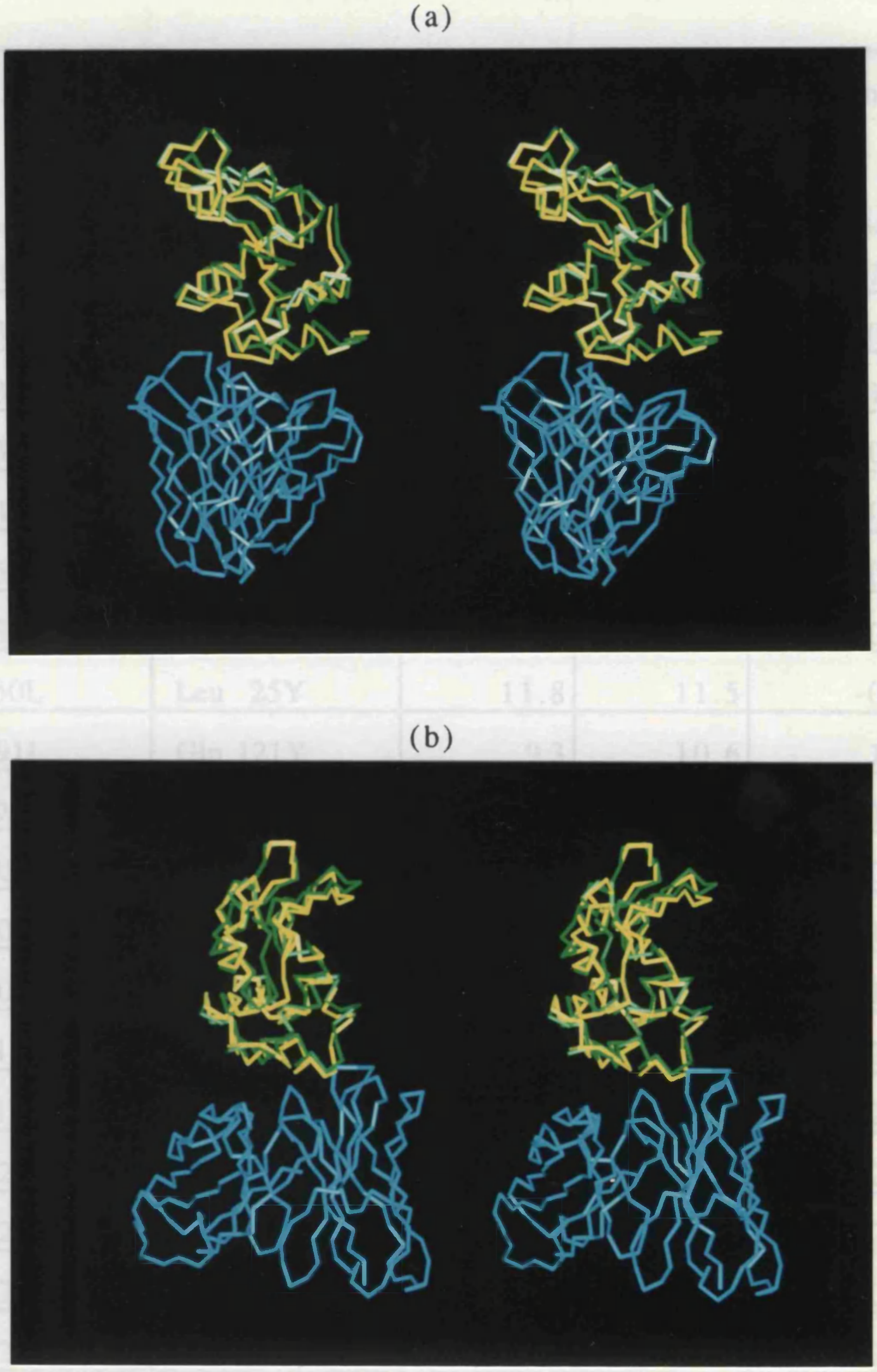


Figure 3.4

Two stereo views comparing the predicted structure of D1.3 to the known crystallographic structure. The Cα traces of D1.3 antibody (cyan) and the complexed form of lysozyme (yellow) are shown; the green Cα-trace shows the lysozyme orientation resulting from the steric matching procedure.

D1.3 Residue	Antigen Residue	Actual Distance (Å)	Predicted Distance (Å)	Difference (Å)
His 30L	Leu 129Y	9.2	8.2	-1.0
Tyr 32L	Leu 25Y	11.5	11.4	-0.1
Tyr 32L	Gln 121Y	9.7	10.8	1.1
Tyr 32L	Ile 124Y	12.1	13.0	0.9
Tyr 49L	Gly 22Y	8.7	9.4	0.7
Tyr 50L	Asp 18Y	9.7	9.7	0.0
Tyr 50L	Asn 19Y	8.1	7.8	-0.3
Tyr 50L	Leu 25Y	11.8	11.5	-0.3
Phe 91L	Gln 121Y	9.3	10.6	1.3
Trp 92L	Ile 124Y	10.5	11.6	1.1
Ser 93L	Gln 121Y	7.8	8.2	0.4
Thr 30H	Lys 116Y	5.9	7.0	1.1
Thr 30H	Gly 117Y	4.9	6.0	1.1
Gly 31H	Lys 116Y	4.0	6.2	2.2
Gly 31H	Gly 117Y	3.4	5.8	2.4
Tyr 32H	Lys 116Y	7.3	9.3	2.0
Tyr 32H	Gly 117Y	5.2	7.7	2.5
Trp 52H	Gly 117Y	5.9	6.1	0.2
Trp 52H	Thr 118Y	5.6	7.7	2.1
Trp 52H	Asp 119Y	7.7	8.6	0.9
Gly 53H	Gly 117Y	4.9	5.1	0.2
Asp 54H	Gly 117Y	7.5	6.2	-1.3
Arg 99H	Arg 21Y	13.0	11.4	-1.6
Arg 99H	Gly 22Y	10.1	8.2	-1.9

Arg 99H	Tyr 23Y	9.9	9.1	-0.8
Asp 100H	Gly 22Y	8.1	6.2	-1.9
Asp 100H	Tyr 23Y	7.2	6.2	-1.0
Asp 100H *	Ser 24Y *	7.0	6.7	-0.3
Asp 100H	Asn 27Y	8.9	9.3	0.4
Tyr 101H	Thr 118Y	10.3	12.8	2.5
Tyr 101H	Asp 119Y	9.5	10.5	1.0
Tyr 101H	Val 120Y	9.5	11.0	1.5
Tyr 101H	Gln 121Y	8.6	10.3	1.7
Arg 102H	Asn 19Y	12.4	11.4	-1.0
Arg 102H	Gly 22Y	10.3	9.3	-1.0

Table 3.5

Residue/residue interaction distances for the crystallographic and predicted D1.3 complex. * denotes the residue pair used as a distance constraint.

seen. Table 3.5 gives the contact distances, again showing a typical error of less than 1Å and a tendency to loose packing.

The steric matching part of the algorithm, which produces possible docking orientations, was developed using the HyHEL-10 structure only (Section 2.13). The subsequent filtering parts of the algorithm were also tested against the D1.3 complex, but this section of the algorithm merely rejects infeasible structures - it cannot improve on the structures generated by steric matching. All the measures quoted show the D1.3 result to be the best agreement between predicted and expected docking orientation. This suggests that the parameters and potential derived during the HyHEL-10 testing phase are applicable to general antibody/antigen docking, and are not specific to the HyHEL-10 system.

3.6.3 HyHEL-5

The prediction for HyHEL-5 is the poorest of the crystallographic complexes. A larger number of solutions remain after the filtering process than for the previous complexes. This is partially because the electrostatic filtering stage discards far fewer sterically feasible solutions - only 10% compared with 26% for D1.3 and 30% for HyHEL-10. A further investigation of this effect will be presented in Section 3.8.

The solution is 30th from a list of 40, and the global r.m.s. deviation is 7.5Å. This result appears little better than random considering the constraints applied. However the portion of the structure examined by the program is much closer to the true solution. The interface deviation is 3.5Å. Figure 3.5 shows the structure comparison. Table 3.6 lists the contact differences, here the

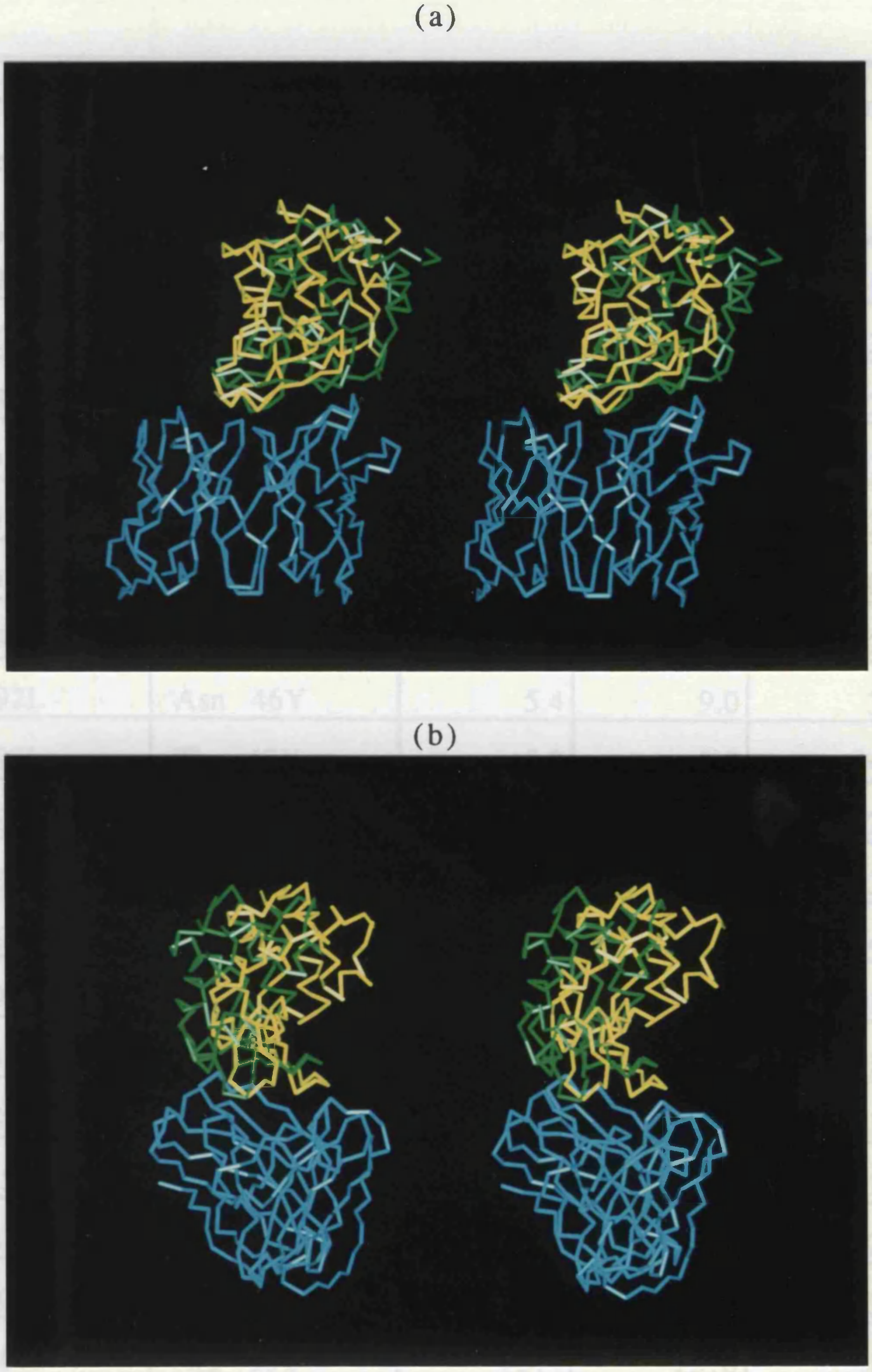


Figure 3.5

Two stereo views comparing the predicted structure of HyHEL-5 to the known crystallographic structure. The Cα traces of HyHEL-5 antibody (cyan) and the complexed form of lysozyme (yellow) are shown; the green Cα-trace shows the lysozyme orientation resulting from the steric matching procedure.

HyHEL-5 Residue	Antigen Residue	Actual Distance (Å)	Predicted Distance (Å)	Difference (Å)
Tyr 31L	Asp 48Y	7.9	9.8	1.9
Met 32L	Pro 70Y	10.8	14.3	3.5
Thr 50L	Pro 70Y	9.9	12.1	2.2
Gly 91L	Arg 45Y	6.9	7.9	1.0
Gly 91L	Gly 49Y	5.8	8.6	2.8
Gly 91L	Arg 68Y	12.8	13.6	0.8
Arg 92L	Arg 45Y	5.1	7.5	2.4
Arg 92L	Asn 46Y	5.4	9.0	3.6
Arg 92L	Thr 47Y	5.2	9.2	4.0
Asn 93L	Arg 45Y	8.0	10.7	2.7
Asn 93L	Asn 46Y	9.1	12.6	3.5
Asn 93L	Thr 47Y	8.3	12.5	4.2
Thr 95L	Arg 45Y	13.1	15.0	1.9
Trp 33H	Tyr 53Y	13.0	11.3	-1.7
Trp 33H	Arg 68Y	9.4	10.8	1.4
Glu 35H	Arg 68Y	12.5	13.8	1.3
Trp 47H	Arg 45Y	13.2	15.6	2.4
Glu 50H *	Arg 45Y *	10.4	12.4	2.0
Glu 50H	Arg 68Y	13.4	15.0	1.6
Ser 55H	Gln 41Y	6.8	3.4	-3.4
Ser 55H	Leu 84Y	6.5	4.1	-2.4
Ser 57H	Gln 41Y	5.6	3.6	-2.0
Ser 57H	Thr 43Y	6.2	6.4	0.2
Thr 58H	Thr 43Y	6.8	7.5	0.7

Asn 59H	Thr 43Y	7.0	8.1	1.1
Asn 59H	Asn 44Y	7.0	8.6	1.6
Tyr 95H	Arg 68Y	19.0	19.8	0.8
Leu 97H	Gly 67Y	15.0	14.3	-0.7
Leu 97H	Arg 68Y	13.2	13.9	0.7
Leu 97H	Thr 69Y	16.2	17.5	1.3
Leu 97H	Pro 70Y	17.8	19.9	2.1

Table 3.6

Residue/residue interaction distances for the crystallographic and predicted HyHEL-5 complex. * denotes the residue pair used as a distance constraint.

differences have risen slightly and are generally in the range 1.5 - 2.5 Å.

3.6.4 D1.3 Model

The predicted structure of D1.3 produces a roughly native-like solution ninth from a list of 15. This solution has an r.m.s deviation of 4.8Å in the interface region. Figure 3.6 shows that there is a degree of correspondence between the predicted and actual solution. Table 3.7 lists the residue/residue contact distances, many of the crucial interactions are still present. In the random r.m.s. distribution for D1.3 none of the 10,000 orientations produced a r.m.s. lower than this, the closest being 5.5Å. Since, before constraints were applied, only 1,100 orientations remained the algorithm has performed significantly better than random.

3.7. Incorrect Structures

For each system the algorithm identified several non-native structures which had a high degree of steric complementarity. Such steric matches are found for all sections of the lysozyme antigen. A selection of such non-native structures are shown in Figure 3.7. For each system two structures are shown, the one with most favourable steric score after the electrostatic filtering stage, and the one with the most favourable score after applying the distance constraint. The structures produced after the electrostatic filtering stage are not restricted to the correct epitope and so several different areas of lysozyme are docked. The algorithm shows a preference for broad, flat interfaces such as the HyHEL-10 interface. The incorrect structures are generally well packed, with a tendency for the two

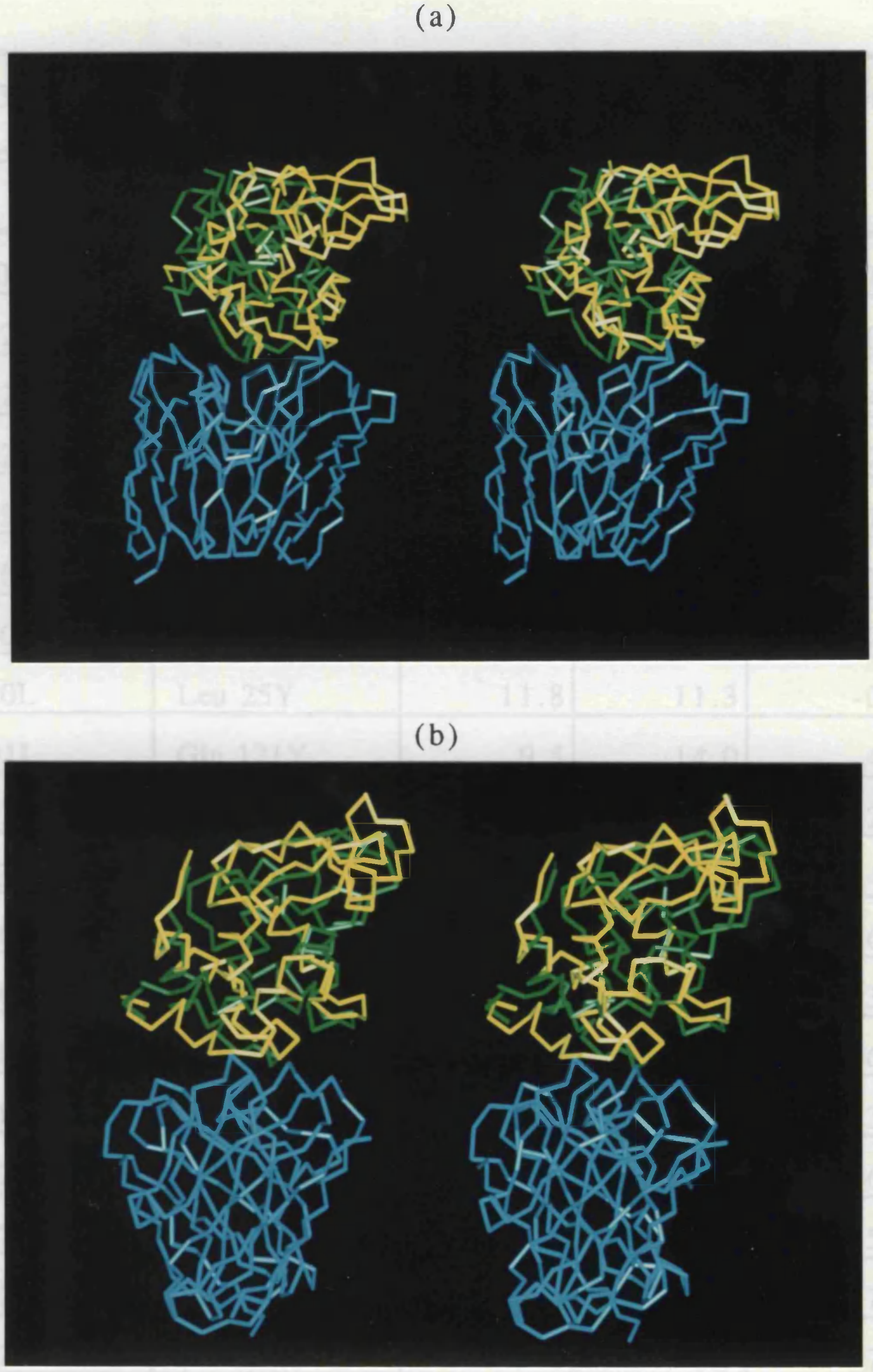


Figure 3.6

Two stereo views comparing the predicted structure of D1.3 to the known crystallographic structure. The Cα traces of D1.3 antibody (cyan) and the complexed form of lysozyme (yellow) are shown; the green Cα-trace shows the lysozyme orientation resulting from the steric matching procedure, using the modelled form of D1.3 as a docking target.

