Analysis and Prediction

of

Protein-Protein Recognition

Matthew James Betts

February 1999

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

and

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

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

Abstract

The aims of the work presented in this thesis were two-fold. Firstly, an existing protein-protein docking algorithm (Walls and Sternberg (1992). *J. Mol. Biol.*, 228:277–297) was re-implemented on a type of computer more available than that used originally, and its behaviour was analysed in detail. This analysis led to changes in the scoring function, a treatment of electrostatic complementarity, and side-chain truncation. The algorithm had problems with its representation of surface, but more generally it pointed to difficulties in dealing with conformational change on association. Thus such changes were the second problem studied. They were measured in thirty-nine pairs of structures of complexed and unbound proteins, averaged over interface and non-interface regions and for individual residues. The significance of the changes was evaluated by comparison with the differences seen in twelve pairs of independently solved structures of identical proteins. Just over half had some substantial overall movement. Movements involved main-chains as well as side-chains, and large changes in the interface were closely involved with complex formation, while those of exposed non-interface residues were caused by flexibility and disorder. Interface movements in enzymes were similar in extent to those of inhibitors. All eight of the complexes that had structures of both components in an unbound form available showed some significant interface movement. An algorithm that was tested on five of these complexes (Gabb et al. (1997). *J. Mol. Biol.*, 272:106–120) was seen to be successful even when some of the largest changes occurred. The situation may be different in systems other than the enzyme-inhibitors which dominate this study. Thus the general model of protein-protein recognition was found to be induced fit. However, because there is only limited conformational change in many systems, recognition can be treated as lock and key to a first approximation.

Acknowledgements

Thanks first of all to Mike Sternberg for his supervision and patience over the years. Thanks to Rob Russell for help with work and continuing friendship, and for letting me stop at your flat, to Richard Jackson for advice and taking on the exciting job of proof-reading, to Peter Walls for taking the time to explain his work, and to Suhail Islam for help with geometry and pictures. Other members of the Biomolecular Modelling Laboratory also made it an interesting place to be: Henry (for telling me I'm easily amused, though I'm not sure it was meant as a complement), Bob, Lawrence, Marcel, Paul B, Paul H, Ross, Gidon, Baldo, Patrick, Hedi, and Jingchu.

My friends from other parts of the ICRF helped keep me going. Shauny, I hope that every time you see that scar on your face you remember the good times. Maria, white wine for the lady? Nicki, fancy a quick pint? (Just the one though.) Bryony, I hope you still have parties when in San Francisco. Andy, toast to the North? The friendship of everybody was invaluable; I wish we never had to go our separate ways. I try not to think too much that I won't often see these and other people from the ICRF who I didn't get the chance to know as well.

Thanks to Tim for being a top flatmate, even though we mainly talked gibberish (cheeky), and for putting me up on my frequent trips back. Also to Stephen for being a top flatmate in the past, and listening to my verbal diarrhoea in London. The canoe polo club meant my social life wasn't taken up entirely with drinking beer, and it was nice to find people who share my own amusement with rude words.

Darren and Torunn, you stopped me from being too lonely when I moved to Cambridge, and you don't know how much that means. Rodger's understanding whilst I was starting my new job and writing up was more than I had the right to expect.

Finally, Mum & Dad. I couldn't have done it without you, for more than the obvious reason.

Contents

Chapter Two - Development of a Protein-Protein Docking Algorithm 45

List of Tables

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

List of Figures

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Abbreviations

ASA	Accessible Surface Area
ASP	Atomic Solvation Parameter
CASP2	second Critical Assessment of Structure Prediction
CDR	Complementarity Determining Region
F_{ab}	Antigen Binding Fragment of antibodies, consisting of F_{C1} and F_V
F_{C1}	first Constant Fragment of antibodies
F_V	Variable Fragment of antibodies, consisting of V_H and V_L
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance spectroscopy
PDB	Protein Data Bank
PTI	Pancreatic Trypsin Inhibitor
RMSD	Root Mean Square Deviation
SCOP	Structural Classification Of Proteins database (Murzin et al., 1995)
V_H	Heavy-chain domain of the Variable fragment of antibodies
V_L	Light-chain domain of the Variable fragment of antibodies

Publication

Betts, M. J. and Sternberg, M. J. E. (1999). An Analysis of Conformational Changes on Protein-Protein Association: Implications for Predictive Docking. *Prot. Eng.,* In the Press.

Chapter One

Introduction

The binding of proteins to other proteins is an important event in many biochemical processes, including enzyme catalysis, the immune response, and signalling. The mechanisms by which protein-protein recognition occurs have therefore received considerable attention from computational biologists, both in the analysis of known complexes (e.g. Jones and Thornton, 1996) and in the prediction of their structures (for a review, see Sternberg et al., 1998). This thesis presents the development of one such prediction method, together with an analysis of a particular aspect of recognition - conformational changes induced by the formation of complexes. To understand this work it is necessary to summarise several related areas. This introduction starts with a brief description of how the structures of proteins are determined experimentally, and of the information needed to assess properly the quality of the data produced. Following this, the structural and chemical features of the interfaces of protein-protein complexes are presented, together with the changes that occur on binding. The next section explains methods of predictive protein-protein docking that have been tested on complexes of unknown structure. Finally, an outline of the contents of the rest of the thesis is given.

1.1 Protein Structure Determination

There are three main experimental methods used to determine protein structures: X-ray crystallography, Nuclear Magnetic Resonance spectroscopy (NMR. For an overview see Wuthrich, 1995), and electron microscopy (for an overview see Stowell et al., 1998). The structures of large proteins are difficult to determine by NMR - the largest NMR structure in the November 14, 1998 release of the Brookhaven Protein Data Bank (PDB) is a serine protease with a chain length of 259. Since protein-protein complexes are large almost by definition, and because the investigation of such complexes is the main topic of this thesis, NMR structures have not been analysed here. Electron microscopy is generally unable to give structures to a resolution at which more than just the overall shape of a protein can be seen, and so differences between protein structures are also difficult to see. Therefore, the focus here is on X-ray crystallography, which can cope with large proteins and complexes - the largest single chain in the PDB that was determined by this method is part of a carbamoyl phosphate synthetase molecule, and is 1058 amino acids long. It can also give structures to a resolution at which differences in amino acid conformation can be seen. Thus although structures solved by X-ray crystallography are frozen in a particular conformation, it is more suitable for looking at detailed conformational differences. This section presents a brief description of the techniques and theory of X-ray crystallography, with a discussion of those aspects needed to assess the quality of the resultant structures. The following section describes how these assessments are performed.