D1.3 Model Residue	Antigen Residue	Actual Distance (Å)	Predicted Distance (Å)	Difference (Å)
His 30L	Leu 129Y	9.2	8.0	-1.2
Tyr 32L	Leu 25Y	11.6	11.4	-0.2
Tyr 32L	Gln 121Y	9.9	13.9	4.0
Tyr 32L	Ile 124Y	12.3	15.0	2.7
Tyr 49L	Gly 22Y	8.7	7.5	-1.2
Tyr 50L	Asp 18Y	10.2	8.8	-1.4
Tyr 50L	Asn 19Y	8.6	6.3	-2.3
Tyr 50L	Leu 25Y	11.8	11.3	-0.5
Phe 91L	Gln 121Y	9.5	14.0	4.5
Trp 92L	Ile 124Y	10.6	13.	2.9
Ser 93L	Gln 121Y	8.8	13.7	4.9
Thr 30H	Lys 116Y	9.0	15.3	6.3
Thr 30H	Gly 117Y	7.6	11.5	3.9
Gly 31H	Lys 116Y	5.4	11.9	6.5
Gly 31H	Gly 117Y	4.4	8.1	3.7
Tyr 32H	Lys 116Y	7.1	13.6	6.5
Tyr 32H	Gly 117Y	4.3	10.2	5.9
Trp 52H	Gly 117Y	7.3	14.9	7.6
Trp 52H	Thr 118Y	6.1	13.9	7.8
Trp 52H	Asp 119Y	7.6	11.4	3.8
Gly 53H	Gly 117Y	6.0	12.8	6.8
Asp 54H	Gly 117Y	8.0	14.3	6.3
Arg 99H	Arg 21Y	13.5	14.1	0.6
Arg 99H	Gly 22Y	10.4	10.4	0.0

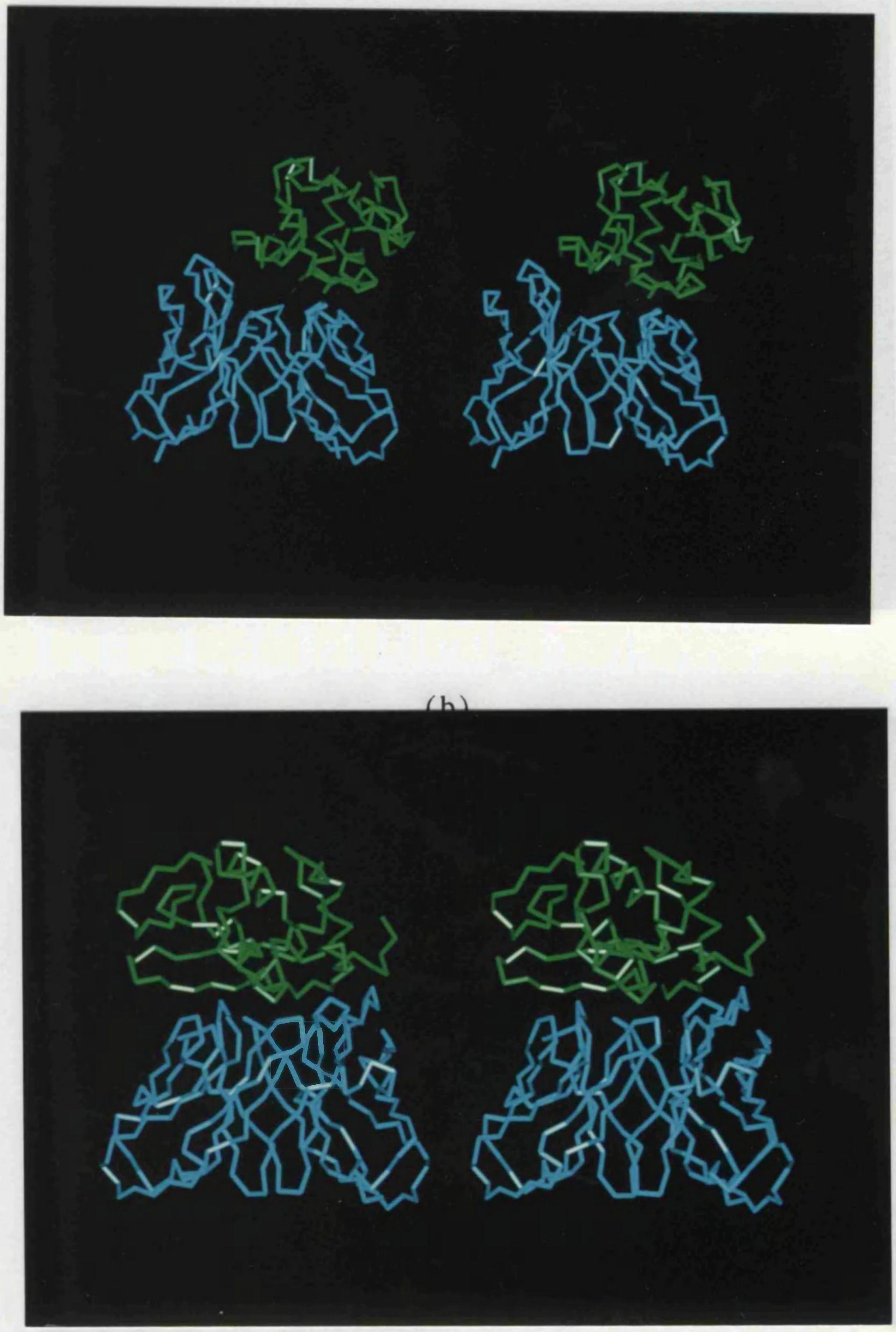
Arg 99H	Tyr 23Y	10.2	11.1	0.9
Asp 100H	Gly 22Y	8.4	8.4	0.0
Asp 100H	Tyr 23Y	7.1	8.1	1.0
Asp 100H *	Ser 24Y *	6.7	7.0	0.3
Asp 100H	Asn 27Y	8.5	10.2	1.7
Tyr 101H	Thr 118Y	10.5	15.4	4.9
Tyr 101H	Asp 119Y	9.4	12.6	3.2
Tyr 101H	Val 120	9.4	12.7	3.3
Tyr 101H	Gln 121Y	8.2	11.1	2.9
Arg 102H	Asn 119Y	12.7	11.2	-1.5
Arg 102H	Gly 122Y	11.0	10.3	-0.7

Table 3.7

A comparison of residue/residue contact distances in the D1.3 complex, using the D1.3 antibody model, and the predicted complex using the model as a docking target. * denotes the residue pair used as a distance constraint.

HyHEL-10 Non-native solutions

(a)



(b)

Figure 3.7

Stereo views of a sample of non-native structures which the DAPMatch algorithm scored favourably. For each antibody/lysozyme complex two incorrect structures are shown. The first view is the most favourable orientation remaining after clustering. No constraints have been applied and so the correct epitope of lysozyme need not be in contact with the antibody binding region (p.t.o.)

D1.3 non-native solutions

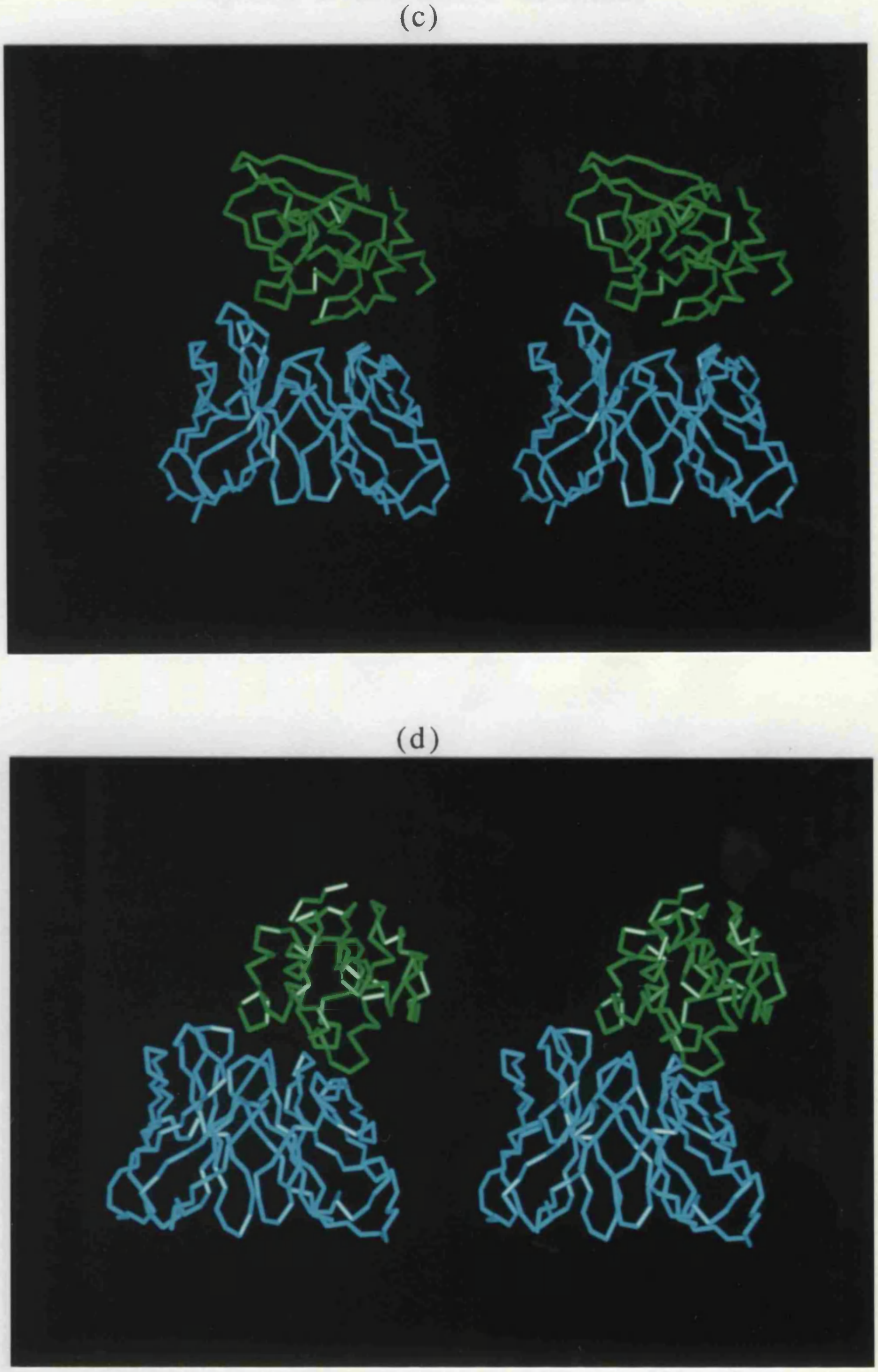


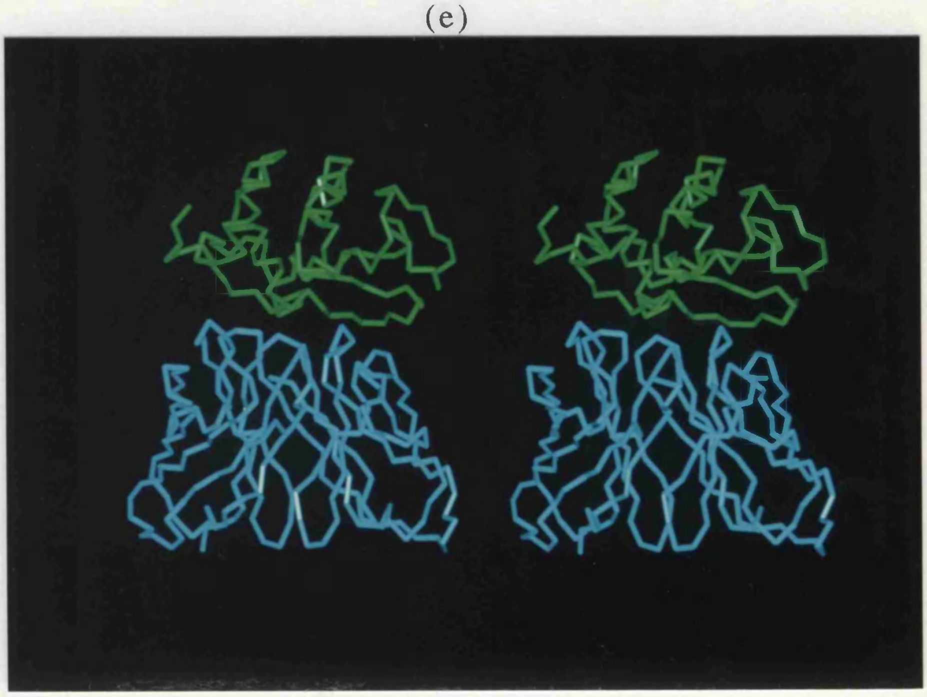
Figure 3.7 (cont)

The second view is of the most favourable orientation remaining after all the filtering procedures have been applied.

The views are labelled as follows; HyHEL-10 (a,b), D1.3 (c,d), HyHEL-5 (e,f) and D1.3 model (g,h). Although these structures are well packed and have reasonable electrostatic interactions some can be eliminated by eye (p.t.o)

HyHEL-5 non-native solutions

(e)



(f)

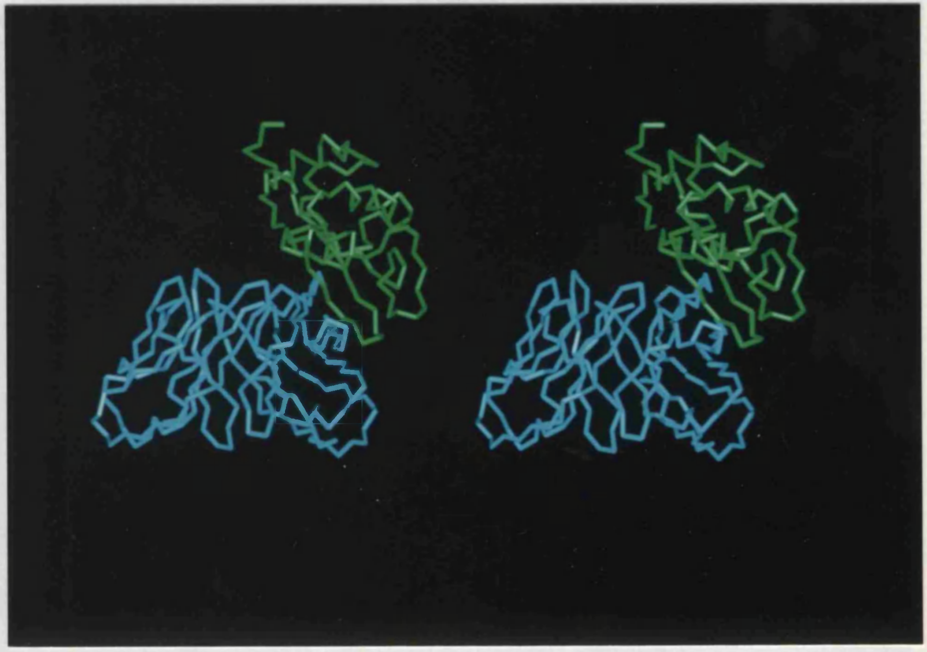
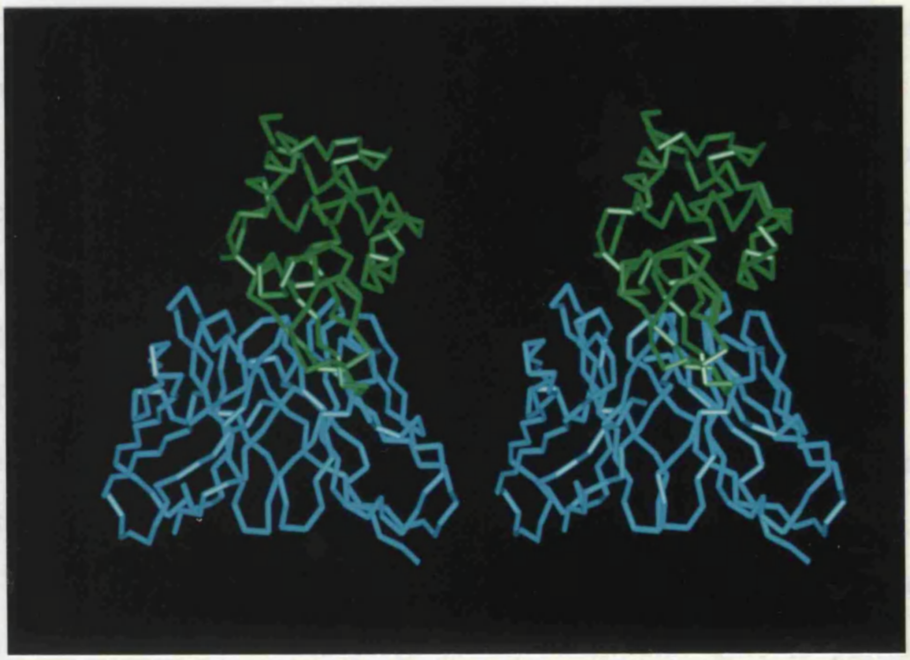


Figure 3.7 (cont)

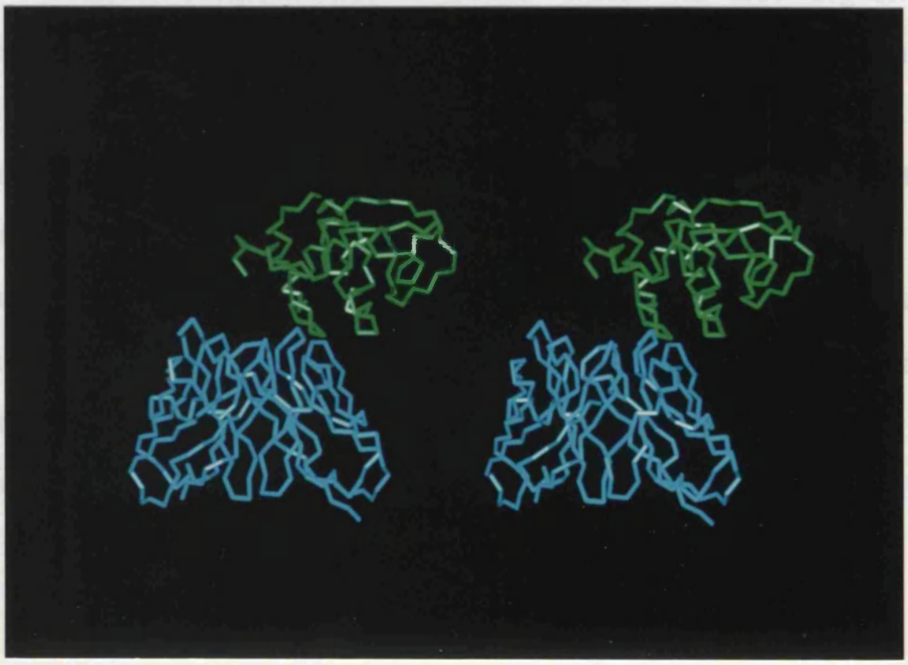
The most favourable constrained solution for HyHEL-5 (f) can easily be eliminated, since here lysozyme packs only against the antibody framework and the light chain CDRs.