1.1.1 Experimental Overview

The first step is to obtain crystals of the protein that are well-ordered and so give good diffraction of X-rays. Since proteins are globular, and therefore do not pack together well, the main contacts between unit cells of the crystals are between disordered solvent molecules that fill the spaces between molecules of the protein. This means that different arrangements of the same protein are possible, and also that the crystal structure closely resembles the structure of the protein when in solution. Crystallisation is something of a black art that requires the experimentalist to try many different combinations of conditions, such as concentration, pH, temperature, and solvent, before decent crystals form.

The crystals are then exposed to X-rays, which are scattered when they interact with electrons in the crystal. This scattering occurs because the electrons are excited by the X-rays and so emit X-rays themselves in all directions as they fall back to a lower energy state. Some of these secondary X-rays interfere constructively with one another, producing diffraction patterns that are recordable on film or electronically and which relate to the structure of the protein. Diffraction produces a representation of the protein in which all the information about its structure is captured in the transverse waves of the X-ray radiation. However, only the amplitudes of these waves can be recorded; the phase is lost. Phases can be inferred from crystals of the protein that are isomorphous to the original but where a strongly diffracting heavy atom has been introduced (see the review by Ke, 1997). Other methods for solving the phase problem include refining phases calculated from a protein of known structure that is thought to be similar to the one of interest (summarised by Turkenburg and Dodson, 1996), and, more recently, multi-wavelength anomalous diffraction (Ogata, 1998). This uses x-rays of varying energies around the absorption edge of atoms attached to the protein, producing differing diffraction patterns that can be compared to determine the phase.

Each spot on the diffraction pattern, termed a reflection, corresponds to interference between X-rays that have been scattered by all atoms with a particular spacing. The more ordered a crystal, the higher the number of atoms with a particular spacing, and so the stronger the diffraction pattern. The resolution of a structure is the minimum spacing of atoms that produces reflections used in the calculation of the structure. If this value is more than about 1.5Å (the length of the carbon-carbon bond in ethane), then individual atoms can not be distinguished. This is often the case, and so the structures must be refined using other information, as described in the next section.

1.1.2 Refinement

An important stage in structure determination is that of refinement. Briefly, the aim is to find the best agreement of a model structure with the observed diffraction data and previously known chemical properties. In other words, to minimise the energy function

$$E_{total} = E_{x\text{-}ray} + E_{chem}$$

$E_{x\text{-}ray}$ describes the differences between the observed structure factors and those calculated from the model. E_{chem} restrains the model to empirically derived values for bonded and non-bonded interactions. Bonded interactions include bond lengths, bond angles, chirality and planarity. The non-bonded term includes van der Waals and electrostatic interactions.

Various refinement methods have been developed, with differences in the details of E_{chem} and $E_{x\text{-}ray}$ and in the techniques used to minimise E_{total}. Of the four methods used to produced the data presented in Data and Methods, Chapter Three, PROLSQ (Konnert and Hendrickson, 1980), TNT (Tronrud et al., 1987), and RESTRAIN (Driessen et al., 1989) all use least-squares refinement. This seeks to minimise the squares of the differences between observed and calculated values. X-PLOR (Brunger et al., 1987) uses a molecular dynamics (MD) simulation, which solves Newton’s equations of motion for every atom, with the forces acting on those atoms given by E_{chem} and $E_{x\text{-}ray}$. Least-squares refinement can only travel down the energy surface, and so is much more likely than MD to get stuck in a local minimum, which increases the need for manual intervention to vary the input parameters and to examine the results (Brunger et al., 1987). Newer methods, reviewed by Brunger et al., 1998, include simulated annealing, which is essentially MD from multiple start points (and which therefore increases the likelihood of finding the global minimum), and torsion angle dynamics, which reduces the number of degrees of freedom and so reduces the computational requirements.

1.1.3 Confidence Values and Error Estimation

There are several measures that can be used to indicate the amount of confidence with which protein structures should be treated. The three most commonly given in structures deposited in the PDB are resolution, R-factor and B-factors. More recent work has considered R-free (defined by Brunger, 1992) and standard uncertainties (SU’s). Resolution, R-factor and R-free are all measures of the overall precision of a structure, B-factors measure the precision of individual atoms, and SU’s are estimates of the precision of refined parameters.

Resolution has already been described above. It is the overall level of detail that can be seen from the diffraction data alone. The R-factor is a measure of the agreement of the observed diffraction data with that calculated from the model. Values range from around 0.6 for no agreement down to zero for perfect agreement. 0.2 is usually considered to be good enough. However, increasing the number of model parameters can reduce the R-factor without any associated improvement in the model (Brunger, 1992). Brunger, 1992, proposed R-free to tackle this problem. R-free measures the agreement of calculated diffraction data with observed data that was not included in the modelling and refinement stages of the structure determination.

The coordinates given in PDB files are the most likely position of the centroids of the atoms. These are taken from the maxima in the electron density, and B-factors indicate the rate at which the density drops off from this position. They are a measure of the expected deviations about the centroids, caused by dynamic and static disorder in the crystal. Dynamic disorder is simply the thermal motion of an atom, and because of this B-factors are often termed ‘temperature factors’. It is a measure of the mobility of the atom. Static disorder arises from the difference in position of two equivalent atoms from different molecules in the crystal. These two types of disorder are difficult to distinguish because X-ray structures are time-averaged, and so B-factors include them both. However, Artymiuk et al., 1979, demonstrated a correlation between the B-factors of lysozyme and its flexibility, with the active site showing high mobility. Figure 1-1 shows the relationship between B-factor and root mean square deviation (RMSD).