D1.3 (Model) non-native solutions

(g)



(h)

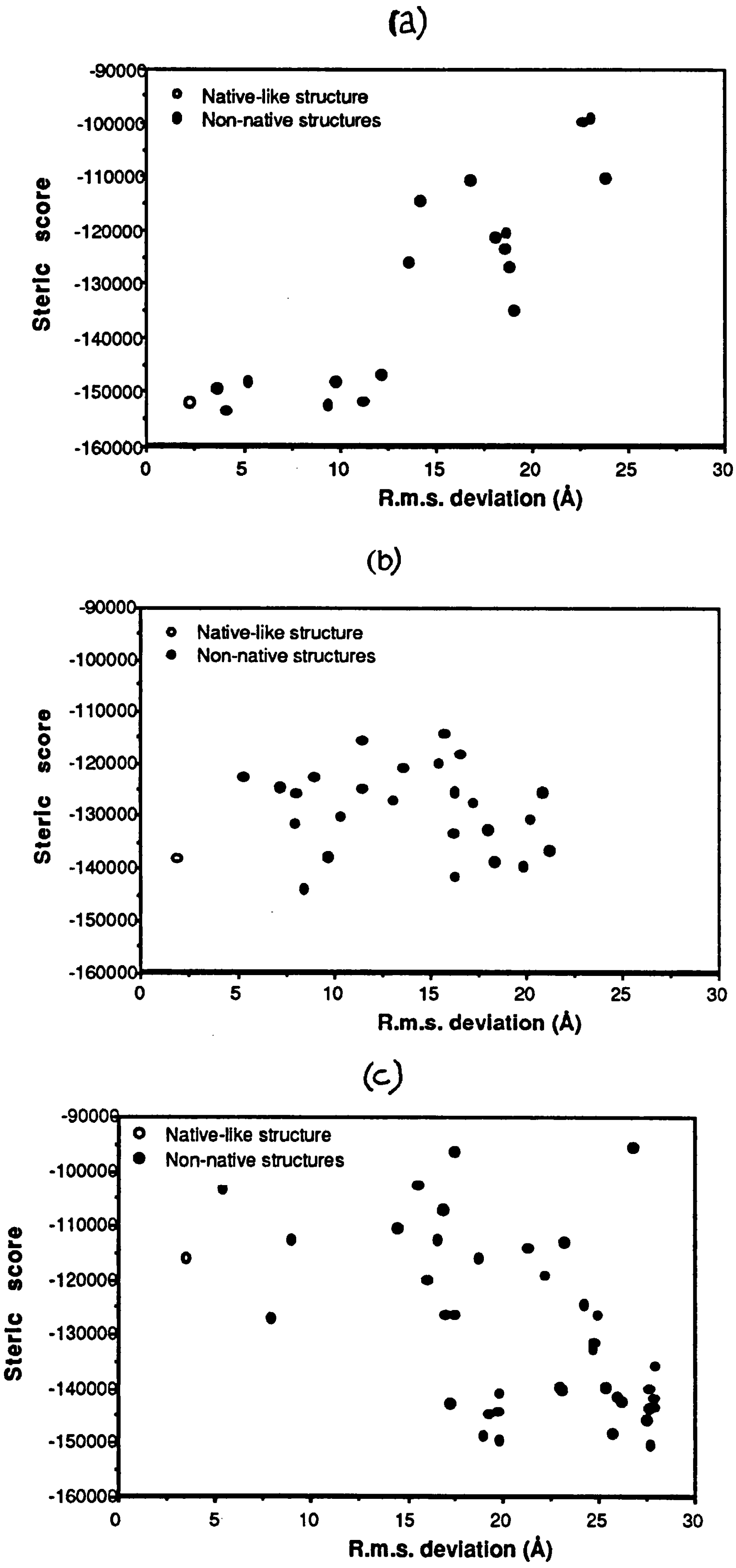


molecules to be slightly too far apart. This effect can be seen in Figure 3.7c where there is no interdigitation of side-chains between the antibody and antigen molecules. The antigen is usually sited above the antigen-binding region of the antibody, however for the distance constrained HyHEL-5 solution (Figure 3.7f) the lysozyme molecule binds with only the heavy chain of the antibody, and makes contact with many residues of the framework region.

Figure 3.8 shows the range of steric scores produced for each system. The score and r.m.s. deviation for each structure remaining after the distance constraint stage is plotted for each system. The graphs for D1.3 and HyHEL-10 show that the near-native structures have a steric score which is close to the minimum. This is particularly apparent for HyHEL-10 which has a broad spread of steric scores and the near-native solution has a score which is not significantly different from the minimum score. For the D1.3 model structure the spread of steric scores is smaller and the near-native score lies in the middle of the spread. The HyHEL-5 graph shows a cluster of structures with very high r.m.s. deviations from the native but which still have favourable steric scores. The structure which is closest to the native has a score on the least favourable half of the scale, and so further information would be necessary to identify this as a good docking orientation.

3.8. The Electrostatic Constraint and HyHEL-5 Docking

The electrostatic filter does not remove as many structures from the HyHEL-5 systems as it does for the other systems. This is because the relatively high number of electrostatic groups at the HyHEL-5 interface provides many favourable possibilities. In the


(a)
Native-like structure
Non-native structures
Steric score
-90000
-100000
-110000
-120000
-130000
-140000
-150000
-160000
0
5
10
15
20
25
30
R.m.s. deviation (Å)
(b)
Native-like structure
Non-native structures
Steric score
-90000
-100000
-110000
-120000
-130000
-140000
-150000
-160000
0
5
10
15
20
25
30
R.m.s. deviation (Å)
(c)
Native-like structure
Non-native structures
Steric score
-90000
-100000
-110000
-120000
-130000
-140000
-150000
-160000
0
5
10
15
20
25
30
R.m.s. deviation (Å)

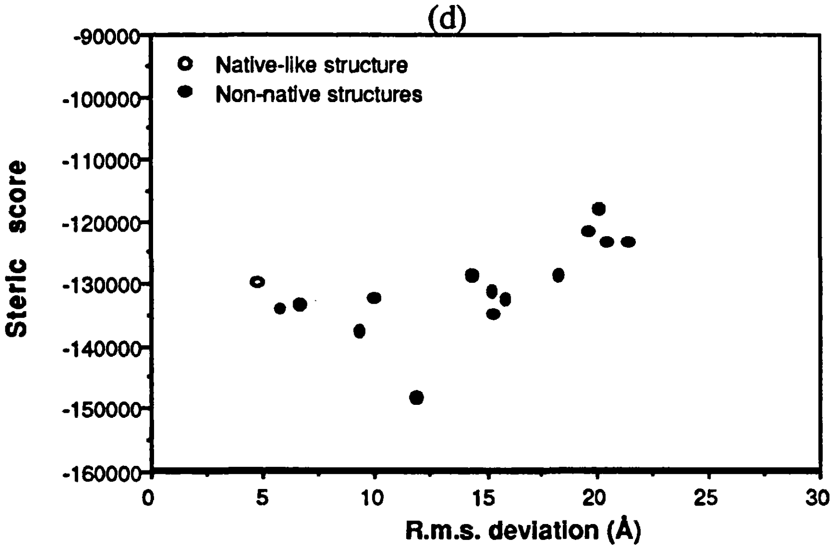


Figure 3.8

Graphs showing the variation of steric score with r.m.s. deviation from the native structure. Each point represents a structure that remained after all the filtering stages. The HyHEL-10 results are plotted in Figure 3.8a. They show strong correlation between score and r.m.s. The structure with the lowest r.m.s. deviation from the native structure has a steric score which is not significantly different from the minimum observed score. The results for the D1.3 system (Figure 3.8b) are similar, though the correlation is not as strong. The most native-like structure has a steric score close to the observed minimum. The results for HyHEL-5 (Figure 3.8c) and the D1.3 model (Figure 3.8d) both show that the most native-like solutions do not have favourable steric scores. This problem is particularly acute for HyHEL-5 which has a cluster of solutions with high r.m.s. deviations from the native but with apparently favourable steric scores.

crystallographic complex three salt bridges are observed between the antibody and lysozyme molecules. The electrostatic cut-off can be altered to take this high degree of electrostatic complementarity into account. When a stronger electrostatic cut-off was used on the HyHEL-5 system, this eliminated 1,432 of the 1,791 orientations that remained after clustering (Table 3.8). This was seven times more than the original cut-off which had only removed 182 orientations. With additional epitope and residue-residue contact constraints, a native-like orientation was then found 13th in a list of 14 (Table 3.8). Thus, far more selective results can be obtained if cut offs are chosen for associations dominated by electrostatics. Similarly, the steric complementarity requirements could be increased by altering the potential. This would favour the D1.3 and HyHEL-10 systems, but would exclude solutions for HyHEL-5 and the predicted D1.3 structure. This highlights the difficulty of obtaining parameters which are generally applicable from a limited set of test complexes. Benchmarking on only one or two of the three available structures can lead to highly specific methods. The steric matching scheme we present here was developed solely on the HyHEL-10 system, and the filtering stages were benchmarked using both the HyHEL-10 and D1.3 systems.

3.9. Discussion and Conclusion

The algorithm produced native-like conformations for the four antibody/antigen systems studied. However a large number of possible orientations were generated because the steric potential and electrostatic constraint used were not sufficiently selective. Only further elimination using biological knowledge reduced this number

Procedure	HyHEL-5
Store good fits	207,360
Cluster	1,791
Electrostatic Cutoff [1]	306 / 359
Antigen Maps [2]	$\frac{86}{432}$
Restrict to Antigen Epitope [1]	79 / 96
Single Distance Constraint [1]	13/ 14
Epitope and Distance Constraints [1]	13 / 14
R.M.S[3]. Overall	7.5
R.M.S.[3] of Epitope Region	3.5

Table 3.8

The HyHEL-5 results with a favourable electrostatic cutoff.

[1] Results presented in the form n/m show that the best structure occurred at position n in a list of m.

[2] The number of antigen slices that contain the epitope, from an initial number of 432.

[3] All R.M.S deviations are calculated using the Cα atoms only

to a manageable size. It is, however, likely that knowledge of this kind would be available for a system actively studied experimentally. The algorithm can only be applied to antibodies and antigens of known structure and so there would be knowledge of surface residues and their relative positions that can guide experiments to identify epitopes and specific antibody/antigen residue contacts.

It is difficult to compare the results of different docking methods. The current approaches yield a rank ordered list and ~~one~~ assess how far down the list a good approximation to the true docking occurs. First the number of alternate docking will increase as the docking search becomes spatially finer. More importantly, as the number of systems successfully studied during the development of the algorithm increases, the criteria of selection tend to be looser and the results of the algorithm are less selective. There have been two recent docking studies (Cherfils, *et al.*, 1991 ; Jiang and Kim, 1991). Jiang & Kim (1991) apply their method to only one antibody/lysozyme complex (HyHEL-5), successfully identifying a native-like solution 12th from a list of 15. Although their method achieves similar results for other protein/protein interactions it is impossible to evaluate it as a general antibody/antigen docking algorithm without obtaining results for the D1.3 and HyHEL-10 systems. Simulated annealing, as used by Cherfils *et al.* (1991), identifies the native conformation as the global energy minimum for the HyHEL-5 system, even when the uncomplexed form of lysozyme is used, but does not clearly identify a native conformation for either D1.3 or HyHEL-10. However, neither of these studies have attempted to dock a predicted antibody structure to an antigen.

The high degree of electrostatic complementarity in the HyHEL-5 system aids methods with a strong electrostatic component. The DAPMatch method concentrates first upon steric complementarity, only later removing electrostatically infeasible solutions. Section 3.8 showed the extent to which the algorithm could be altered for a single antibody complex to obtain a more selective algorithm.

The strength of the distance constraint was carefully selected. Too tolerant and the constraint would lose selectivity. Too harsh and the constraint would become unrealistic in comparison with the degree of error in the method, and would hence risk eliminating a native-like orientation. If the steric matching section of the algorithm is unsuccessful, then the structures remaining after the distance constraint has been applied will all have reasonable contact distances for the residue pair used as a constraint, but inaccurate distances for all other pairs. In all the distance tables (Tables 3.4, 3.5, 3.6 and 3.7) the residue pair used as a distance constraint has been highlighted with an asterix. The differences seen display the loose nature of the constraint. The error for these pairs is never abnormally low. In the case of HyHEL-10 the distance error for the distance constraint pair is actually one of the poorest. This suggests that the steric matching section of the algorithm is correctly identifying reasonable surface interactions and that the distance constraint parameters have been set at reasonable values.

The crystal structure of uncomplexed D1.3 antibody has been solved and a comparison with the complexed structure made (Bhat, *et al.*, 1990). Antigen binding introduces no major conformational changes to the V_L and V_H individually; a Cα r.m.s. deviation of 0.37Å is reported. However, there is a bulk rearrangement of the two domains, resulting in the V_L domain moving 0.5Å closer to the antigen binding site, relative to the V_H domain (or the V_H moving 0.7Å closer relative to V_L). This change is large in comparison with the expected crystallographic error (0.3Å), and may be too large to be accomodated by the DAPMatch soft potential. The D1.3 structure is, as yet, the only antibody structure to be solved in both complexed and uncomplexed forms and it may be that the observed movement is atypical, or that the rearrangement follows a predictable pattern.

Recent progress in antibody modelling (Chothia and Lesk, 1987; Martin, *et al.*, 1989) suggests it may soon be possible routinely to predict an antibody structure from its sequence with confidence and accuracy. The use of sequence homology for modelling a protein of unknown structure from one of known structure is widespread e.g. Blundell *et al* .(1987). Docking studies could, therefore, be undertaken using protein models of both the antibody and the antigen. These models would be sufficiently accurate for soft potential docking methods to produce meaningful results. Such a technique would allow fast and simple evaluation of antibody design and alteration, useful for antibody engineering experiments (Jones, *et al.*, 1986; Pollack, *et al.*, 1988; Riechmann, *et al.*, 1988; Roberts, *et al.*, 1990; Shokat, *et al.*, 1989).

Chapter 4

A Practical Application of DAPMatch to the modelling of the High Molecular Weight Epidermal Growth Factor complex

4.1 Synopsis

This chapter discusses the use of the DAPMatch algorithm to predict a complex of epidermal growth factor (EGF) and epidermal growth factor binding protein (EGFBP). The full complex is formed from two EGF molecules and two EGFBP molecules and is known as high molecular weight EGF (HMWEGF). The structure of this complex has not yet been solved, by either crystallography or N.M.R. This work was undertaken in collaboration with Dr B. Bax (Birkbeck College, University of London) who modelled the EGF binding protein using the structures of tonin and kallikrein. The DAPMatch program was used to suggest possible modes of binding. Careful examination of the data revealed a problem with the protein structures used, details of this problem are given and its solution is described. A variety of biochemical data was used to narrow the search and a single binding orientation was chosen. Since the observed structure of the complex is not known, the result presented in this chapter is the proposed model of the complex.

4.2 Introduction

4.2.1 Epidermal Growth Factor

Epidermal growth factor (EGF) was one of the first growth factors to be discovered and has since been widely studied, it stimulates the division of epidermal, epithelial and connective tissue cells. EGF transmits a signal to the interior of a cell by binding to EGF receptors in the cell membrane and inducing structural changes to the receptor. EGF is found in large quantities in the mouse submandibular gland, where it exists as a high molecular weight complex (HMWEGF). This complex appears to consist of two EGF molecules bound to two molecules of a much larger protein, known as EGF binding protein (EGFBP). EGFBP has been sequenced (Blaber, *et al.*, 1987) and thereby identified as a member of serine proteinase family, with close homology to the protein kallikrein. Three glandular kallikreins were suspected of binding EGF, and were called EGFBP types A,B and C. In fact, only EGFBP type C has been shown to bind EGF (Isackson, *et al.*, 1987), in this chapter type C will be referred to as simply EGFBP.

An N.M.R. structure of EGF has been determined (Montelione, *et al.*, 1992). The structure of EGFBP has been modelled by Dr B. Bax and Ms G. Ferguson using the program COMPOSER (Sutcliffe, *et al.*, 1987a; Sutcliffe and Hayes, 1987; Sutcliffe, *et al.*, 1987b) and the graphics package SYBYL (Tripos Associates inc.). The known structures of porcine pancreatic kallikrein A (PPK) (Bode, *et al.*, 1983) and rat submaxillary gland tonin (Fujinaga and James, 1987) were used as the basis for the modelling procedure. These structures are almost identical, except for seven structurally variable regions (see Section 1.4). For each of these loops a comparison of the EGFBP sequence and the tonin and kallikrein sequences was made, and the most homologous loop was used in the EGFBP structure. In each case a loop of the correct length, size and properties was found in either tonin or kallikrien and it was not necessary to make a more general search of known protein structures. The model was then energy minimised to remove any short range clashes between atoms.

4.2.2 Application of the DAPMatch algorithm

The DAPMatch program was used to examine the docking of EGF to EGFBP. The N.M.R. structure of EGF and the model of EGFBP were the essential starting point for this investigation. As well as these structures, constraints were necessary before the DAPMatch program could be used with confidence. Several pieces of biochemical data were known,

i. The N-terminal domain of EGFBP interacts with EGF (Blaber *et al.* , unpublished results).

ii. The C-terminal residues of EGF are retained in the active site of EGFBP. This is suggested by the fact that the removal of the C-terminal arginine residue of EGF prevents formation of the HMWEGF complex (Server, *et al.*, 1976).

iii. HMWEGF is known to be formed from two EGF/EGFBP complexes (Taylor, *et al.*, 1974). A two-fold axis of symmetry is therefore almost certain to be present in HMWEGF.

iv. Neither EGFBP nor EGF normally exists as dimers. This implies that the full complex can only be formed after independent EGF/EGFBP complexes have been formed.