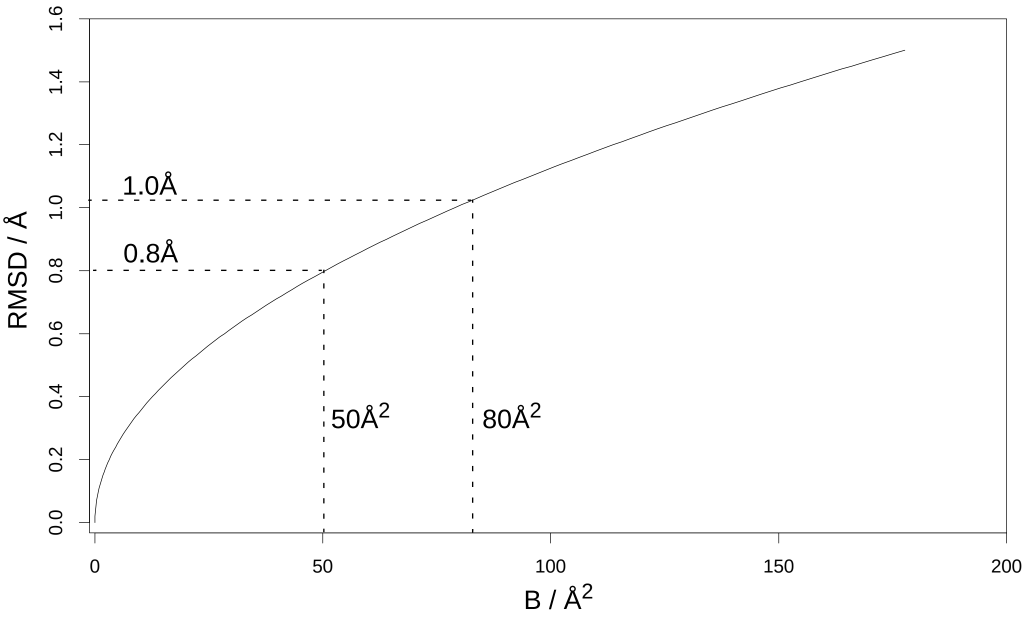
Daopin and Davies, 1994, compared two structures of transforming growth factor β (TGF−β), and used four different methods to estimate the coordinate errors. Three of these methods require knowledge of the diffraction data, which is not generally available in the public domain. Hence they are not discussed further, except to say that they cannot give a value for systematic differences in the determination of structures; these can be found only by comparing independently solved structures, as presented in this thesis (see Chapter Four). The fourth method was based on such a comparison, but of only one pair of structures. Tickle et al., 1998 calculated standard uncertainties for two crystallin structures from full-matrix least-squares refinement. This also requires generally unavailable data and can not quantify systematic differences. For these measures to become more widely used, they either need to be given in PDB files, or the data from which they are calculated should be distributed.

Figure 1-1 - The Relationship Between Temperature Factor and RMSD

Temperature factor = $B = 8\pi^2\overline{u^2}$, where u is the atomic displacement amplitude.

$\sqrt{\overline{u^2}} = \mathrm{RMSD}$, therefore $\mathrm{RMSD} = \sqrt{\frac{B}{8\pi^2}}$

Thus when the temperature factor of an atom equals 80Å^2, the RMSD = 1Å, and the position of the atom is unlikely to be determined precisely (Cruickshank, 1996). A temperature factor of 50Å^2 gives an RMSD of 0.8Å, which is approximately half the length of a carbon-carbon bond (Engh and Huber, 1991).



1.1.4 Structure Validation

Structure validation is the process of testing the correctness of a model and assigning confidence values to it, by an assessor who is independent of those who determined the model (Dodson et al., 1998). This can be broken down into two questions: i) do the experimental data justify the model?; and ii) does the model agree with empirical criteria?

Point ii) obviously requires that the empirical criteria themselves are reliable. The values for E_{chem} are derived from crystal structures of small organic molecules (Engh and Huber, 1991), which do not suffer as much from the problems seen with macromolecular crystallography - mobile solvent, weak crystal contacts, and variable periodicity (Dodson et al., 1998). This means that they can be determined to atomic resolution, and that the variation of their stereochemical properties can be measured. As proteins are also organic molecules, it is assumed that their stereochemical properties will be similar to these. However, inclusion of data from the determination of protein structures at atomic resolution, as these become more available, will obviously increase the reliability of these parameters (Wilson et al., 1998).

Wilson et al., 1998, applied four different validation tools to eight atomic resolution structures. The distinction between two types of stereochemical properties was made - those that are used in refinement (see above), and others that were termed ‘conformational’ properties. These included backbone and side-chain torsion angles, ring-pucker and residue packing. The different environment of proteins compared to small organic molecules means that it is unreasonable to share standard values for these other parameters between the two. Therefore the validation tools examined derive them from structures in the PDB. They are not restrained in refinement and so are good features to check in new structures; values of the properties used in refinement are biased towards the values to which they were restrained. However, as the authors point out, bias in the conformational parameters could creep into the database if structures are validated in this manner before deposition, but without careful attention as to whether the corrections agree with the diffraction data. Structures at atomic resolutions have little ambiguity in where atoms should be placed in the electron density. Wilson et al., 1998, tested the performance of four validation tools on eight such structures, and found that standard uncertainties for the conformational parameters were generally lower than expected. The torsion angle defined around the peptide bond had a higher standard uncertainty than expected, close to that seen in small organic molecules. This analysis indicates the need for the tables of target values for stereochemical parameters used in refinement and validation to be updated with information from atomic resolution protein structures.

1.2 Characteristics of Protein-Protein Interfaces

Several different features of protein-protein interfaces have been investigated in the past, and can be divided broadly into two overlapping categories: i) structural properties - size (measured by the burial of accessible surface area, or ‘ASA’ - see Chapter Two, figure 2-4), shape, and shape complementarity; and ii) chemical properties - solvation potential (linked to ∆ASA), hydrophobicity, electrostatic potential, hydrogen bonds and salt bridges. Residue propensities have also been examined, and are related to all the other properties. Jones and Thornton, 1996 calculated the propensity of different amino acid types to be in an interface, and saw in general that hydrophobic residues were more common than in other parts of the surface. This section presents the findings of studies of these structural and chemical properties, with particular emphasis on hetero-protein complexes. Oligomeric proteins are not usually found in a dissociated state, and so it is reasonable to assume that their interfaces have peculiarities that are not necessarily true in the area of interest.