These conditions suggest a complex of the form illustrated in Figure 4.1. Here the first step of forming the complex is the interaction of the C-terminus of EGF(A1) with EGFBP(B1). This binding is weak, but can be stabilised by the presence of another

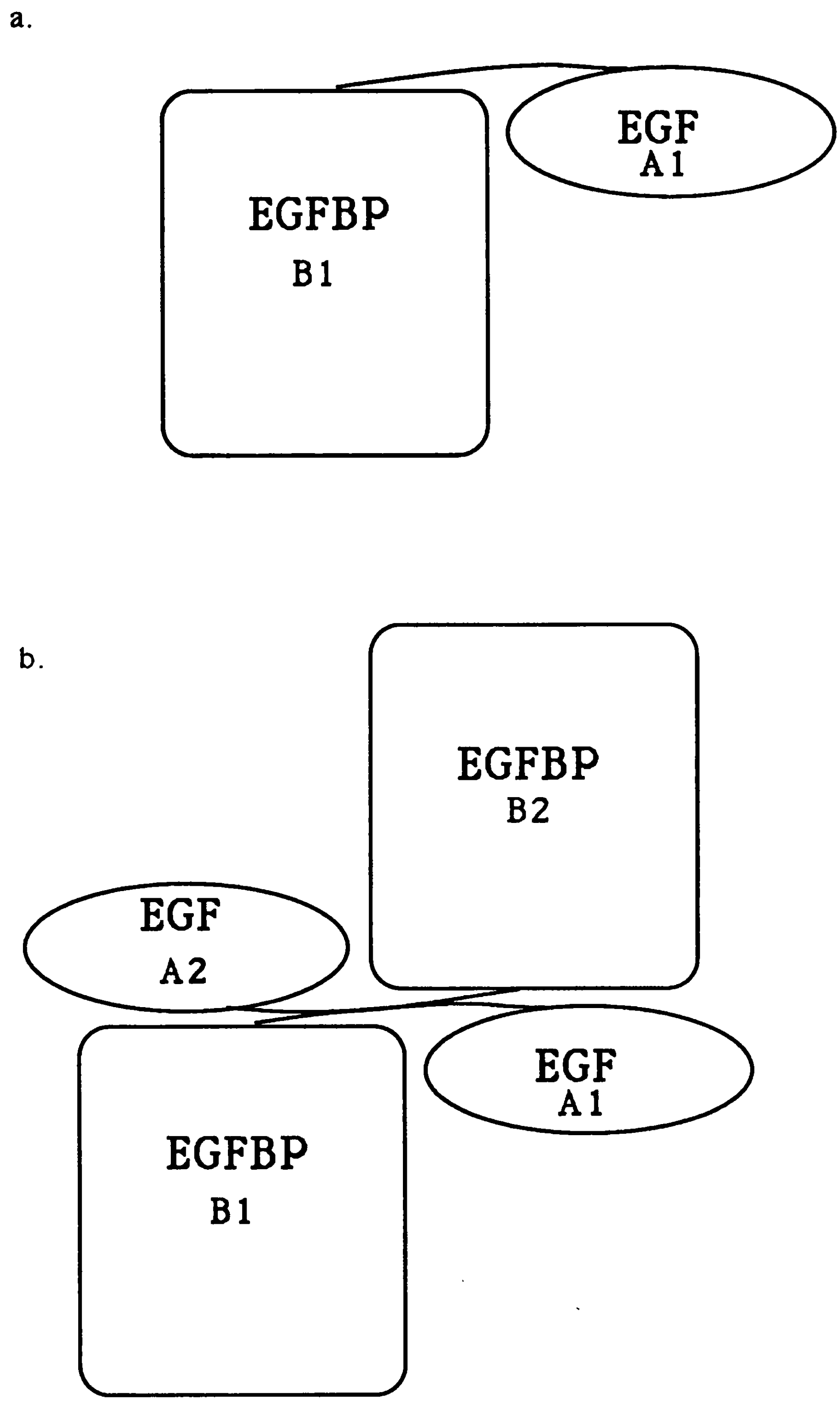


Figure 4.1

A schematic binding model for the HMWEGF complex. The first step (a) is a relatively unstable association between an EGF molecule (A1) and and EGFBP molecule (B1), involving the binding of the C-terminal arm of EGF(A1) to the specificity pocket of EGFBP(B1). Two such dimers come together (b) to form the final, symmetrical, HWMEGF complex.

EGF/EGFBP complex (A2/B2). The final, stable HMWEGF complex involves the binding of EGF(A1) with EGFBP(B2) and, by symmetry, EGF(A2) and EGFBP(B1). Hence each EGF molecule has large contact areas with both EGFBP molecules. However, the amount of contact between the two EGF molecules (A1/A2) and between the two EGFBP molecules (B1/B2) would be much smaller, explaining the absence of homodimers of both EGF and EGFBP. The A1/B1 and A2/B2 intermediate complexes are identical and this will form a two-fold axis of symmetry in the final HMWEGF complex.

The DAPMatch program was used to investigate the binding of the guest molecule EGF(A2) with host molecule EGFBP(B1). This binding was complicated by the presence of the C-terminus of EGF(A1). The C-terminal residues of EGF(A1) were ill-defined in the N.M.R. structure. However, since it was known that these residues bound to the active site of EGFBP, it was possible to model the three C-terminal residues into the specificity pocket of EGFBP, using the residues occupying the P1,P2 and P3 sites (Section 1.9) in the structure between PPK and BPTI. An additional two residues from the C-terminal region were also modelled leaving the EGFBP binding cleft (Figure 4.2). These residues were chosen to be in an extended conformation. This augmented EGFBP model could then be used as a host molecule in the DAPMatch algorithm.

4.2.3 Constraints

The known biochemical data of the EGF/EGFBP system had to be translated into a set of constraints. Previously two types of constraint had been used (Section 2.12). The first was the binding site constraint, which could be applied to both the host and guest molecules and stipulated that certain residues be present in the

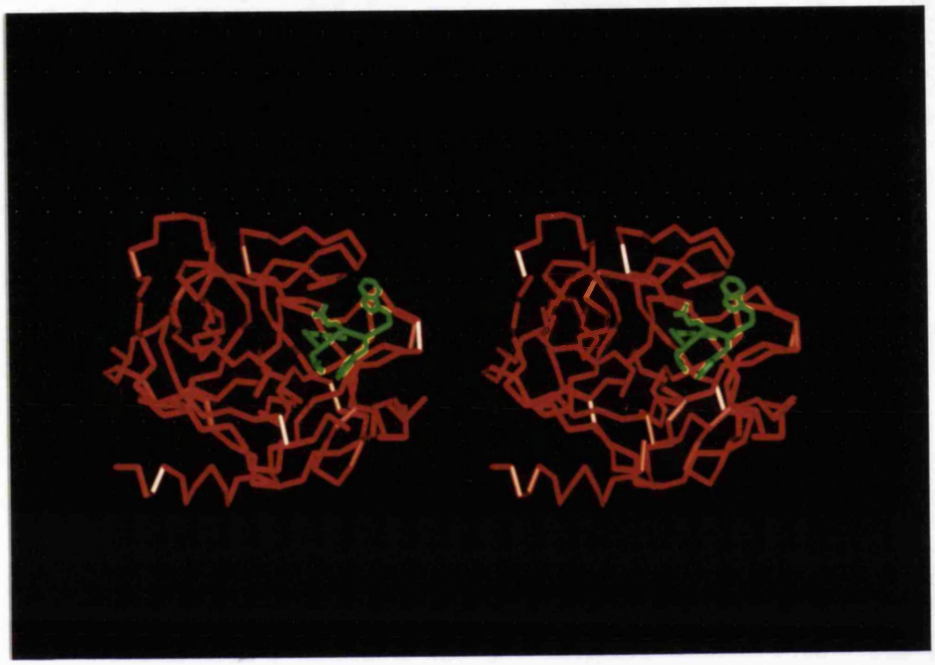


Figure 4.2

The C-terminal arm of EGF (green trace, all atoms shown) as modelled into the specificity sites of EGFBP (red trace, Cα only) binding region. The second was the loose distance constraint, which stipulated that a particular residue of the host molecule be within interacting distance of a certain residue of the guest molecule. A new constraint was now used, which was similar to the binding site constraint but more specific. This constraint required that a particular residue of one of the molecule was not only present in the binding region, but also made an interaction with an undefined residue from the other molecule. This constraint could be applied to residues that were known to form part of the binding region, but whose precise interaction was not known, and so a distance constraint could not be used.

This interaction constraint was applied so that,

i. At least one of the three C-terminal residues of EGF(A1) interacted with the guest molecule, EGF(A2).

ii. At least one of the EGFBP residues 39,40 or 41 interacted with the guest molecule. These residues form a surface loop which is strongly implicated to play a part in the complex formation (Blaber *et al.*, unpublished results).

Binding region and loose distance constraints were also applied.

iii. The binding region of the EGFBP was known. The area around the catalytic triad (Ser 57, His 107, Asp 195) was involved, as was the C-terminal section of EGF(A1) (Trp 49, Trp 50, Glu 51, Leu 52, Arg 53). Hence surface slices taken from the EGFBP molecule were only considered if they contained at least 90% of these residues.

These three constraints proved inadequate, and several hundred binding orientations still remained. One final piece of biochemical data still to be used was the implied symmetry of the HWMEGF complex (Section 4.2.1). One way for this symmetry to be satisfied would be for the two C-terminal regions of the EGF molecules, between the section bound to EGFBP and the bulk of the EGF molecule, to form a section of anti-parallel β-sheet. Similar structures are seen in prealbumin (Blake, *et al.*, 1978) and concanavalin A (Hardman and Ainsworth, 1972). This type of symmetry was present in several of the DAPMatch orientations, and it was decided to choose only structures of this type. To allow the formation of a β-sheet at the centre of symmetry it was necessary for the C-terminal residues of the guest EGF molecule (A2) to be in contact with the C-terminal residues of EGF(A1) which had been modelled into the specificity pocket of EGFBP(B1). Two constraints were used to allow for two different ways of associating the individual β–strands to form a sheet. These were that either residue 51(A1) was within 10Å of residue 47(A2) or that residue 53(A2) was within 15Å of residue 47(A2). The two β–sheets allowed by these constraints place a centre of symmetry at either residue 49 or residue 50 of the EGF C-terminus.

4.3 Method

4.3.1 The First Application of DAPMatch

The initial data for the DAPMatch program was the EGF structure, as supplied by Dr G. Montelione and the model of EGF binding protein, as supplied by Dr Bax. The algorithm presented in Chapter 2, and applied to antibody/antigen complexes in Chapter 3, was followed. The same parameters were used throughout. Surface slices were taken from both structures and the steric matching procedure followed. After clustering 1858 orientations were suggested by the DAPMatch program. The standard electrostatic filtering was then used. The increased selectivity filter presented in Section 3.8 could not be used in this case since there was no evidence for strong electrostatic linkages in the EGF/EGFBP complex. The constraints detailed in Section 4.2.3 were then applied.

The remaining 84 complexes were examined by eye on a graphics workstation. It was clear to both Dr Bax and myself that none of the structures proposed by the DAPMatch program were satisfactory. The complexes produced fell into three broad categories.

i. Straddling (Figure 4.3a,b). This class of structures brought the EGF in contact with the cleft walls of EGFBP, but not with the cleft itself. Two patches of surface area were buried, but the cleft was left unoccupied and a hole formed.

ii. End on approach (Figure 4.3c). These structures bury an area around one end of the major axis of the roughly ellipsoidal EGF molecule. This surface of EGF is bound to the cleft of EGFBP, but the total surface area buried is small and the binding cleft is not fully occupied.

iii. Overhanging (Figure 4.3d). These structures brought a variety of areas of the EGF molecule into contact with EGFBP. However none of the complexes fully occupied the binding cleft, instead one or other of the binding cleft walls formed the centre of the interaction region. This meant that the EGF molecule wrapped around EGFBP, contacting with areas that were not implicated in binding.

None of the structures suggested by the DAPMatch program fully occupied the EGFBP binding cleft. The structures of EGF and EGFBP were closely examined for features that might cause this and it was noted that some of the surface side-chains on the EGFBP model

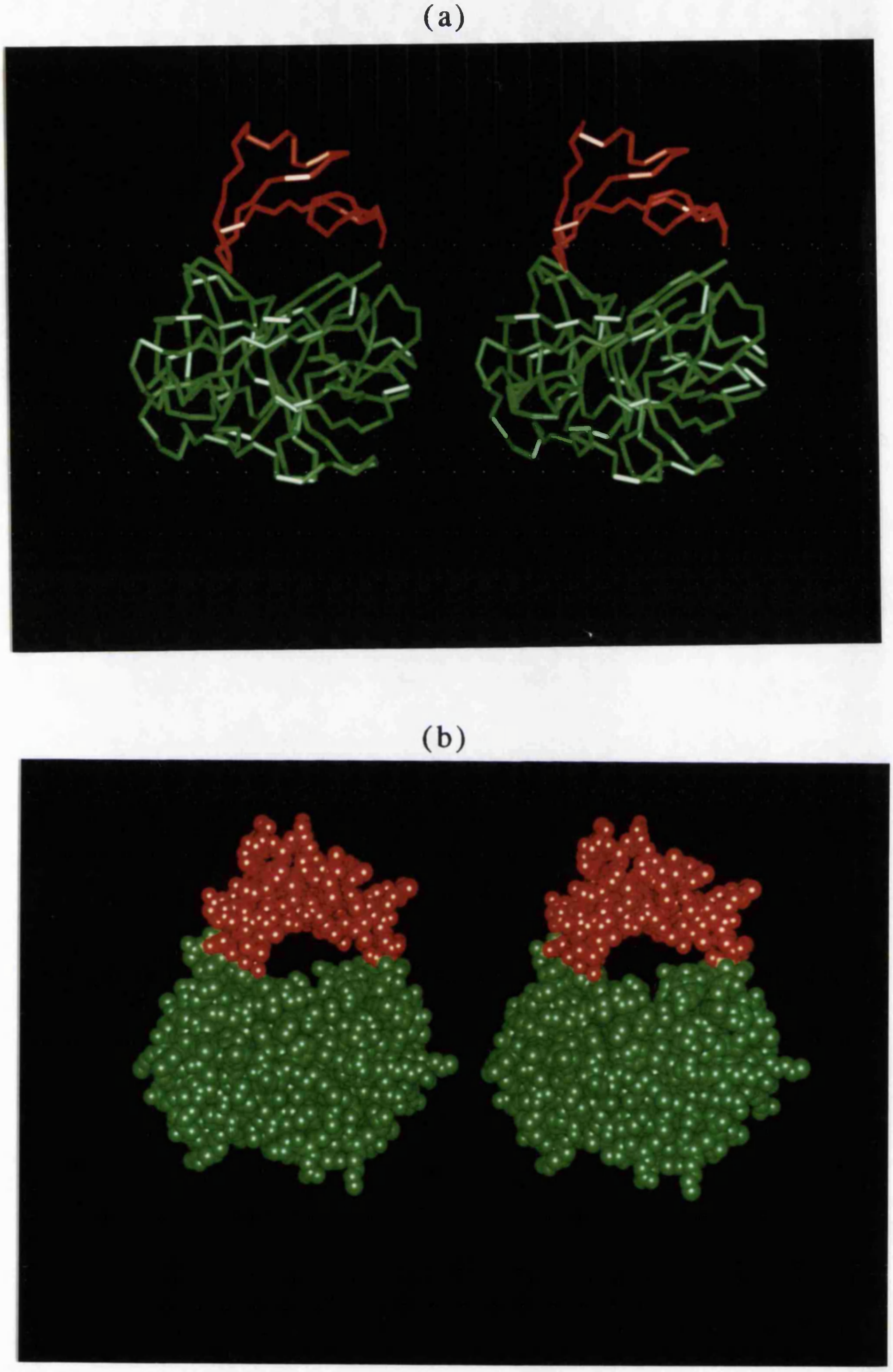


Figure 4.3a,b

Two stereo views of a type *i* docking, produced by the first application of DAPMatch to the EGF/EGFBP system. View (a) shows the Cα trace of EGF (red) straddling the binding cleft of EGFBP (green). View (b) shows a van der Waals sphere representation of the same complex, highlighting the lack of contact between EGF and EGFBP in the central region.

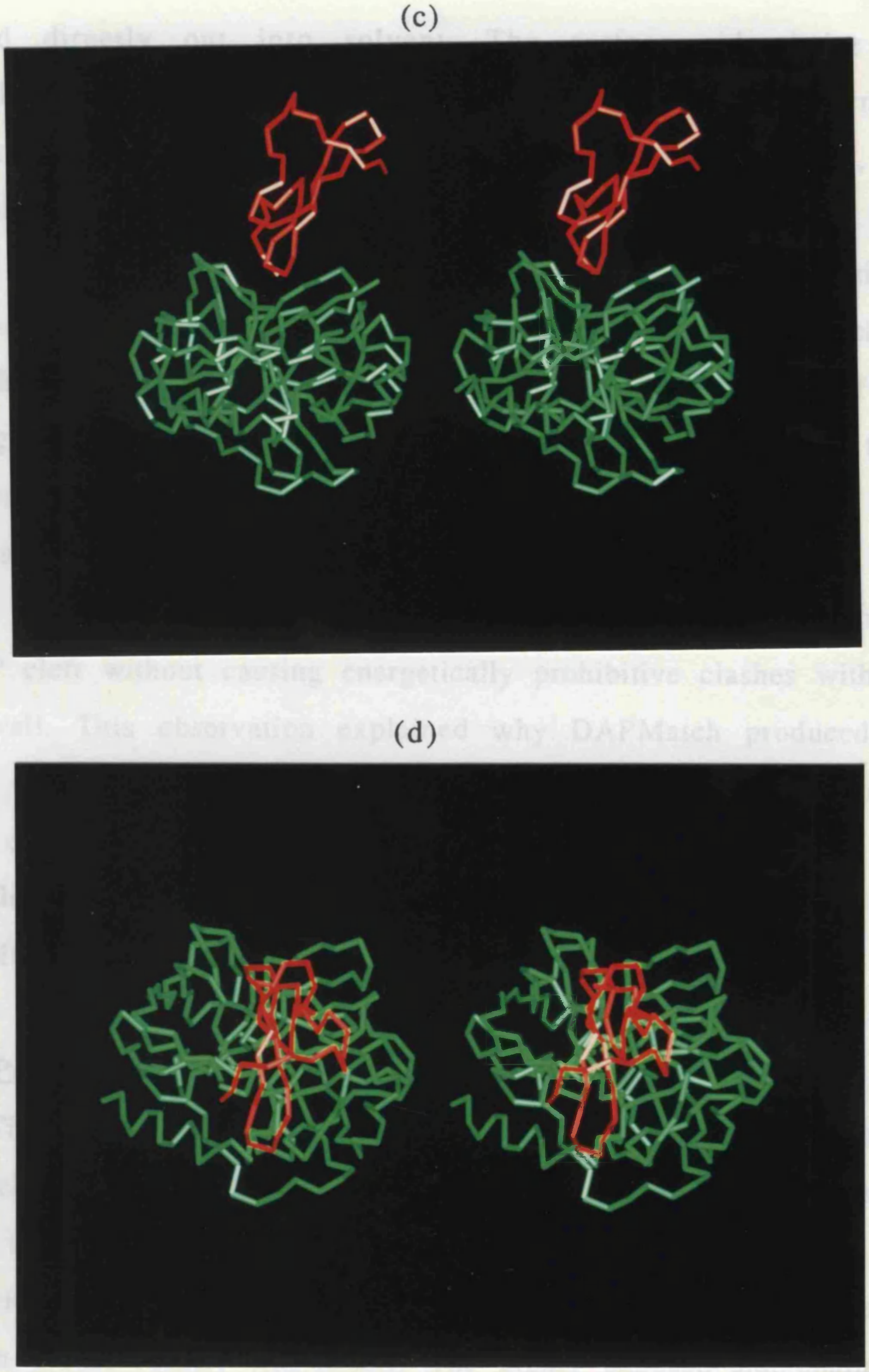


Figure 4.3c,d

Stereo views of a type *ii* docking (c) and a type *iii* docking (d), produced by the first application of DAPMatch. The type *ii* docking results in a small surface area of contact between the molecules and the type *iii* docking overhangs the binding cleft.

pointed directly out into solvent. The surface side-chains are generally mobile and hence the modelling process cannot normally predict their orientations. Due to this lack of information, the orientation of many of the solvent exposed side-chains had been chosen in an arbitrary manner. In particular, a cluster of residues were noted on one wall of the EGFBP binding cleft whose side-chains were highly solvent exposed (Figure 4.4). Surface slices of the binding cleft, as taken by the DAPMatch program, included these residues and consequently the size of the binding cleft wall was artificially enhanced. For this reason the DAPMatch program was unable to bring the EGF molecule into satisfactory contact with the EGFBP cleft without causing energetically prohibitive clashes with the cleft wall. This observation explained why DAPMatch produced the classes of complex described above: types *i* and *iii* avoided the centre of the cleft completely and type *ii* brought a narrow strip of the EGF molecule into contact with the cleft without making any contact with the cleft wall.