1.2.1 Structural Properties

The size of the interfaces of protein-protein complexes is usually given as the difference between the ASA of the complex and the separated components. This gives an indication of binding strength (Jones and Thornton, 1996), because the burial of surface area is related to the hydrophobic energy of desolvation (Chothia, 1974). Both Janin and Chothia, 1990, and Jones and Thornton, 1996, with similar data sets, observed that the mean ∆ASA for enzyme-protein inhibitor complexes and for antibody-protein antigen complexes was similar at approximately 800Å^2 per component. The antibody-antigen complexes showed more variation towards greater values from this mean (up to around 875Å^2 for the complex between Fab NC41 and neuraminidase, Janin and Chothia, 1990), with a standard deviation of 135Å^2 compared to one of 75Å^2 for the enzyme-inhibitor complexes (Jones and Thornton, 1996).

Jones and Thornton, 1996, also found a higher mean and standard deviation (849Å^2 and 244Å^2 respectively) in seven hetero-complexes of other types, reflecting the greater diversity of molecular weights of the components and the nature of their interfaces. ∆ASA was higher still for permanent complexes.

Whilst considering shape, Jones and Thornton, 1996, found that antibody-protein interfaces were more planar than those of enzyme-inhibitor complexes, indicating that catalytic residues are usually located in surface clefts. The mean planarity for other hetero-complexes was approximately half way between these two, but with more variation.

The requirement for close packing at protein-protein interfaces has been known for a long time (Chothia and Janin, 1975). Janin and Chothia, 1990 saw close packing in their analysis of enzyme-inhibitor and antibody-antigen complexes. Lawrence and Colman, 1993, developed a measure with which shape complementarity could be quantified. The measure combines the distances between points on each surface in the interface with the angles between surface normals at these points, to give a value known as ‘Sc’. Sc is equal to one for a perfect fit, and tends to zero for very poor fits. Enzyme-inhibitor complexes gave higher values than antibody-antigen complexes (0.75 against 0.65). The authors suggest that this is a consequence of the necessity for antibodies to recognise modified or previously unseen antigens. Jones and Thornton, 1996, confirmed this work with measurements of the extent of gaps in interfaces.

Ysern et al., 1998, calculated Sc for the interface of another type of immune system complex - that of a T-cell receptor (TCR) bound to a self-peptide-MHC. The value of 0.45 indicates significantly worse packing than the other complexes. Once again this is related to the biological function of the molecules involved. During development, T-cells are selected based on the binding of their receptors to self-peptide-MHC. If this is too tight then the cell is not allowed to proliferate, so that auto-immune reactions are avoided. If binding is too weak then foreign-peptide-MHC may not be recognised (since the MHC provides the majority of the binding surface), and so such T-cells are also selected against.

These examples indicate that methods for predicting the structure of complexes may need to be tuned to the problem at hand.

1.2.2 Chemical Properties

Solvation potentials measure the preference of amino acids to be exposed to solvent or to be buried. Jones and Thornton, 1997a, used an empirical scale (based on the average ASA seen for each amino acid type in a set of non-homologous proteins) to measure the differences in the solvation potentials of interface regions with those of other surface patches. The results for hetero-protein complexes showed no particular trend, except that in general they had higher solvation potentials than homo-dimers. This reflects the fact that the components of a hetero-protein complex must be able to exist independently in solution. A quality related to solvation potential is that of hydrophobicity. Jones and Thornton, 1996, calculated hydrophobicity using the empirical scale of Janin et al., 1988. Exposed residues of all the different types of hetero-protein complex had roughly the same negative values. The interfaces were slightly less hydrophilic, significantly so for the enzyme-inhibitor complexes. This explains the higher-binding affinities between enzymes and inhibitors.

The general analyses of Janin and Chothia, 1990, and Jones and Thornton, 1996, both comment on the electrostatic complementarity in protein-protein interfaces, relating it mainly to observed residue-residue interactions. Honig and Nicholls, 1995 looked at the electrostatic field across protein surfaces. This models the propagation of charge through the protein and solvent environments, and the effect that the shape of the protein has on the electrostatic surface. It was seen that the electrostatic surfaces generated also showed a high degree of complementarity.

Antibody-protein and enzyme-inhibitor complexes both involve an average of ten intermolecular hydrogen bonds (Janin and Chothia, 1990). The apparent disagreement between this and the different levels of hydrophobicity seen in the two types of complex (above) can be explained by the fact that the majority of hydrogen bonds in enzyme-inhibitors are between main-chain atoms, and so do not require polar or charged residues. Jones and Thornton, 1996, observed more hydrogen bonds per 100Å^2 ΔASA of enzyme-inhibitor complexes than of antibody-protein complexes. This disagreement with Janin and Chothia, 1990, is presumably a consequence of the differences in the data sets. Along with the results for other hetero-complexes, in which Jones and Thornton, 1996, saw less hydrogen bonds per 100Å^2 ΔASA of other hetero-complexes than they did with the first two types, the variable nature of protein-protein interfaces is highlighted.

What both studies (Janin and Chothia, 1990, and Jones and Thornton, 1996) lack, however, is an analysis of hydrogen bonds that are mediated by bound water molecules. This is probably because of the difficulties in locating ordered water molecules in electron densities (Savage and Wlodawer, 1986). Ordered water has been seen in the interfaces of antibody-protein complexes (Bhat et al., 1994). It is likely to be more common than in enzyme-inhibitor interfaces because such interfaces have a better fit (see section 1.2), with little room for water. An analysis of newer structures, at resolutions that are high enough to resolve bound water molecules, will probably show that the packing and number of hydrogen bonds in all interfaces is largely proportional to the sizes of the interacting surfaces. This was seen by Xu et al., 1997, who examined over 300 protein interfaces (though most of these were between chains that do not exist independently). Thus docking algorithms could benefit from a consideration of such water molecules, though at increased computational cost. Xu et al., 1997 also saw about two salt-bridges per interface.