4.3.2 Pruning

The DAPMatch program was designed to allow for areas of mismatch during the docking process. The nature of the soft potential results in a greater tolerance for cavities between surface slices than for steric clashes (Section 2.9). This feature was used to overcome the problem of the mobile side-chains. The highly exposed side-chains of EGFBP (listed in Table 4.1) which had been identified as blocking the formation of reasonable EGF/EGFBP complexes, were removed, leaving only the main-chain atoms and either just the first side-chain carbon atom (Cβ) or the first and second side-chain carbon atoms (Cβ,Cγ). The choice of which residues to prune, and to what extent,

Molecule	Extent	Pruned Residues
EGFBP	Cβ	Arg 35, Tyr 36, Asn 38, Glu 39, Ile 41 Glu 61, Tyr 74, Phe 147, Lys 148 Ile 173
	Cγ	Lys 192
EGF	Cβ	Asn 1, Tyr 3, Asp 11, Glu 24, Leu 26, Asp 27, Ile 35, Gln 43, Arg 45, Asp 46, Leu 47 Trp 49, Trp 50, Glu 51
	Cγ	Tyr 37, Arg 41

Table 4.1

The residues from EGF and EGFBP which had their side chains removed. The extent of the pruning, whether back to Cβ atom or also leaving the Cγ atom, is indicated. All these residues were highly solvent exposed (Figure 4.4) and hence their side chain orientation would be highly variable. Note that residues 49 to 51 of EGF were the C terminal residues which were built onto the EGFBP model and hence formed part of the host molecule, not the guest.

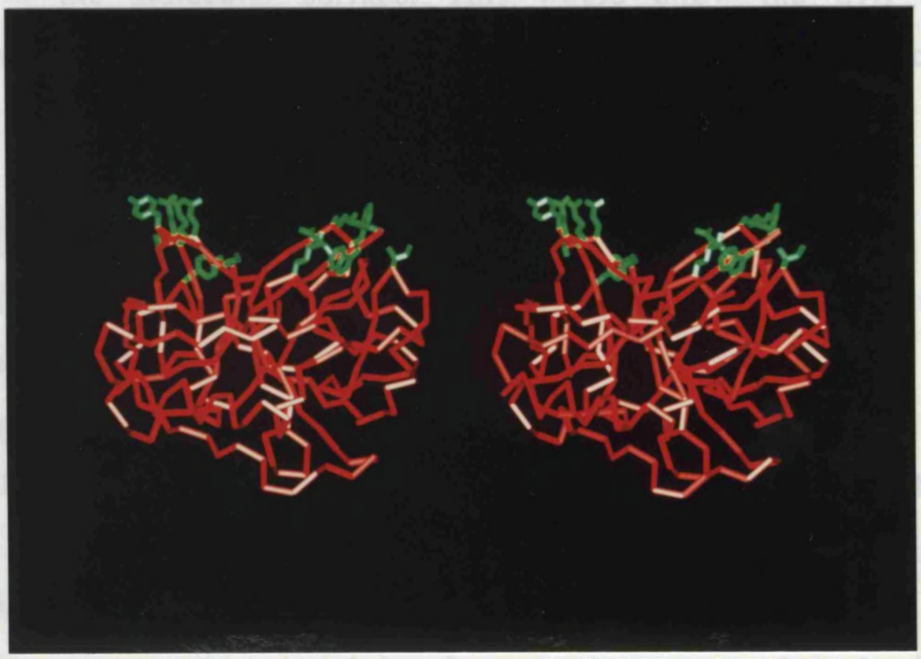


Figure 4.4

A stereo diagram of the highly solvent exposed side-chains (green trace) on the EGFBP model (red Cα trace).

was made using knowledge of the flexibility of each of the residues. Only those atoms whose positions were unpredictable were removed. As some of the residues of EGF were also highly solvent exposed these residues were also pruned (Table 4.1).

The pruning of side-chains was intended to alleviate the problem of erroneous steric clashes without drastically altering the shape of the molecular surface. This would allow the DAPMatch program to find steric matches between host and guest molecules, leaving some cavities in the region of the pruned side-chains.

4.3.3 The Second Application of DAPMatch

The algorithm was re-applied to the EGF/EGFBP system using the pruned structures. Since the DAPMatch algorithm was being used to suggest possible modes of interaction, the clustering process was strengthened so that more orientations were rejected as being similar to other, more favourable, orientations. After clustering 29 structure remained, each of which was examined by eye.

The structures produced by the second run proved to be better packed than those of the original run. In many cases the binding cleft was completely occupied by the guest EGF. Despite this better packing, most of the structures suggested were quickly rejected. Table 4.2 briefly gives the reasoning behind the rejection of each of these structures. As with the antibody/antigen results, the least favourable packings generally had low contact areas. The fourteen packings ranked lowest were all rejected because of this. Three of the remaining structures repeated orientations that had been seen in the first run, two resembled type *i* and one type *iii* (Section 4.3.1 defines this notation). These three structures were therefore rejected.

Structure	Comment
1	Insufficient space to replace side-chains.
2	A good orientation, but without the required β sheet conformation.
3	*Chosen orientation.*
4	Insufficient space to replace side-chains.
5	EGF packed across the EGFBP cleft. Type *i* structure, (Figure 4.3a)
6	Insufficient space to replace side-chains.
7	A good orientation, but without the required β sheet conformation.
8	A good orientation, but without the required β sheet conformation. Also an insufficient contact area.
9	Small contact area with no possibility of forming the required β sheet.
10	EGF packed across the EGFBP cleft. Type *i* structure, (Figure 4.3a)
11	A good orientation, but without the required β sheet conformation. Also an insufficent contact area.
12	Despite the constraints used no contact was made between EGF and the 39-41 loop of EGFBP.
13	Despite the constraints used no contact was made between EGF and the 39-41 loop of EGFBP.
14	Despite the constraints used no contact was made between EGF and the 39-41 loop of EGFBP.
15	Small contact area.

16	Small contact area.
17	Small contact area, also this orientation overhung the expect EGFBP binding region. A type *iii* structure, Figure 4.3c.
18	Small contact area.
19	Small contact area.
20	Badly packed, with prohibitively close contact between the 39-41 loop of EGFBP and the EGF molecule.
21	Small contact area.
22	Small contact area. Without the required β sheet conformation
23	Small contact area.
24	Small contact area.
25	Small contact area. Generally badly packed.
26	Small contact area.
27	Small contact area.
28	Small contact area.
29	Small contact area.

Table 4.2

The reason(s) for rejecting all but one of the 29 structures produced by the DAPMATCH program. A brief description of the reason is given, Section 4.3.2 gives more detailed reasoning. These structures are ordered according to their DAPMatch potential, in general this correlates with the amount of surface area buried by each structure.

This left twelve structures to be examined closely. Each one of these structures was well packed, filled the EGFBP binding cleft and involved the burial of a large, continuous surface of each molecule. The constraints used were as loose as possible, to avoid eliminating feasible packings. However, in several cases the constraint was found to be too loose, and packings were generated that did not fulfil the intended condition. Three packings were discarded due to lack of contact between the 39-41 loop of EGFBP and the EGF molecule. The removal of side-chains allowed the host and guest molecules to approach more closely, unfortunately this meant that it was possible for them to pack together too tightly and leave no room for the side-chains to be replaced. Four orientations were rejected because of this.

Each of the five remaining structures provided a very plausible way of constructing the HMWEGF complex. However, it had been decided to choose a structure that allowed an anti-parallel β-sheet to be formed between the two EGF components of the complex. The loose distance constraints only required that certain residues be close in space, there was no way of ensuring that the direction of the two strands was correct. Despite this one of the structure resulting from the DAPMatch program did have the strands correctly orientated for the formation of a β-sheet. This structure was chosen as the single most plausible packing orientation for EGF and EGFBP.

This structure was built into the proposed HMWEGF model (Figure 4.5). Figure 4.6 shows a van der Waals sphere representation of the HMWEGF model, clearly showing that the complex associates the four molecules into a single, well packed, globular structure. Measurements were made of the accessible surface area buried by the formation of this complex (Table 4.3a,b). These tables show that

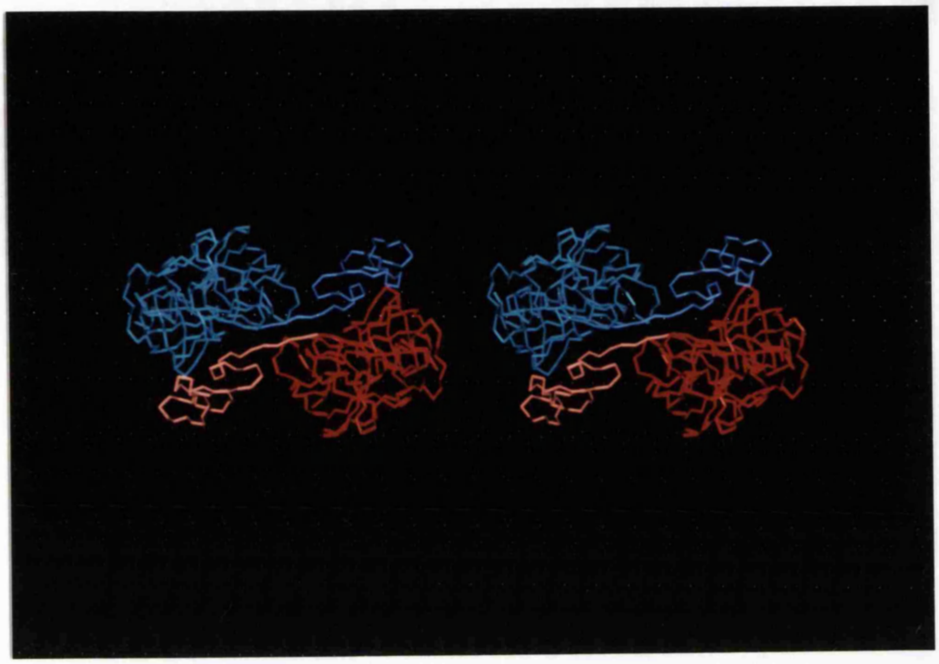


Figure 4.5

The proposed model for HMWEGF. The Cα traces of all four constituent molecules are shown (in stereo). The EGFBP molecules are in dark red and light blue, the EGF molecules are in light red and dark blue.

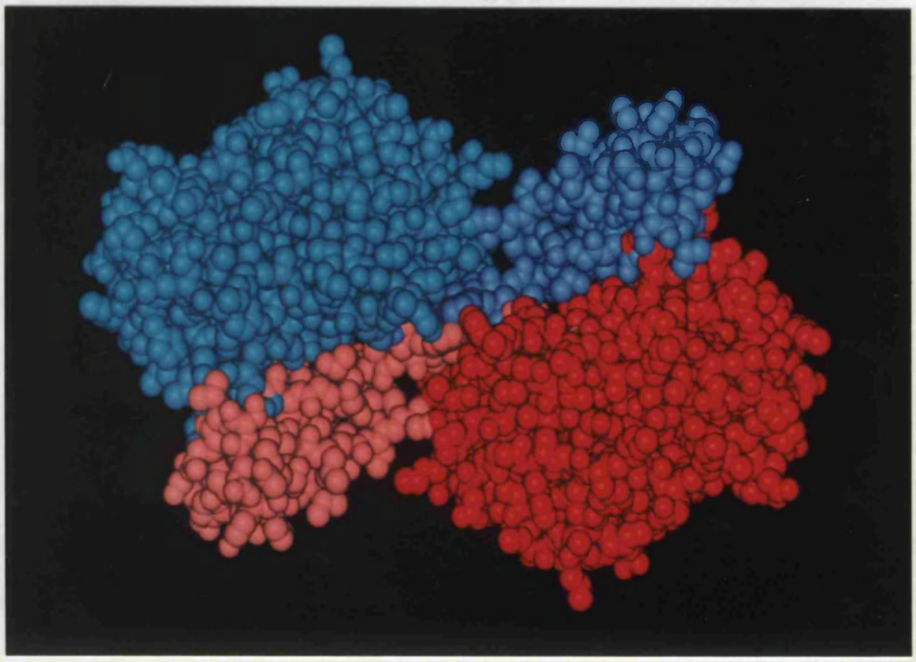


Figure 4.6

The proposed model for HMWEGF. A van der Waals sphere representation in shown. The colour scheme remains the same as Figure 4.5 (previous page). This representation shows HMWEGF as a well packed, compact complex.

Molecule	A1	A2	B1	B2
A1		533	489	939
A2	533		933	489
B1	1,077	855		85
B2	855	1,077	85	

Table 4.3a

The accessible surface area (in square Ångstroms) buried between molecules on formation of HMWEGF (see Figure 4.2 for nomenclature). The area indicated is the A.S.A. buried on the row molecule on coming into contact with the column molecule. e.g. EGF(A1) buries 855Å^2 of surface when bound to EGFBP(B2).

Molecule	A1	A2	B1	B2
A1		1,065	1,566	1,794
A2	1,065		1,794	1,566
B1	1,566	1,794		170
B2	1,794	1,566	170	

Table 4.3b

The total accessible surface area (in square Ångstroms) buried between molecules on formation of the HMWEGF complex.

each contact area in the HMWEGF molecule buries a large amount of surface area, with the exception of the B1/B2 interface which only buries 170Å^2 of surface. This fact explains the absence of an EGFBP dimer. The ASA buried between the two EGF molecules (A1/A2) is much larger, 1,065Å^2. This is mostly due to the formation of the β-sheet. The C-terminal arm of EGF is highly flexible, and the β-sheet association can only be formed once its conformation is fixed by the interaction with EGFBP. Hence a homodimer of EGF is unstable in the absence of EGFBP. The initial EGF(A1)/EGFBP(B1) (and A2/B2) dimer buries a total ASA of 1,566Å^2. The largest amount of ASA is buried by the formation of the full HMWEGF complex from two EGF/EGFBP dimers.

4.4 Conclusion

This chapter has illustrated the utility of the DAPMatch algorithm in a genuine modelling situation. The program was not used as a 'black box', resulting in a single suggested structure, but as as one of many tools used to model the HMWEGF complex. Human intervention was required at several stages, most particularly in deciding that the first attempt at steric matching had failed, understanding why it had failed and correcting the structures accordingly. Used in this way the DAPMatch program suggested many structures, any one of which could have been the basis for a HMWEGF model. The eventual choice of a single structure was based upon the formation of a well packed complex with a β-sheet at the centre of symmetry. These choices were made Dr Bax and myself, based upon experience and knowledge of other protein structures.

The concept of pruning side-chains is a useful one, which may have a more general application to systems with ill-defined or highly flexible side-chains. A similar approach was taken by Shoichet & Kuntz (1991) (Section 1.12) in their work on protease/inhibitor complexes. In this work the decision of which residues to prune, and to what extent, was taken using knowledge of both structures and chemical experience. Surface loop residues with high B-values were removed. A more rigorous method of pruning, that could be used for crystal structure, N.M.R. structures and protein models, would be necessary before the method could be used generally. It is a fundamental assumption of the DAPMatch method, and all other fixed structure docking algorithms, that induced fit plays only a small part in the docking process and hence that rigid-body docking is a useful approximation. Flexible side-chains may be removed without any loss to the steric matching procedure, since it is these side-chains that are most likely to be involved in the induced fitting of two molecules. The loss of these side-chains alters only those parts of the molecular surface that are likely to change due to induced fit anyway.

The structure presented in this chapter as a model for the HMWEGF complex fits all known biochemical data. However, this model should only be seen as a intermediate step on the way to the discovery of the HMWEGF structure. Further experiments could be designed to test the model, and in particular the presence of the central β-sheet structure should be tested.

Chapter 5

α-Helix Packing

5.1. Synopsis

This chapter describes the application of the DAPMatch algorithm to α-helix/α-helix packing. Idealised α-helices are generated and used to investigate the effect of varying side-chain sizes. This problem differs from the work in Chapters 3 and 4 since the goal is analysis not prediction, α-helices with specific sequences will not be considered. Another, more practical difference is that the interaction area for this problem is smaller than for the protein/protein interactions previously described. To allow for these differences several changes are made to the DAPMatch procedure. These changes, and the reasoning behind them, are described. The results are presented, discussed and compared to the ideal α-helix/α-helix packing models presented in Section 1.10.

5.2. Ideal α-helices

The first step in investigating the α-helix/α-helix docking problem was the creation of a series of ideal α-helices, suitable for use with the DAPMatch program. Using the graphical molecular modelling package QUANTA (Polygen Corporation, Massachusetts, USA) it was possible to fold a given protein sequence into an ideal α-helix by setting each set of main-chain φ, ψ angles to the values -57° and -47° respectively. Five ideal α-helices were produced in this way. To investigate the effect of side-chain size on α-helix/α-helix docking α-helices were generated which were 30 residues long and whose side-chains were all of a single amino acid type. The nature of the ridges and grooves on the α-helix surface are dependent upon the steric properties of its side-chains (Figure 5.1), and so the amino acids glycine, alanine, cystine and valine were chosen to represent a broad range of side-chain shape and size. The side-chain orientations were set at the most common α-helical conformation, as reported by McGregor *et al.* (1987).

A further degree of simplification was then considered. The side-chain of alanine is a single methyl group. This was normally assigned a van der Waals radius of 2.0Å (Table 2.3). By varying this radius the size of the side-chains of the poly-Ala α-helix could be controlled. The side-chain sphere radius was varied between 1.0Å and 10.0Å in 1Å steps. Although this model does not reproduce the effects of side-chain shape it does model the effects of varying the dimensions of the ridges and grooves on the α-helix surface.