1.3 Conformational Changes Upon Protein-Protein Association

When the structures of a complex and of its components in isolation have been determined, the workers report the conformational change on association (e.g. Hecht et al., 1991, Hecht et al., 1992, Bhat et al., 1994, Chantalat et al., 1995). On the limited data sets available at the time, Huber, 1979, Janin and Wodak, 1983, Bennett and Huber, 1984, and Janin and Chothia, 1990, described general features of conformational changes in proteins. More recently, Stanfield and Wilson, 1994, have reviewed conformational changes in antibody-antigen association, and in a series of papers by Lesk and Chothia, 1988, Gerstein and Chothia, 1991, and Gerstein et al., 1994, the nature of domain movements in proteins has been analysed. However, these studies are dominated by the conformational change induced by small molecules binding to proteins. The topic of this thesis is a single type of recognition - the formation of heteroprotein complexes. The lack of literature about conformational changes on the formation of such complexes forces this section to summarise the general modes of flexibility seen in all cases, and to indicate how heteroprotein complexes fit into this scheme.

The studies listed above identify five main types of flexibility. These can be associated with different types of function, as outlined below.

Movement Between Rigid Domains Connected by a Flexible Linker

Domains of this type have minimal contacts with each other. Such cases, for example the F_V and F_{C1} domains of antibodies, show a wide range of motion. This enables multi-site proteins to adapt to recognise macromolecular antigens or cell-surface motifs (Janin and Wodak, 1983). However, it is unclear whether binding causes these changes (Stanfield and Wilson, 1994), especially as similar differences have been seen between different crystal forms of the same antibody (Lesk and Chothia, 1988). This type of flexibility would not necessarily affect the performance of docking algorithms, as the conformation of the interface is largely unchanged.

Movement Between Rigid Domains Linked by a Short Flexible Hinge

Domains of this type are in close contact. This type of movement allows only a few different conformations. Gerstein et al., 1994, reviewed domain closure movements that fit this description. The movements exclude water and improve the position of the catalytic residues around the substrate. They can be characterised further as ‘shear’, where the domains slide across each other (for example citrate synthase upon citrate binding), and ‘hinge’, where one domain rotates towards the other about the hinge (for example lactoferrin upon iron binding).

Movement Which Occurs when Disordered Domains Become Ordered

Huber, 1979, saw this in a comparison of trypsinogen, trypsin and their complexes with pacreatic trypsin inhibitor (PTI). The complexed proenzyme had a remarkably similar structure to the bound enzyme, despite differences between their free structures. Sufficiently strong ligands (such as PTI) were able to induce a conformational change in four disordered loops of trypsinogen. This enabled binding in the same manner as trypsin, though with lower association energy. The conformational change is the same as occurs when the proenzyme is converted to the active form by proteolysis. In both cases this should be thought of as the freezing out of one particular conformation, rather than a conformational change. The transition from disorder to order is one of the mechanisms by which catalytic activity is regulated.

Movement of Secondary Structural Elements

Gerstein and Chothia, 1991, examined an association that involved conformational changes at this level, in the loops and helices of lactate dehydrogenase that move when it binds lactate and NAD. Binding caused the 10Å shift of a loop to a position that covers the active site, together with smaller movements in five helices and some other loops. These lesser changes were often away from the binding site, in regions connected to the loop with the large movement. These sorts of coupled movements in this case may be for no other reason than they allow the large movement, though in other systems it can allow allosteric binding.

Small Movements of Side-chains

An additional level of domain motion, discussed by Janin and Wodak, 1983, is essentially none at all, but with a few side-chain movements. Serine-proteases binding to macromolecules, where the substrate itself excludes water from the active site, are an example of this. Janin and Chothia, 1990, examined conformational changes in the limited number of enzyme-protein inhibitor complexes and one antibody-protein complex where the structures of both components were available in an unbound form. Recognition sites on the enzyme-inhibitor complexes showed low mobility, but still had small, low-energy conformational changes that improved packing and hydrogen bonding. The antibody-lysozyme complex behaved in a very similar manner. Stanfield and Wilson, 1994, looked at the same complex and saw small rearrangements of V_L with respect to V_H. Antibodies that bound non-protein molecules, such as progesterone-like steroids, short peptides, DNA, and haptens, showed a wide variation of movement, from none to substantial V_H-V_L rearrangements and movements of CDR loops.

Conclusion

From the data available, it appears that protein-protein association often involves much less conformational change than is the case when proteins bind other types of molecules.

1.4 Prediction of the Structures of Protein-Protein Complexes

The prediction of the structure of protein-protein complexes, known as the protein-protein docking problem, is usually defined as follows: given the unbound structures of two proteins that are known to associate, can we predict the structure of their complex? Most attempted solutions to this problem can be separated into two main parts: the generation of many different structures of the complex, and then the selection of a structure from this set that closely resembles the real structure. Other methods use directed searches, such as simulated annealing, but these are not guaranteed to include a structure close to the real complex amongst all of the structures analysed. The aim of this section of the thesis is to review methods that have been entered into two blind trials of predictive protein-protein docking (Strynadka et al., 1996, Dixon, 1997), along with more recent developments that have been tested using unbound structures of components. Methods tested only by re-docking structures taken from the structure of a complex are not examined. Such tests do not give a proper assessment of the likely performance of a method when the structure of the complex is unknown, as would obviously be the case in a biologically useful prediction. For reviews of protein-nucleic acid and protein-small molecule docking algorithms, see Sternberg et al., 1998, and Dixon, 1997. The DAPMATCH protein-protein docking program (Walls and Sternberg, 1992) is reviewed in detail in Chapter Two, as its development was a major part of the work undertaken for this thesis.

1.4.1 Rigid-body Docking Algorithms

All of the algorithms described in this section use the rigid-body approximation, at least initially. This means that the conformation of each protein is kept fixed, and only the six degrees of freedom (three rotations and three translations) that define the orientation of one protein with respect to the other are sampled. This cuts down the number of different possible structures that need to be considered, but necessitates the use of ‘soft’ scoring functions to score those that are generated. Soft functions allow a certain amount of poor complementarity so that small conformational changes that occur on association, and which are not considered by the rigid-body approximation, can be tolerated.

Katchalski-Katzir et al., 1992, digitise the two starting molecules onto a regularly spaced three-dimensional grid. Grid points containing no atoms are given a value of zero, those on the surface are given the value ‘1’, those in the interior of one of the proteins are given large negative values, and finally those in the interior of the other protein are given small and positive numbers. Then all transformations of one molecule with respect to the other are scored by a summation of the products of the values in all grid points. Thus if the surfaces are just touching then the score will be positive, but if there is severe overlap the score will be large and negative. The algorithm is speeded up greatly by Fourier correlation techniques, which are used to calculate simultaneously the scores for every possible translation of the proteins at a fixed rotation. The algorithm is performed in two stages: an initial ‘scan’ stage, using a large grid size for speed, and a second ‘discrimination’ stage at a higher grid resolution, where promising areas from the scan are examined in more detail. In the original paper, this method was tested on only one complex starting from structures of the unbound components. This was a trypsin-trypsin inhibitor complex, and no structure close to the real one was found. This was thought to be a consequence of conformational changes that occur on binding.

The Fourier correlation approach was extended from an assessment of shape complementarity only, as above, by the inclusion of an attempt to match hydrophobic surfaces (Vakser and Aflalo, 1994). The same trypsin-trypsin inhibitor as before was the only complex where a prediction was attempted starting with unbound components, and only a marginal improvement in performance was seen. The overriding problem with this complex appears to be the conformational differences between the bound and unbound structures.

Vakser, 1995, also modified the Fourier correlation approach, but in a different way: by using low resolution grids (with 7Å spacing), he hoped to allow for larger conformational changes than previously. However, this was not tested by starting with structures of unbound proteins.

Gabb et al., 1997, developed the approach of Katchalski-Katzir et al., 1992, and included an electrostatic scoring function that used the same Fourier correlation technique as the original score of shape complementarity. In addition, the algorithm was applied to eight complexes where the unbound structures of both components were available. They found that shape complementarity alone did not provide enough information with which to pick out a structure close to the real complex. The inclusion of electrostatics halved the number of geometries with good scores and placed a correct solution further up the ranking. Different levels of filtering based on biochemical knowledge, from general regions of surface in contact to specific residue-residue interactions, also drastically reduced the number of false positives, and therefore increased the rank of good solutions. Such information could very well be available in a real docking experiment where the structure of the complex is unknown. This is especially true of the types of complex tested by Gabb et al., 1997; the catalytic residues of serine proteases are well known, and antibodies are known to bind antigens on specific parts of their surface called complementarity determining regions.

The DOCK algorithm of Shoichet and Kuntz, 1991, uses a method that attempts to match grooves on the surface of one protein with ridges on the surface of the other. The surface of the first protein is covered with spheres, and clusters of overlapping spheres are kept. These overlaps occur in concave areas of the surface. The size and depth of the concave regions identified depends on the radius of the spheres. The same method is used to identify ridges on the other protein, this time by covering the inside of the surface with spheres. Then the two proteins are brought together by a superposition of each sphere cluster from the first protein onto each sphere cluster of the second. Each superposition is scored based on all atom-atom contacts.

For all three protease-inhibitor complexes considered, the algorithm was able to generate structures within 1Å of the real complex. This was true even when starting from unbound components, although only after selective pruning of some problem side-chains. The challenge, then, is to select these structures from the thousands of others also produced. The authors used various established methods of association energy to evaluate the possible complexes, such as the degree of surface area burial, solvation free energy (which extends the measure of buried surface area through consideration of atom types), packing, biochemical restraints (i.e. only allowing matches which have certain residues, identified from experimental data, in the interface), energy minimisation, and electrostatic interactions (see section 1.2). None of these were found to discriminate reliably between the real structure and false positives, though electrostatic complementarity and energy minimisation performed best. The authors suggest that some of the false positives may represent transient complex structures that could occur on the way to formation of the known structure. The inability of the methods to disregard them as realistic is, however, more likely to be because of missing information or inaccurate representations, as the authors acknowledge.

Cherfils et al., 1991 simplify the structures of the proteins by representing each amino acid as a single sphere, the size of which is proportional to the size of the residue. Five of the six rigid-body degrees of freedom - those that defined the orientation of the molecules with respect to each other - are held fixed, and the simplified representations are brought together along the sixth degree, which is the separation distance of the two molecules. The conformation kept is that given by the smallest separation for which no spheres overlap more than a certain amount. This ‘certain amount’ can be varied, allowing different degrees of soft docking. Once this conformation has been obtained, it is scored by the degree of surface area burial and an approximation of the amount of atomic overlap (rather than the amount of overlap of spheres that was used in the generation stage). Surface area burial is treated as an attractive force and atomic overlap is treated as a repulsive force, and the two measurements are combined together into a pseudo-energy function. At the beginning of the docking simulation, the initial values for the first five degrees of freedom are chosen at random, and the energy of the best conformation, as defined above, is calculated. Then one or two of the angles are changed and the calculation is repeated. The new conformation is accepted if it has lower energy than the previous one. If it has higher energy it is accepted or rejected by a Boltzmann weighted probability that depends on the temperature - the higher the temperature, the more likely it is to be accepted. The cycle then repeats, with gradually decreasing temperature, until no new conformations are accepted. The process is then re-started from another random location. All minima, including the global minimum, should be explored if enough starting points are used. The final step, which does not use the rigid-body approximation, is a refinement of the side-chains of the interface residues of all the resultant conformations. This is done by energy minimisation using the program ‘X-PLOR’ (Brunger et al., 1987). The results when trying to dock unbound trypsin with unbound BPTI, and bound antibody with unbound lysozyme, are close to the native structures, but have fewer hydrogen bonds. It is difficult to compare these results with those of the other algorithms discussed, because no RMSD's or numbers of correctly reproduced interface interactions are reported.