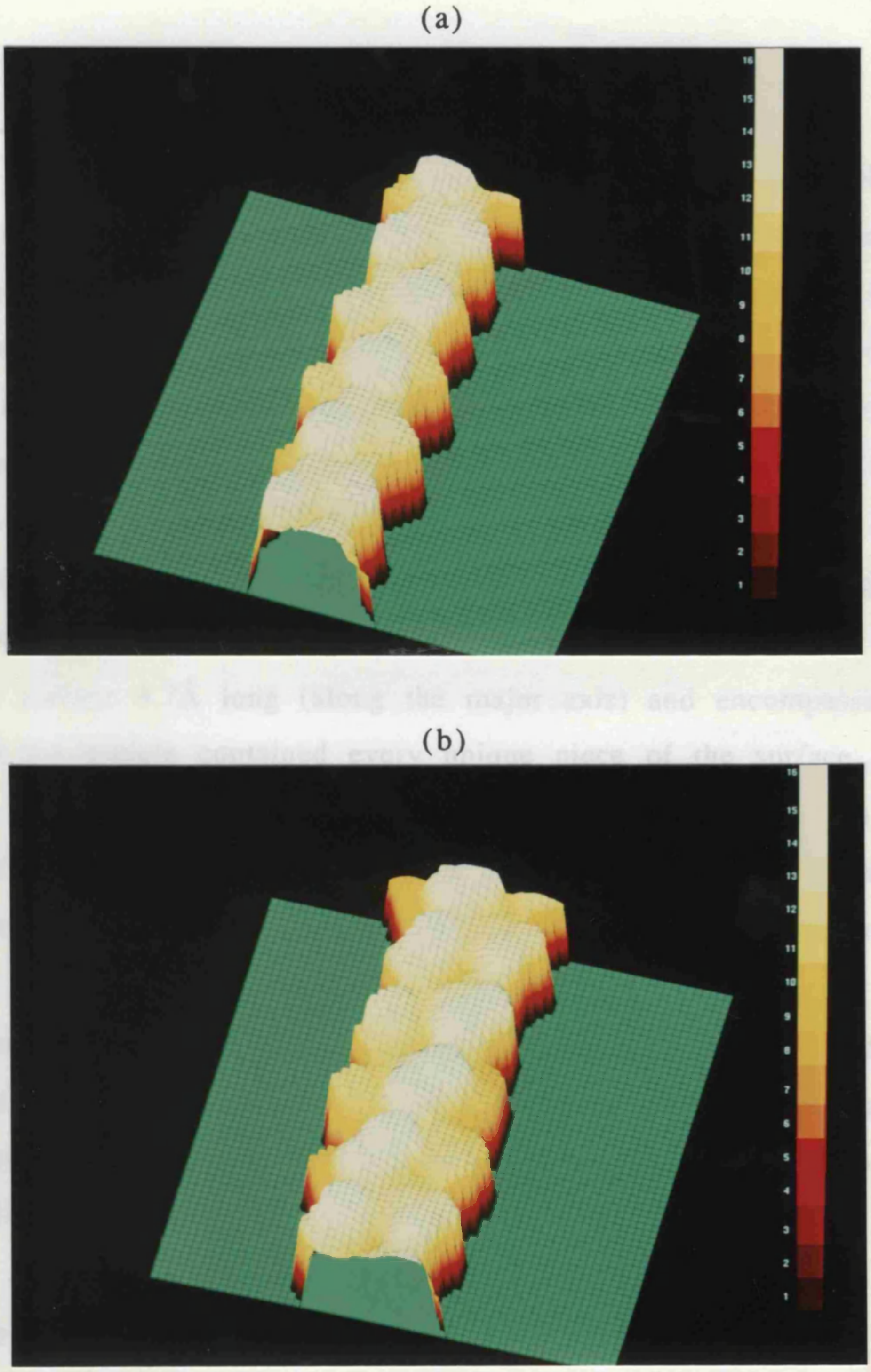


Figure 5.1

The character of the ridges and grooves on an α-helix surface change with the side-chains present. Surface (a) shows a poly-Ala α-helix. The height scale (far right) is in Ångstroms. The side-chains protrude from the surface and tend to be 14-16Å (coloured white) above the base plane (green). These protrusions form a visible, but broken, 4 groove. Surface (b) is of poly-Val, here the larger side-chains form a nearly continuous 4 groove.

5.3. Method

5.3.1 Changes Made to the Search of Configuration Space

The nature of the ideal α-helix/α-helix docking problem is different to that of the general protein/protein docking problem. Figure 5.2 shows the degrees of freedom involved. The main degrees of freedom to be considered are ω, the angle between the α-helices and z the distance between them. The α-helices used had set main-chain angles and were comprised of a single residue type, making them highly symmetric. The four other degrees of freedom can be confined to specific ranges by considering this symmetry. In particular the pitch of the α-helices was such that a patch of the α-helix surface 5.7Å long (along the major axis) and encompassing 96° about the τ angle contained every unique piece of the surface (Figure 5.3). Thus, the search ranges for these variables were chosen to be 0Å $\leq x \leq$ 6Å (in 0.5Å steps) and -48° $\leq \tau \leq$ 48° (in 6° steps). The surface of one was completely searched using the τ and x parameters. No y translations were necessary. However, as in the case of the normal DAPMatch algorithm, planar translations could be made very quickly and so shifts in the y direction were made to increase the search coverage. The range chosen was -1Å $\leq y \leq$ 1Å, in 1Å steps. Finally, -45° $\leq \theta \leq$ 45° was chosen, and this range was ample to ensure that all reasonable α-helix/α-helix packing configurations were sampled.

Apart from these theoretical differences in the α-helix docking problem there are also several physical differences. A typical α-helix/α-helix interaction buries between 200Å^2 and 400Å^2 of accessible surface area (60-140 Å^2 of contact surface area) from each α-helix (Richmond & Richards,1978) considerably less than the 750Å^2 involved at the antibody/lysozyme interfaces

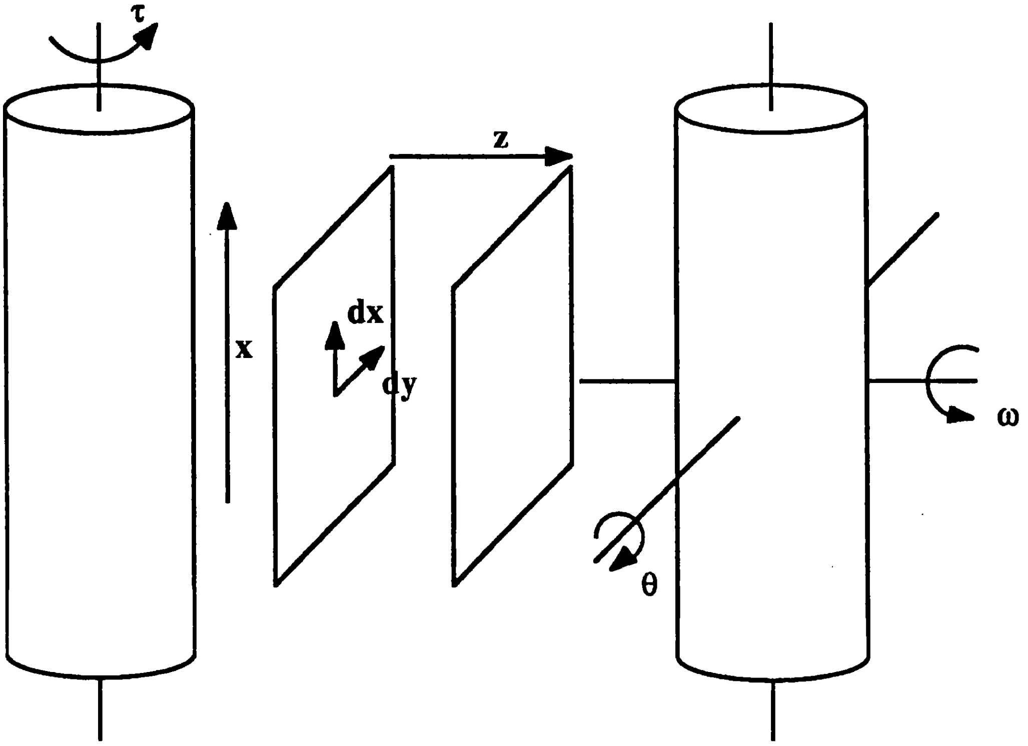


Figure 5.2

The degrees of freedom involved in the helix docking problem. The parameters dx,dy and z are varies within the DAPMATCH program, the angular parameters θ, τ, ω and the translation x are accounted for in the surface slicing procedure.

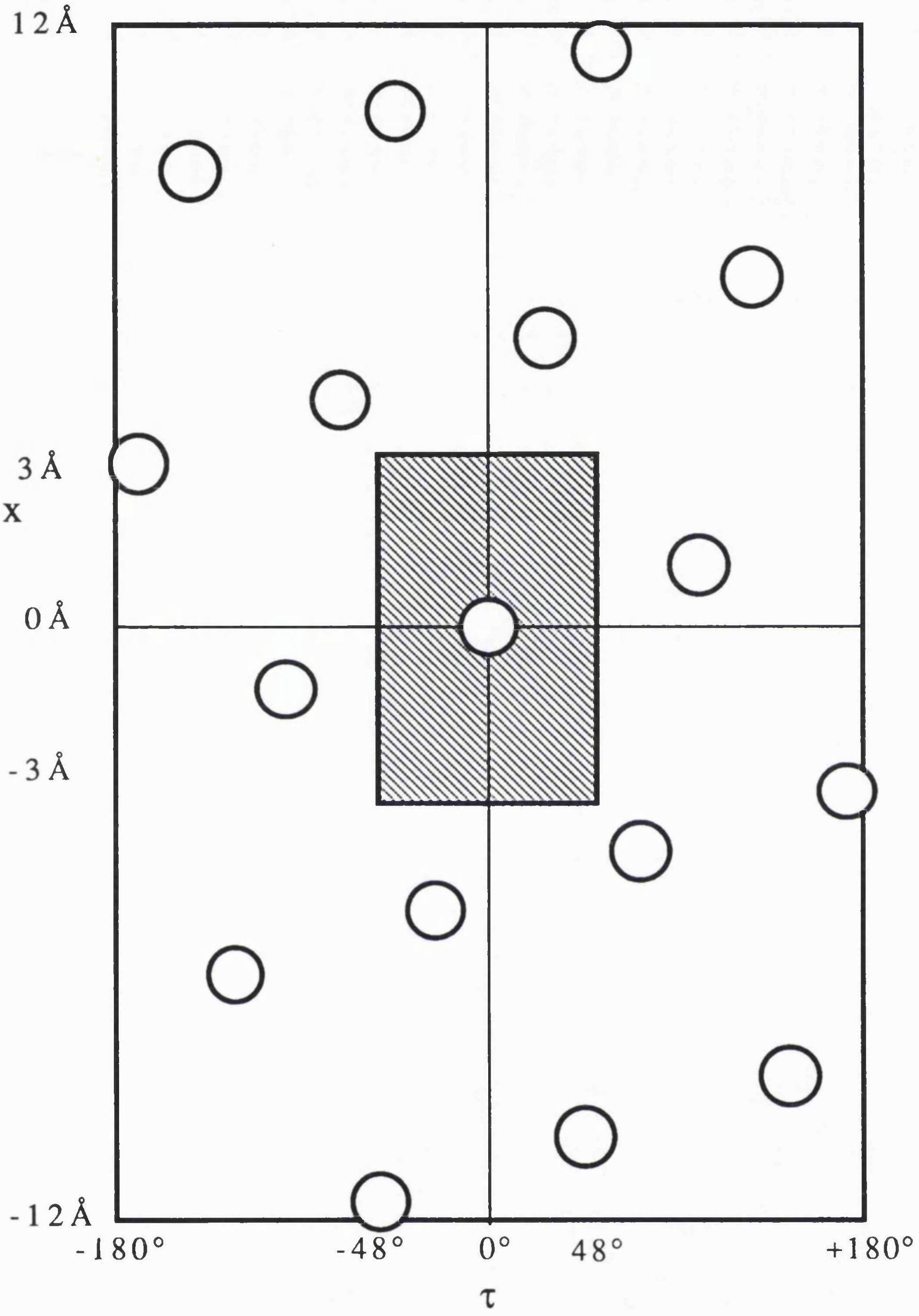


Figure 5.3

A flat α-helix representation of a poly-X α-helix. The shaded box shows an entire, unique section of surface, all other points on the helix are identical to some point within this box. The DAPMatch algorithm takes slices centred upon points within the shaded box, and therefore searches all possible x, τ orientations. N.B. The box is shown centred upon the origin, in fact the algorithm searches 0Å< x < 6Å.

(Section 1.7.4). To account for this difference the DAPMatch algorithm was altered. The minimum overlap criterion (Section 2.8) was removed to enable packing orientations with very low α-helix/α-helix interaction areas to be considered.

5.3.2 Changes made to the Matching Potential

The examination of theoretical α-helix/α-helix interactions does not require a soft docking algorithm. A much simpler potential was used which conformed closely to a rigid sphere packing model. This model treats the van der Waals radii atoms as hard spheres. As two α-helices dock together their atomic spheres come into contact, closer interaction is completely forbidden. This situation was modelled with a potential of the form

$$V_{\text{hard}}(x) = \begin{cases} \infty & x < -0.5\ \text{Å} \\ -1 & -0.5\text{Å} \le x \le 1\ \text{Å} \\ 0 & x > 1\text{Å} \end{cases} ,$$

as shown in Figure 5.4. Here, as in Section 2.9, x, is a measure of the displacement between the surface elements. Overlaps between two atomic spheres of more than 0.5Å are disallowed ($V_{\text{hard}}(x < -0.5\text{Å}) = \infty$). This condition is implemented by changing the search range for the perpendicular height. Previously (Section 2.8) the height search started with the surfaces touching and only terminated when two elements overlapped by 5Å. This ensured that the search extended to all configurations allowed by the soft potential. In the case of α-helix/α-helix docking the search was terminated when two elements overlapped by more than 0.5Å. No change to the outer height limit was necessary. Two surface elements were deemed to be interacting

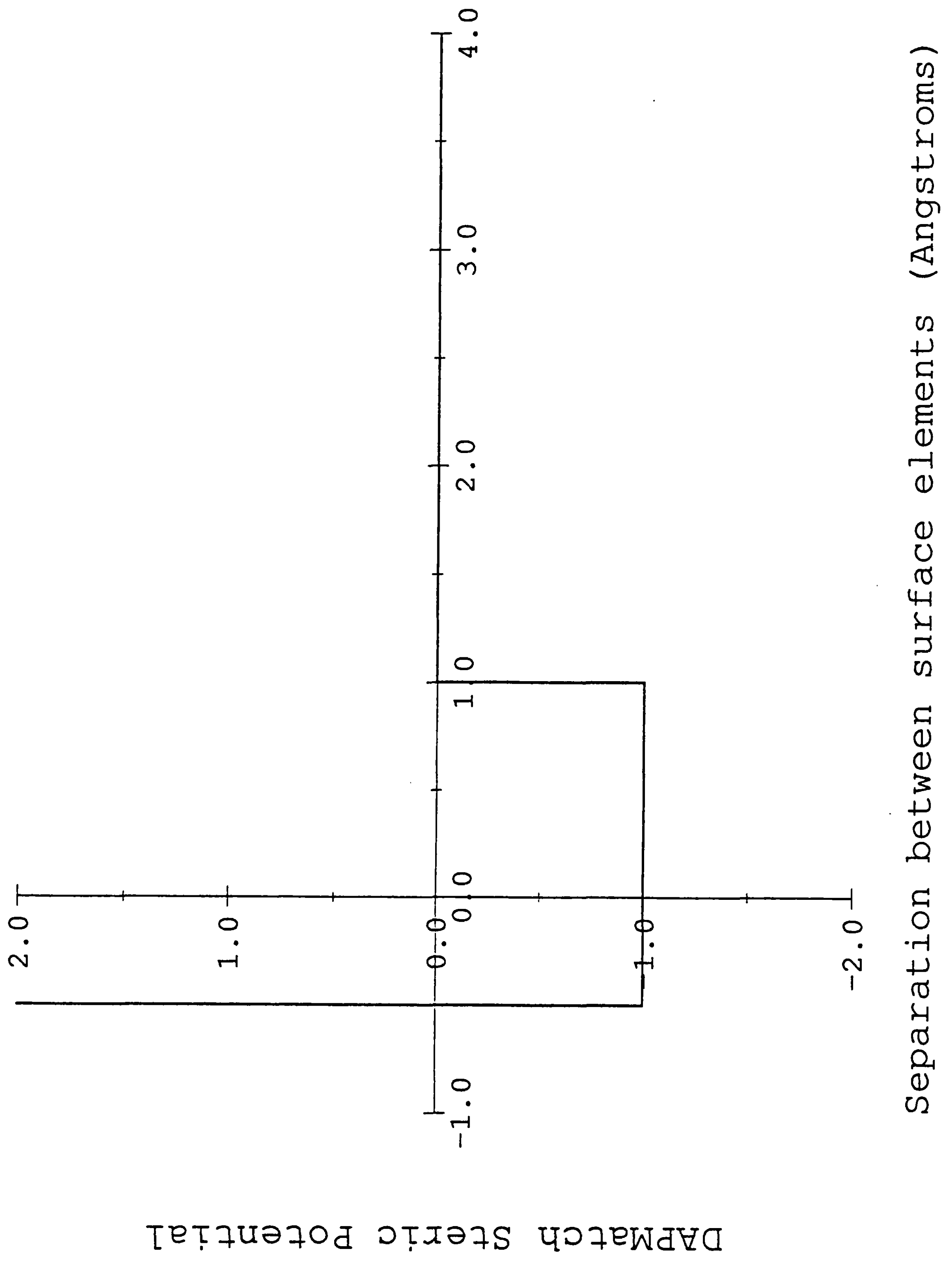


Figure 5.4

The steric matching potential used for the docking of α-helices. The separation between surface elements is plotted on the x-axis, this variable is zero when the elements touch, grows increasingly negative as the elements overlap and increasingly positive as the elements are separated.

favourably when they clashed by less than 0.5Å and were closer than 1Å from their ideal (x=0) spacing. This gave a stable potential well with a width of 1.5Å, narrower than that of the soft potential but still wider than that of a typical potential of the Leonard-Jones form. A simple square well proved sufficient to examine α-helix/α-helix docking. Both softer and harder potentials were tested. Soft potentials, such as the one used previously, produced similar results to those that will be quoted for the square well, precise preferences for docking angles were less well defined. Although a global search was made, sampling all possible orientations, the step sizes used for each degree of freedom (e.g. the along-axis translation x , was varied from 0Å to 5Å in 0.5Å steps) imposed a level of resolution on the search. The potential used could only be made harder by decreasing the width of the potential well, so that favourable dockings fell into a narrower range. This, however, brought the width of the minimum close to the resolution of the search. This meant that favourable configurations could be missed, being stepped over completely. Harder potentials were therefore impractical.