Webster and Rees, 1993, use an approach based on graph theory to match 'key topological features', and therefore to limit the search to more likely areas of interest. For each protein, an ellipsoid containing half of the atoms is generated. Large distances to atoms along normals from the ellipsoid surface identify potentially interesting topological features of the proteins, namely ridges and grooves. Then graphs that connect these features are generated, and the program looks for matching subgraphs between the two proteins. The structures corresponding to these matching subgraphs are scored on their van der Waals and electrostatic interaction energies. A loose constraint on the graph edges (i.e. the distances between features) means that the algorithm is able to generate structures with surfaces that do not match exactly, and so can allow for changes in shape that occur on binding, although with an associated increase in the number of false positives. The original paper did not report results when starting from the unbound structures of proteins, but the approach was tested in the second blind trial of predictive protein-protein docking (Dixon, 1997), which is discussed in section 1.4.3.

1.4.2 Energy and Flexibility Based Filtering

Other work has concentrated on more sophisticated methods of assessing putative complexes than is the case with the above rigid-body soft docking techniques. This has often included explicit allowance of molecular flexibility. Rigid body docking is less computationally intensive, and has been shown to be able to reject many unreasonable structures, and therefore the methods described below have generally been used to filter those structures that remain after a rigid-body search.

Jackson and Sternberg, 1995, developed a description of the thermodynamic processes involved in protein-protein recognition, based mainly on the hydrophobic effect caused by the loss of molecular surface area. This description included electrostatic free energy, hydrophobic free energy, and the loss of conformational entropy cause by the burial of side-chains that were previously accessible to solvent. Lost van der Waals contacts with water are assumed to be compensated for by van der Waals contacts that are gained between the two proteins. Thus the enthalpic contribution to association free energy is completely electrostatic. The electrostatic energy was calculated by the loss of interaction between each protein and the solvent, plus the gain in interaction between the two components in the presence of the solvent. Hydrophobic free energy was modelled as the energy required to make a cavity in the solvent with the same shape and size as the complex, minus that required to make cavities for the two separated proteins. Hydrogen bonding in the interface was optimised by placing polar hydrogen atoms (OH and SH), which have non-specific rotamers in solution, in conformations that gave the lowest energy when interacting with local atoms with hydrogen bonding ability. The loss of conformational entropy was calculated from the empirical scale of Pickett and Sternberg, 1993. This model assumes that side-chains are free to move when they are solvent accessible, but that their conformations are restricted when they become part of the interface. Small amounts of flexibility were allowed for by modifying the interaction energy of atoms that clashed, giving them a value dependent on their separation and that which they would have if their van der Waals surfaces were just touching. This method was used to assess putative complex structures that had previously been found to be indistinguishable by commonly used energy evaluation methods (Shoichet and Kuntz, 1991), and was able to select good structures from false positives in all cases.

Weng et al., 1996, developed a slightly different empirical method to calculate the thermodynamic properties involved in protein-protein association. Atomic solvation parameters (ASPs) were used to relate the area and chemical nature of the solvent-accessible surface to the solvation free energy. The ASPs were derived from experimental free energies of transferring individual amino acids from hydrocarbon to water. These were then used to calculate the transfer free energy of the complex minus the sum of the transfer free energies of the individual components, which gives the solvation free energy. Flexibility of residues that were substantially buried in the interface was modelled by the optimisation of side-chain torsional angles, through minimisation of their van der Waals and electrostatic energies. This algorithm was applied to the same complexes evaluated by Shoichet and Kuntz, 1991, and Jackson and Sternberg, 1995. The lowest energy structures always had an all-atom RMSD between 1Å and 2Å from the real structure of the complex.

Duncan and Olson, 1993, use the electron density of every atom, represented as a Gaussian distribution centred on the centre of the atom. The molecular surface is then defined by a contour over these electron densities, which has fewer discontinuities than the traditional definition of molecular surface given by Richards, 1977. Normals and gradients at different points are calculated by integrating over surrounding points, and the detail described can be altered by contouring at different values of electron density. Complementarity is evaluated by volume overlap and the matching of gradients and normals. The search of conformational space is directed by simulated annealing and an evolutionary algorithm. The paper did not report the use of this algorithm in protein-protein docking, but the procedure was tested in the first blind trial (Strynadka et al., 1996, and see section 1.4.3).

Totrov and Abagyan, 1994, do not use the rigid-body approximation, but simulate molecular flexibility from the outset. Flexibility is confined to the side-chain torsional angles of surface residues to reduce the computational requirements. 120 initial orientations of one protein with respect to the other are chosen by an even sampling of the relevant conformational space. The simulation proceeds from each of these positions by a pseudo-Brownian motion Monte Carlo procedure: the orientation is altered randomly, within certain constraints, and then the conformations of the side-chains are optimised by energy minimisation. This new structure is accepted if it has lower energy than the previous one, or by a Boltzmann weighted probability if it is of higher energy. The procedure is repeated until no new structures are accepted. Thus there are now 120, hopefully improved, orientations. The thirty lowest energy structures are subjected to further local optimisation by more detailed energy minimisation, using interaction and desolvation energy and the loss of side-chain entropy, biased to side-chain conformations that have been seen to be statistically and energetically preferred. This algorithm was applied to the prediction of a lysozyme-antibody complex, starting from the unbound structure of lysozyme. The starting structure of the antibody was that seen in the complex. The lysozyme of the best structure from the first round of the simulation had a backbone RMSD of 5.5Å from the real complex, with only slightly lower energy than poorer predictions. This energy gap, and therefore the discrimination of real from false positives, was improved by the more detailed refinement, in parallel with an improvement of the lysozyme backbone RMSD to 1.6Å.

Jackson et al., 1998, developed another method for filtering putative complex structures, in which rigid-body movements were refined along with side-chain torsion angles. The method used a microscopic treatment of thermodynamics, rather than the continuum description developed previously (Jackson and Sternberg, 1995). This was achieved by the representation of individual water molecules as dipoles. Proteins were modelled with rigid backbones and flexible side-chains, the latter by the use of rotamer libraries, in which all combinations of known possible side-chain torsion angles for each amino acid type are represented. The interaction between water and protein was described by electrostatic, van der Waals and hydrophobic energies. Of the five protease-inhibitor complexes on which the method was tested, a good solution was placed in the top four of up to 364 alternatives. Only two of the four antibody-lysozyme complexes showed reasonable discrimination between true and false positives. This result was attributed to higher conformational change in the interfaces when compared to those of enzyme-inhibitor complexes, and / or the lower specificity of interaction.