5.4. Interpretation

To compare the results of the DAPMatch program with those of theoretical α-helix/α-helix packing models it was necessary to evaluate the α-helix/α-helix packing angle, Ω. This angle was calculated by finding the line of closest approach between the two α-helical axes. The angle Ω was then the twist of one α-helix relative to the other in the plane perpendicular to this line of closest approach. If the line of closest approach did not lie within both α-helices, then the relevant end-to-end line was used. Although the ideal α-helices used were longer than usually observed in protein structures, the line of closest approach still lay off the α-helices in cases where they were arranged nearly parallel (or anti-parallel). The length of the line of closest approach, *d*, was also calculated. This distance represented the extent of intercalation between the α-helices, and hence gave some information about the stability of that α-helix packing orientation. If the relative directions of the α-helices (parallel or anti-parallel) are ignored, then the torsion angle can be confined in the range -90°< *Ω* < 90°. Although, for clarity, Chothia *et al.* (1981) quote the 1-4 packing angle as -105°, as it ranges between -80° and -110° in observed packings. The DAPMatch program searches both parallel and anti-parallel orientations, and the packing angle therefore ranges between -180° and 180°. Parallel packings have angles in the range -90° to 90° and anti-parallel packings in the range -180° to -90° or 90° to 180° (Figure 5.5).

The angle, *Ω*, was not, by itself, sufficient to confirm the type of packing the algorithm was producing. For example, although value of *Ω* = -50° suggested a 4-4 packing of the α-helices, the only true test was to examine the structure by eye and to confirm that the 4 ridge of one α-helix was docked into the 4 groove of the other. This inspection was carried out wherever possible, and was also useful in confirming that the α-helix/α-helix separation, *d*, was correct.

There was no need to cluster the results obtained from the DAPMatch program. Instead the packing angle and distance of each orientation was calculated and the packing score compared with other orientations with similar packing angles. A table was formed detailing the most favourable packing orientations, grouped into five degree ranges. In this way it was possible to compare the results of

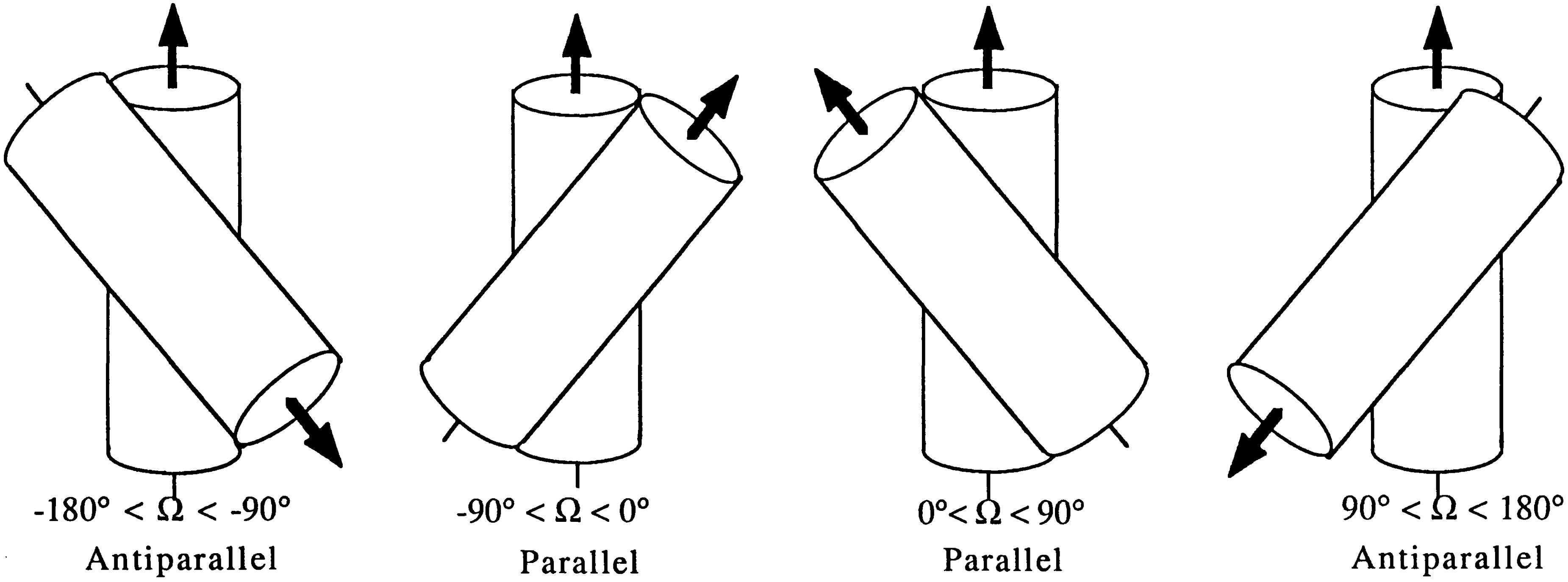


Figure 5.5

The variation of helix/helix binding orientation with the packing angle Ω. A full 360° sweep of Ω encompasses antiparallel and parallel forms of all possible binding orientations the DAPMatch program with those predicted by theoretical calculations.

5.5. Results

The results obtained for the α-helix docking problem will now be presented. The results are split into two sections, one for the effect of changing the amino acid group and one for the effect of changing the van der Waals radius of the side-chain of alanine. In each case a graph of packing potential against angle will be shown along with a selection of the structure of several packings shown to be favourable by the DAPMatch algorithm.

5.5.1 Variation in Side-chain Type

Figure 5.6 shows the variation of DAPMatch potential with α-helix packing angle for the residues glycine, alanine, cystine and valine. All four traces show a clear minimum in the region of -40°. These minima are generally broad, with widths of up to 30°. Packing angles in this region correspond closely to an parallel 4-4 packing. The corresponding antiparallel 4-4 packing angle should be in the range Ω=130° - 140°. Minima in this region are seen for alanine, cystine and valine, but the potential well is not as favourable as for the parallel form. The glycine potential, however, has a strong minimum centred at Ω= 140°. The poly-Gly α-helix has no side-chain atoms. This increases the symmetry of the α-helix and hence the potential trace has a greater symmetry. All the other α-helices have side-chains atoms. This causes an asymmetry of the surface and produces asymmetric potential traces.

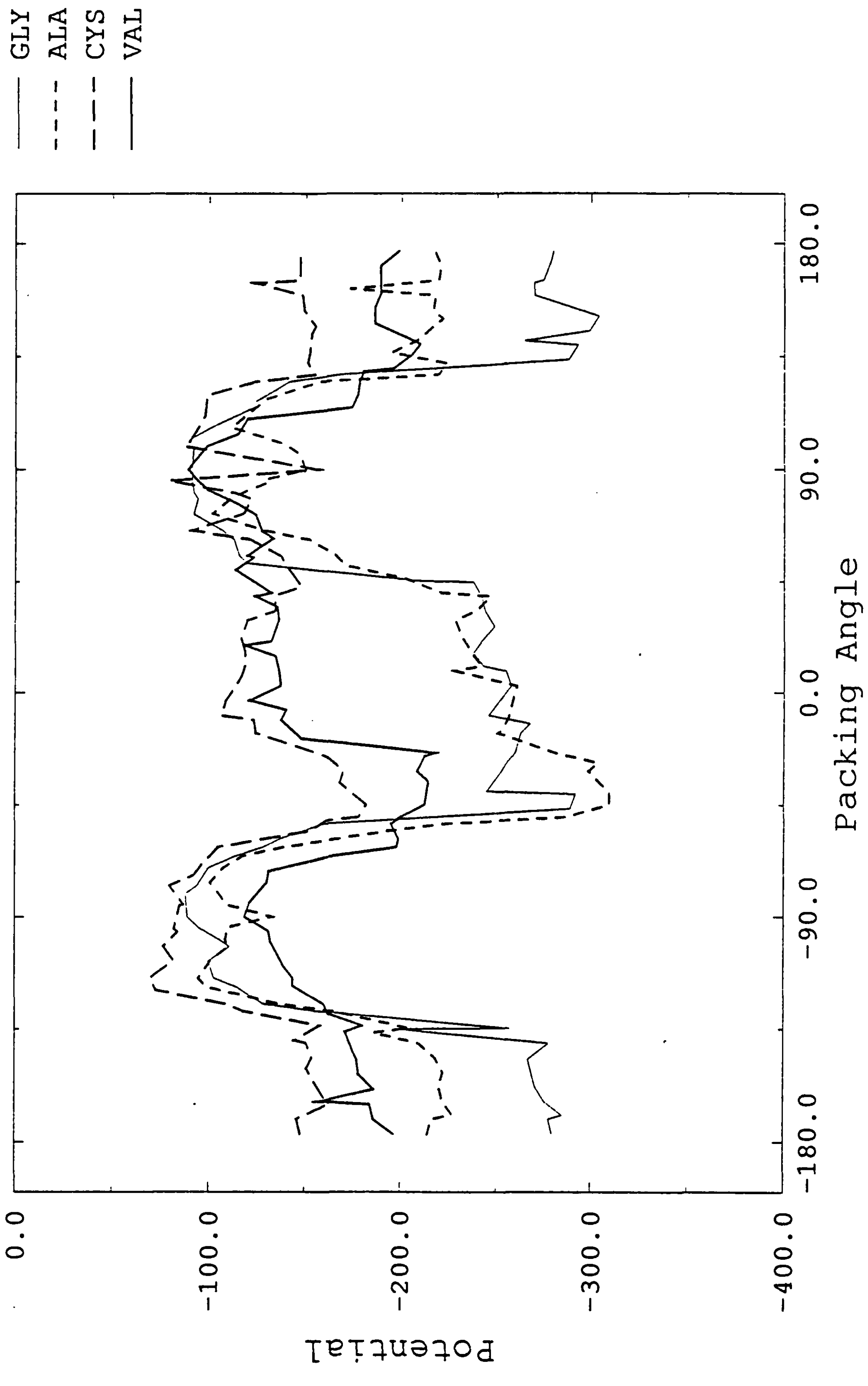


Figure 5.6

The variation of DAPMatch steric potential with packing angle for the four α-helical systems, poly-Gly; poly-Ala; poly-Cys and poly-Val.

5.5.2 Variation in Side-chain Size

The variation of the DAPMatch potential with side-chain radius (*r*) and α-helix/α-helix packing angle is shown in Figure 5.7. The trace for r=0Å is identical to the poly-Gly trace and the trace for r=2Å matches the poly-Ala trace. As the side-chain radius increases the DAPMatch potential becomes more favourable (more negative). Again, the only clear conclusion to be drawn is that the 4-4 packing is sterically favourable. As the side-chain radius becomes larger the amount of asymmetry in the potential trace reduces. At large radii the α-helix surface becomes completely dominated by the side-chain sphere, and hence the α-helix surface becomes symmetric again.

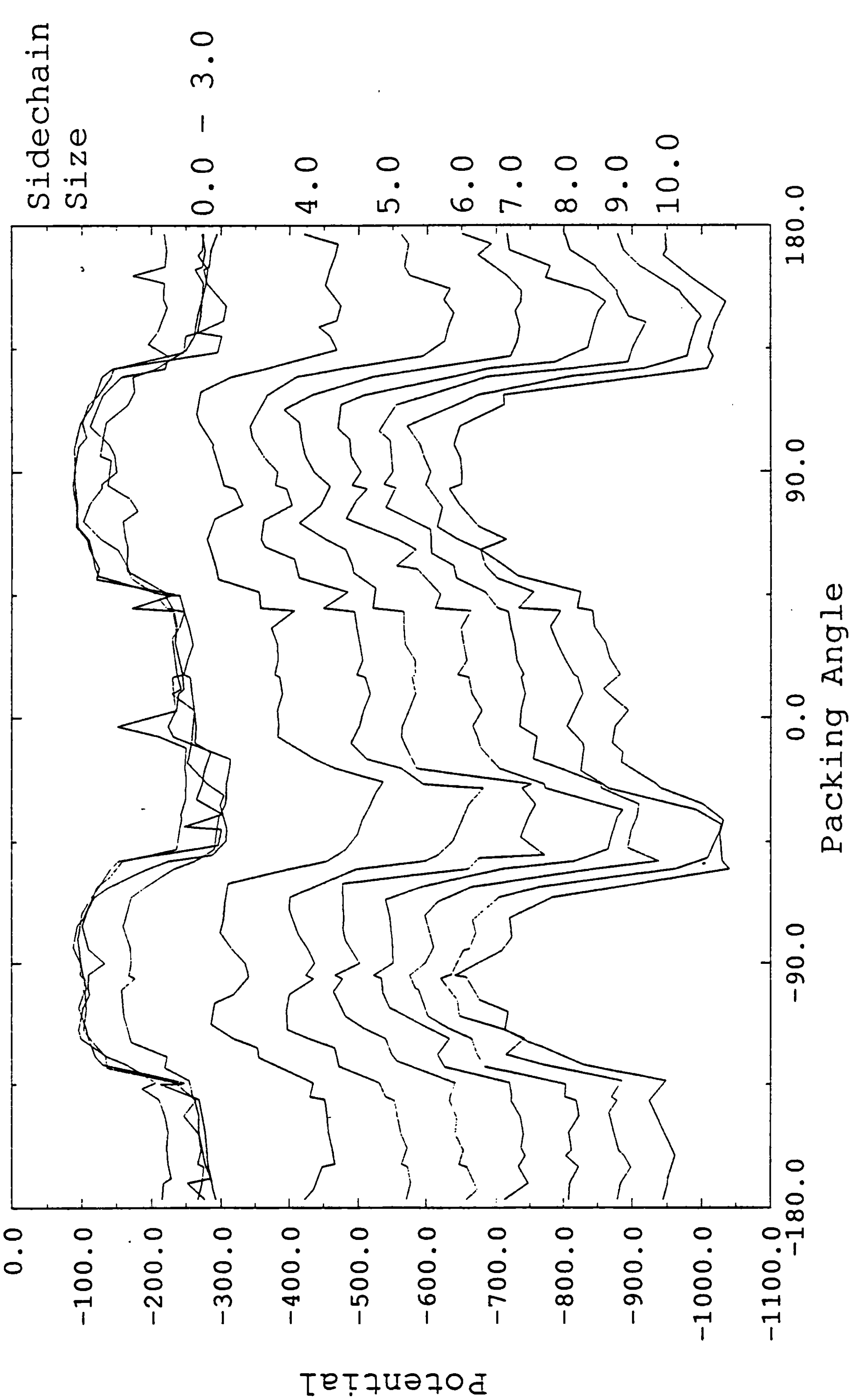


Figure 5.7

The variation of DAPMatch steric potential with packing angle for α-helix side-chain radii in the range 0Å to 10Å. The side-chain radius is indicated (far left). For radii in the range 0Å to 3Å the potential traces overlap, this section corresponds closely to the variable side-chain type potential (Figure 5.6).

5.6. Conclusion

The results presented in this chapter clearly show the steric favourability of the 4-4 α-helix packing. The use of simple single residue models proved ineffective in locating any other packing modes as favourable. Figure 1.13c shows that the 4 ridge and 4 groove are the most important prominent; hence the DAPMatch algorithm has correctly located only the most sterically favourable packing geometry. Other geometries may depend more upon the sequence of the α-helices involved, and hence would only be located if more realistic α-helix models were used. Also, the use of heterogeneous α-helix sequences would introduce electrostatic considerations to the problem.

The simplistic model used in this chapter represents a preliminary attempt at applying the DAPMatch algorithm to α-helix/α-helix packing. Although the method shows some promise further work, using real α-helices, will have to be carried out before it can be decided whether it will be genuinely useful for the docking of secondary structure elements.

Chapter 6

Conclusion

This thesis has described the development, testing and use of a novel docking method. The main objective of the method was to allow implicitly for structural change by using a soft docking potential. Chapter 2 gave a full description of the method, including a discussion of the aims of the method and details of its implementation on a parallel architecture machine, the DAP. The use of the DAP allows a global search to be carried out in a reasonable time, roughly 3 days, but severely limits the portability of the program. The search algorithm is written in DAPFortran, a highly specific, parallel form of FORTRAN that can only be compiled for DAP machines. As very few DAP machines exist, the number of people able to use DAPMatch as an effective tool is small. However, the speed of modern computers is constantly increasing, often doubling in a single year. There are already widely used, more conventional architecture machines, such as the Convex (Convex computer Ltd.(UK), Leatherhead, Surrey), which achieve speeds comparable to the DAP using standard programming languages. A version of the DAPMatch program could be written in a standard programming language, and it would soon become viable to execute this code on a wide range of computers.

Chapter 3 demonstrated the success of the algorithm on antibody/lysozyme systems. Comparison with other work (Section 1.12) is difficult since no other methods have yet been tested on all three of the antibody/lysozyme crystal structures, using the uncomplexed form of lysozyme. However, the level of accuracy achieved by the method is clearly similar to that of other methods (Table 1.6). The DAPmatch algorithm requires biochemical data to reduce the number of orientations produced to a reasonable number. This loss in specificity may be due to the softness of the potential used. A harder potential could be used, and the specificity of the algorithm might be improved, for cases where induced fit is a small effect, or test cases using complexed forms of the host and guest molecule. The soft potential is necessary, however, to match surfaces that are not perfectly complementary; as would be the case if induced fit played any part in the docking process, or if model structures were being used. There is a balance between,

i. searching for perfect steric complementarity and eliminating the majority of possible binding orientations, but running the risk eliminating the correct orientation and,

ii. searching for weak steric complementarity, being almost certain to select the correct orientation but eliminating too few orientations.

In tests on antibody/lysozyme complexes many of the docking algorithms reliably produce a binding orientation close to the known structure. However many other orientations are generally produced and no reliable way of discriminating between them has been found. This suggests that the only realistic way of using any docking algorithm is as a tool which must be backed up by other modelling tools and human experience. Chapter 4 presented a practical use of the DAPMatch algorithm, where the method was used in this way.

The concept of side-chain pruning was introduced to the algorithm only after particular problems with the EGF/EGFBP systems had been identified (Chapter 4). This idea, however, can be applied to general docking problems and could improve results by removing mobile areas of the protein surface. Pruning is particularly useful to the DAPMatch algorithm, since the soft potential (Figure 2.8) is tolerant of cavities forming at the protein/protein interface.

The results presented in Chapter 5 show the importance of steric complementarity in α-helix/α-helix docking. The simple α-helix model used limited the amount of information that could be derived from the results.

Possible Future Work

The whole DAPMatch procedure, slicing, searching, clustering and constraining, could be implemented in standard code on a single machine. This would improve the portability of the program, allow better integration of the different stages and eliminate the need for cumbersome, temporary files.

The clustering formula used to measure the similarity between orientations (Section 2.10) is not mathematically rigorous. A better formula, using rotation and translation matrices, could be found. Although this may slightly improve the results obtained it is unlikely to have a large effect. If the steric search is sufficiently discriminating then the exact clustering method used is not important.

Chapter 4 details the use of the DAPmatch algorithm in a 'real life' modelling situation. The proposed model for HMWEGF may have to be modified as new biochemical data becomes available. In particular, more information on the type of symmetry that relates the two EGF/EGFBP dimers would be useful, since the β-sheet constraint (Section 4.2.3) is not backed up by any experimental data.

Chapter 5 describes preliminary results obtained by applying the DAPMatch algorithm to helix/helix docking. These initial results are promising, the most common, 4-4, packing is clearly found. However, none of the other packing geometries are selected, this may be due to the simple helix model used. Further work, using helices with heterogeneous sequences, could be carried out to discover whether other packing geometries are sequence specific.

Many of the docking algorithms reviewed in Section 1.12 produce similar results, with near-native solutions being found within a relatively small list of possibilities. It would be useful to develop additional filters to reject incorrect orientations. These filters could encode some of the 'human experience' information which was used in Chapter 4, such as burying a continuous section of surface.

In contrast to other methods, the algorithm presented in this thesis concentrates on the soft docking of structures, to enable models to be used. Despite the limitations to the DAPMatch algorithm which are discussed above, this thesis has shown that the goal of docking a modelled antigen structure to a modelled antibody is attainable.

Bibliography

Akrigg, D., Bleasby, A.J., Dix, N.I.M., Findlay, J.B.C., North, A.C.T., Parry, S.D., Wootton, J.C., Blundell, T.L., Gardner, S.P., Hayes, F., Islam, S.A., Sternberg, M.J.E., Thornton, J.M. and Tickle, I.J. (1988). A protein sequence / structure database. *Nature,* **335**, 745-746.

Amit, A.G., Mariuzza, R.A., Phillips, S.E. and Poljak, R.J. (1985). Three-dimensional structure of an antigen-antibody complex at 6Å resolution. *Nature,* **313**, 156-158.

Amit, A.G., Mariuzza, R.A., Phillips, S. and Poljak, R.J. (1986). Three-dimensional structure of an Antigen-Antibody Complex at 2.8Å Resolution. *Science,* **233**, 747-753.

Åqvist, J., Van Gunsteren, W.F., Leifonmark, M. and Tapia, O. (1985). A Molecular Dynamics Study of the C-terminal Fragment of the L7/L12 Ribosomal Protein: Secondary Structure Motion in a 150 picosecond Trajectory. *J. Mol. Biol.,* **183**, 461-477.

Artymiuk, P.J., Blake, C.C., Grace, D.E., Oatley, S.J., Phillips, D.C. and Sternberg, M.J.E. (1979). Crystallographic studies of the dynamic properties of lysozyme. *Nature,* **280**, 563-568.

Bajaj, M. and Blundell, T. (1987). Evolution and the Tertiary Structure of Proteins. *Annu. Rev. Biophys. Bioeng,* **13**, 453-492.

Baker, E.N. and Hubbard, R.E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol,.* **44**, 97-179.

Baldwin, J. and Chothia, C. (1979). Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. *J. Mol .Biol.,* **129**, 175-220.

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C. and Wilson, I.A. (1976). Atomic coordinates for Triose Phosphate Isomerase from Chicken Muscle. *Biochem..Biophys. Res. Comm.,* **72**, 146-165

Barlow, D.J. and Thornton, J.M. (1983). Ion-pairs in proteins. *J. Mol. Biol.,* **168**, 867-885.

Barlow, D.J. and Thornton, J.M. (1988). Helix Geometry in Proteins. *J. Mol. Biol.,* **201**, 601-619.

Bennett, W.S. and Steitz, T.A. (1980). The Structure of A Complex between Yeast Hexokinase A and Glucose. I. Structure Determination and Refinement at 3.5Å Resolution. *J. Mol. Biol.,* **140**, 183-207.

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

Bhat, T.N., Bentley, G.A., Fischmann, T.O., Boulot, G. and Poljak, R.J. (1990). Small rearrangements in structures of Fv and Fab fragments of antibody D1.3 on antigen binding. *Nature,* **347**, 483-485.

Billeter, M., Kline, A.D., Braun, W., Huber, R. and Wurtärich, K. (1989). Comparison of the High-resolution Structures of the Alpha Amylase Inhibitor Tendamistat Determined by Nuclear Magnetic Resonance in Solution and by X-ray Diffraction in Single Crystals. *J. Mol. Biol.,* **206**, 677-687.

Blaber, M., Isackson, P.J. and Bradshaw, R.A. (1987). A Complete CDNA Sequence for the Major Epidermal Growth Factor in the Male Mouse Submandibular Gland. *Biochem.,* **26**, 6742-6749.

Blake, C., Geisow, M.J., Oatley, S.J., Rerat, B. and Rerat, C. (1978). Structure of prealbumin: secondary, tertiary and quarternary interactions determined by Fourier refinement at 1.8Å. *J. Mol. Biol,.* **121**, 339-356.

Blake, C.C.F., Koenig, D.F., Mair, G.A., North, A.C.T., Phillips, D.C. and Sarma, V.R. (1965). Structure of Hen Egg White Lysozyme: A Three dimensional Fourier Synthesis at 2Å Resolution. *Nature,* **206**, 757-761.

Blundell, T.L. and Johnson, L.N. (1976). Protein Crystallography., Academic Press, London.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature,* **326**, 347-352.

Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidkastner, G. and Bartunic, H. (1983). Refined 2Å X-ray Crystal Structure of Porcine

Pancreatic Kallikrein-A: A Specific Trypsin-like Serine Proteinase. *J. Mol. Biol.,* **164**, 237-282.

Branden, C. and Tooze, J. (1991). Introduction to Protein Structure., Garland Publishing Inc., New York and London.

Browne, W.J., North, A., Phillips, D.C., Brew, K., Vanaman, T.C. and Hill, R.L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.,* **42**, 65-86.

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

Bruenger, A.T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature,* **355**, 472-475.

Carter, C.W.J., Kraut, J., Freer, S.T., Nguyen, H.X., Alden, R.A. and Bartsch, R.G. (1974). Two Ångstrom crystal structure of oxidized Chromatium high potential iron protein. *J. Biol.Chem.,* **249**, 4212-4225.

Cherfils, J., Duquerroy, S. and Janin, J. (1991). Protein-Protein Recognition Analysed by Docking Simulation. *Proteins,* **11**, 271-280.

Chothia, C. (1973). Conformation of twisted beta-pleated sheets in proteins. *J. Mol. Biol.,* **75**, 295-302.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature,* **248**, 338-339.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.,* **105**, 1-12.

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature ,* **256**, 705-708.

Chothia, C. and Lesk, A.M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol,.* **196**, 901-917.

Chothia, C., Lesk, A.M., Levitt, M., Amit, A.G., Mariuzza, R.A., Phillips, S.E. and Poljak, R.J. (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, G.S., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M. and Poljak, R.J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature,* **342**, 877-883.

Chothia, C., Levitt, M. and Richardson, D. (1977). Structure of proteins: packing of alpha-helices and pleated sheets. *Proc. Natl. Acad. Sci. USA,* **74**, 4130-4134.

Chothia, C., Levitt, M. and Richardson, D. (1981). Helix to helix Packing in Proteins. *J. Mol. Biol.,* **145**, 215-250.

Chothia, C., Wodak, S. and Janin, J. (1976). Role of subunit interfaces in the allosteric mechanism of haemoglobin. *Proc. Natl. Acad. Sci. USA,* **73**, 3793-3797.

Clore, G.M., Appella, E., Yamada, M., Matsushima, K. and Gronenborn, A.M. (1990). Three-Dimensional Structure of Interleukin 8 in Solution. *Biochem.,* **29**, 1689-1696.

Crick, F.H.C. (1953). The Packing of Alpha Helices: Simple Coiled-Coils. *Acta. Crystallogr.,* **6**, 689-697.

Davies, D.R. and Padlan, E.A. (1990). Antibody-Antigen Complexes. *Annu. Rev. Biochem.,* **59**, 439-473.

Deo, N. (1974). Graph Theory with Applications to Engineering and Computer Science., Prentice-Hall, NJ.

Diamond, R. (1974). Real Space Refinement of the Structure of Hen Egg White Lysozyme. *J. Mol. Biol.,* **82**, 371-391.

Dill, K.A. (1990). Dominant Forces in Protein Folding. *Biochem.,* **29**, 7133-7155.

Efimov, A.V. (1979). Packing of Alpha-Helices in Globular Proteins. Layer-structure of Globin Hydrophobic Cores. *J. Mol. Biol.,* **134**, 23-40.

Evans, P.R. and Hudson, P.J. (1979). Structure and control of phosphofructokinase from Bacillus stearothermophilus. *Nature,* **279**, 500-504.

Fauchere, J.L. and Pliska, V. (1983). Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.,* **18**, 369-375.

Fersht, A.R. (1972). Conformational Equilibria in alpha and delta Chymotrypsin. *J. Mol. Biol.,* **64**, 497-509.

Fersht, A.R. (1987). The hydrogen bond in molecular recognition. *TIBS,* **12**, 301-304.

Fujinaga, M. and James, M.N. (1987). Rat submaxillary gland serine protease, tonin. Structure solution and refinement at 1.8 Å resolution. *J. Mol. Biol.,* **195**, 373-396.

Gellatly, B.J. and Finney, J.L. (1982). Calculation of Protein Volumes: An Alternative to the Voronoi Procedure. *J. Mol. Biol.,* **161**, 305-322.

Goodsell, D.S. and Olson, A.J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins,* **8**, 195-202.

Greer, J. (1990). Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins,* **7**, 317-334.

Gregoret, L.M. and Cohen, F.E. (1990). Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.,* **211**, 959-974.

Hardman, K.D. and Ainsworth, C.F. (1972). The Structure of Concanalvin A at 2.4Å Resolution. *Biochem.,* **11**, 4910-4914.

Hendrickson, W.A., Love, W.E. and Karle, J. (1973). Crystal Structure Analysis of Sea Lamprey Hæmoglobin at 2Å Resolution. *J. Mol. Biol.,* **74**, 331-361.

Hubbard, S.J., Campbell, S.F. and Thornton, J.M. (1991). Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.,* **220**, 507-530.

Isackson, P.J., Silverman, R.E., Blaber, M., Server, A.C., Shooter, E.M. and Bradshaw, R.A. (1987). Epidermal Growth Factor Binding Protein: Identification of a different protein. *Biochem.*, **26**, 2082-2085

James, M.N.G., Sielecki, A.R., Brayer, G.D., Delbaere, L.T. and Bauer, C.A. (1980). Structures of product and inhibitor complexes of Streptomyces griseus proteus A at 1.8Å resolution. *J. Mol. Biol.,* **144**, 43-88.

Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature,* **277**, 491-492.

Jiang, F. and Kim, S. (1991). "Soft Docking": Matching of Molecular Surface Cubes. *J. Mol. Biol.,* **219**, 79-102.

Jones, P.T., Dear, P.H., Foote, J., Neuberger, M.S. and Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature,* **321**, 522-525.

Jones, T.A. (1978). A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.,* **11**, 268-272.

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

Kabat, E.A. (1978). The Structural Basis of Antibody Complementarity. *Adv. Protein Chem.,* **252**, 1-75.

Kasinos, N., Lilley, G.A., Subbarao, N. and Haneef, I. (1992). A robust and efficient automated docking algorithm for molecular recognition. *Protein Engineering,* **5**, 69-75.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimisation by Simulated Annealing. *Science,* **220**, 671-680.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E. (1982). A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.,* **161**, 269-288.

Laskowski, M. and Kato, I. (1980). Protein inhibitors of proteinases. *Annu. Rev. Biochem.,* **46**, 593-626.

Leder, P. (1982). The genetics of antibody diversity. *Sci. Am.,* **246**, 72-83.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol,* **104**, 59-107.

Levitt, M. and Chothia, C. (1976). Structural patterns in globular proteins. *Nature,* **261**, 552-558.

Lewis, P.N., Momany, F.A. and Scheraga, H.A. (1973). Chain Reversals in proteins. *Biochem. Biophys. Acta.,* **303**, 211-299.

MacArthur, M.W. and Thornton, J.M. (1991). Influence of proline residues on protein conformation. *J. Mol. Biol.,* **218**, 397-412.

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

Matthews, B.W., Weaver, L.H. and Kester, W.R. (1974). The Conformation of Thermolysin. *J. Biol. Chem.,* **249**, 8030-8044.

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**, 295-310.

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

Miller, S., Janin, J., Lesk, A.M. and Chothia, C. (1987a). Interior and surface of monomeric proteins. *J. Mol. Biol.,* **196**, 641-656.

Miller, S., Lesk, A.M., Janin, J. and Chothia, C. (1987b). The accessible surface area and stability of oligomeric proteins. *Nature* **328**, 834-836.

Milstein, C. (1980). Monoclonal antibodies. *Sci. Am.,* **243**, 66-74.

Moews, P.C. and Kretsinger, R.H. (1975). Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference fourier analysis. *J. Mol. Biol.,* **91**, 201-228.

Montelione, G.T., Wuthrich, K., Burgess, A.W., Nice, E.C., Wagner, G., Gibson, K.D. and Scheraga, H.A. (1992). Solution structure of murine epidermal growth factor determined by NMR spectroscopy and refined by energy minimization with restraints. *Biochem.,* **31**, 236-249.

Moult, J. and James, M.N.G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins,* **1**, 146-163.

Needleman, S.B. and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.,* **48**, 443-453.

Novotny, J., Bruccoleri, R.E. and Saul, F.A. (1989). On the Attribution of Binding Energy in Antigen-Antibody Complexes McPC 603, D1.3, and HyHEL-5. *Biochem.,* **28**, 4735-4749.

Padlan, E.A. (1977). Structural Implications of Sequence Variability in Immunoglobulins. *Proc. Natl. Acad. Sci. USA,* **74**, 2551-2555.

Padlan, E.A. (1990). On the nature of antibody combining sites: unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins,* **7**, 112-124.

Pauling, L. , Corey, R.B. , and Branson, H.R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA.* **37**, 205-211.

Phillips, S.E. (1980). Structure and refinement of oxymyoglobin at 1.6 Å resolution. *J. Mol. Biol.,* **142**, 531-54.

Pollack, S.J., Nakayama, G.R. and Schultz, P.G. (1988). Introduction of nucleophiles and spectroscopic probes into antibody combining sites. *Science,* **242**, 1038-1040.

Ramachandran, G.N. and Sasisekharan, V. (1968). Conformation of Polypeptides and Proteins. *Adv. Protein Chem.,* **23**, 283-438.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.,* **34**, 167-339.

Richmond, T.J. and Richards, F.M. (1978). Packing of Alpha Helices: Geometrical Constraints and Contact Areas. *J. Mol. Biol.,* **119**, 537-555.

Riechmann, L., Clark, M., Waldmann, H. and Winter, G. (1988). Reshaping human antibodies for therapy. *Nature,* **332**, 323-327.

Ripka, W.C. (1986). Computer-assisted model building. *Nature,* **321**, 93-94.

Roberts, V., Iverson, B.L., Iverson, S.A., Benkovic, S.J., Lerner, R.A., Getzoff, E.D. and Tainer, J.A. (1990). Antibody remodeling: a general solution to the design of a metal-coordination site in an antibody binding pocket. *Proc. Natl. Acad. Sci. USA,* **87**, 6654-6658.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science,* **229**, 834-838.

Schulz, G.E. and Schirmer, R.H. (1979). Principles of Protein Structure., Springer Verlag, New York.

Server, A.C., Sutter, A. and Shooter, E.M. (1976). Modification of the Epidermal Growth Factor affecting the stability of its High Molecular Weight Complex. *J. Biol. Chem.,* **251**, 1188-1196.

Shoichet, B.K. and Kuntz, I.D. (1991). Protein Docking and Complementarity. *J. Mol. Biol.,* **221**, 327-346.

Shokat, K.M., Leumann, C.J., Sugasawara, R. and Schultz, P.G. (1989). A new strategy for the generation of catalytic antibodies. *Nature,* **338**, 269-271.

Sternberg, M.J.E. and Thornton, J.M. (1977). On the conformation of proteins: an analysis of beta-pleated sheets. *J. Mol. Biol.,* **110**, 285-296.

Subbarao, N. and Haneef, I. (1991). Defining topological equivalences in macromolecules. *Prot. Eng.,* **4**, 877-884.

Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. (1987a). Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Prot. Eng.,* **1**, 377-84.

Sutcliffe, M.J. and Hayes, F. (1987). Knowledge based modelling of homologous proteins, partII: rules for the conformation of substituted sidechains. *Prot Eng.,* **1**, 377-384.

Sutcliffe, M.J., Hayes, F. and Blundell, T.L. (1987b). Knowledge based modelling of homologous proteins, part II: rules for the conformation of substituted sidechains. *Prot Eng.,* **1**, 385-392.

Taylor, J.M., Mitchell, W.M. and Cohen, S. (1974). Characterisation of the High Molecular Weight Form of EGF. *J. Biol. Chem.*, **249**, 3198-3203

Tramontano, A., Chothia, C. and Lesk, A.M. (1989). Structural determinants of the conformations of medium-sized loops in proteins. *Proteins,* **6**, 382-394.

Ullman, J.R. (1976). An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach.,* **23**, 31-42.

Venkatachalam, C.M. (1968). Stereochemical Criteria for Polypeptides and Proteins. V. Conformation of a System of Three Linked Peptide Units. *Biopolymers,* **6**, 1425-1436.

Warwicker, J. (1989). Investigating Protein-Protein Interaction Surfaces using a reduced Stereochemical and Electrostatic Model. *J. Mol. Biol.,* **206**, 381-395.

Watenpaugh, K.D., Sieker, L.C. and Jensen, L.H. (1979). The structure of Rubredoxin at 1.2Å Resolution. *J. Mol. Biol.,* **131**, 509-522.

Waterman, M.S. (1983). Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc. Natl. Acad. Sci. USA,* **80**, 3123-3124.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984). A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Amer. Chem. Soc.,* **107**, 765-78.

Wetlaufer, D.B. (1973). Nucleation, Rapid Folding and Globular Intrachain Regions in Proteins. *Proc. Natl. Acad. Sci USA* **70**, 697-701.

William, W.V., Weiner, D.B., Kieber-Emmons, T. and Greene, M.I. (1990). Antibody geometry and form: three-dimensional relationships between anti-idiotypic antibodies and external antigens. *TIBTECH,* **8**, 256-263.

Wilmot, C.M. and Thornton, J.M. (1988). Analysis and Prediction of the Different Types of Beta-turn in Proteins. *J. Mol. Biol.,* **203**, 221-232.

Wodak, S.J. and Janin, J. (1978). Computer Analysis of Protein-Protein Interaction. *J. Mol. Biol.,* **124**, 323-342.

Wolfenden, R., Andersson, L., Cullis, P.M. and Southgate, C.C.B. (1981). Affinities of Amino Acid Side Chains for Solvent Water. *Biochemistry,* **20**, 849-855.

Wut~~h~~rich, K. (1990). Protein Structure Determination in Solution by NMR Spectroscopy. *J. Biol. Chem.,* **265**, 22059-22062.

Yue, S. (1990). Distance-constrained molecular docking by simulated annealing. *Prot. Eng.,* **4**, 177-184.

Zielenkiewicz, P. and Rabczenko, A. (1984). Protein-Protein recognition: Methods for Finding Complementary Surfaces of Interacting Proteins. *J. Theor. Biol.,* **111**, 17-30.