1.4.3 Blind Trials of Protein-Protein Docking Algorithms

The methods described above have all been tested in at least one of the two blind trials of predictive docking, and their performances are discussed below. The two trials were the Alberta challenge, Strynadka et al., 1996, and the docking section of the second Critical Assessment of Structure Prediction (CASP2), Dixon, 1997. While the attempt to recreate known structures of complexes from the unbound structures of their components is the only proper way of developing protein-protein docking algorithms, blind trials are vital to ensure that the algorithms have not been unconsciously biased by knowledge of the complexes used. Such trials require that the structure of a complex has been or is about to be solved (but is currently unpublished), and that structures of the unbound components are available. Unfortunately this is a rare situation, ironically because of the difficulty in solving structures of complexes. This explains why there have been only two such trials to date.

The Alberta Challenge

The Alberta challenge (Strynadka et al., 1996) was to predict the structure of the complex between β-lactamase and an inhibitory protein. Two criteria were used to assess the predictions: i) the main-chain RMSD of the whole complex when superposed on the structure of the real complex; and ii) the main-chain RMSD of the inhibitor only, after the predicted and real complex structures had been superposed using just the coordinates of the main-chain atoms of the enzyme.

Of the six groups that entered the challenge, some submitted several geometries that they had ranked by their own particular scoring function (discussed in the sections above) and in some cases by expert knowledge, including knowledge of the location of the active site of the enzyme. One group submitted only the structure that they considered to be the most likely structure of the complex. The best ranked structures in the multiple entries were always those closest to the real complex. They and the single entry all had a whole-complex main-chain RMSD of between 1Å and 2.5Å. Measurement ii) gave higher values for these structures, at between 3Å and 6Å. This reflects the fact that the structure of the enzyme is relatively unchanged by binding, whilst the inhibitor undergoes a small global hinge-bending motion and has some conformational changes in interface loops. These changes were not predicted by any of the six groups, which suggests that it is not necessary to simulate small changes to successfully predict overall complex structure.

Additional, lower-ranked geometries had widely varying all main-chain atom RMSD’s (2-18Å), which illustrates the difficulty in selecting real from false positives. However, the fact that the highest ranked structures were generally good is encouraging, as is the fact that four of the entries contained less than five structures each - this is a reasonable number of predictions to test experimentally.

.

Table 1-1 - Results of the β-Lactamase to Inhibitor Docking Challenge

	Model Generation and / or Search Method	Scoring Method	Reference	Number of Models Submitted	Top Ranked Model[i]		Range of RMSD’s of Other Models (Whole Complex Main-chain) / Å
					Main-chain RMSD of whole complex / Å	RMSD of Main-chain of Inhibitor Only[ii] / Å	
1	Monte Carlo pseudo-Brownian motion, with torsion angle flexibility of surface side-chains. Energy minimisation of side-chains.	Interaction and desolvation energy + loss of side-chain entropy.	Abagyan et al., 1994	3	1.9	4.6	11.3 - 16.2
2	Grid representation of molecules. Even sampling of all relevant search space by Fourier correlation.	Shape complementarity.	Katchalski-Katzir et al., 1992	3	1.1	3.4	13.4 - 14.1
3	Residues as spheres. Monte carlo simulated annealing, then energy minimisation of interface side-chains.	Atomic overlap + burial of surface area.	Cherfils et al., 1991	4	2.5	6.1	2.5 - 16.0
4	Matching of surface grooves and ridges.	All atom-atom contacts.	Shoichet and Kuntz, 1991	15	1.8	3.8	2.3 - 18.7
5	Simulated annealing + evolutionary algorithm.	Volume overlap + surface normal and gradient matching.	Duncan and Olson, 1993	14	1.9	4.5	2.0 - 17.7
6	As per method 4.	Continuum model of thermodynamics.	Jackson and Sternberg, 1995	1	1.9	4.0	N/A

i. Ranked by the group that made the submission.

ii. RMSD of main-chain atoms of the inhibitor, after superposition of the complex using only the main-chain atoms of the enzyme.

CASP2

The target for the protein-protein docking section of CASP2 (Dixon, 1997) was a haemagluttinin-antibody complex. The number of residues in the antigen binding domain of the antibody is greater than 400, and the number in the haemagluttinin is greater than 500. This large size increases the number of different geometries to be considered, and so presents difficulties to predictive docking. This was offset to some extent by two things: i) constraints on the possible sites of interaction were available from a preliminary crystallographic report (Gigant et al., 1995), and from general knowledge of the location of complementarity determining regions on antibodies; and ii) the predictors were provided with the complexed structure of antibody, as no unbound form was available. This further illustrates problems in staging blind trials of predictive docking, namely the lack of suitable test structures. Haemagluttinin was, however, given in an unbound form.

Entries were given a confidence value by their submitters, so that the combined confidence of the structures submitted by each group was equal to one. Each entry was then evaluated by the RMSD of the antibody Cα atoms within 8Å of the interface in the experimental structure, weighted by the specified confidence value. This method concentrates on accuracy in the interface region, and the confidence weighting prevented groups from being evaluated favourably if they employed a scatter-gun approach, i.e. they had hedged their bets by submitting several structures, only one of which was good. However, it could also mean that a poorer geometry could score well if submitted on its own. The message from this is that there is no easy and completely fair way of precisely comparing submissions.

None of the four groups that accepted the challenge were able to predict accurately the real structure of the complex (table 1-2). The best structure submitted had an RMSD of 8.5Å, but was given with several much poorer ones - the weighted RMSD of this entry was 20.5Å. The best weighted RMSD was 9.5Å, from an entry with a single structure. However, this entry correctly predicted less of the epitope residues of haemagluttinin, and none of the residue-residue contacts. Another group got half of the haemagluttinin epitope residues right, but with a very poor RMSD for the structure (15.1Å). This geometry presumably had either the correct part of haemagluttinin bound to the wrong part of the antibody, or bound in approximately the right place but with a severe rotation from the true structure.

The optimistic view from the CASP2 challenge is that even with a large and therefore difficult target, predictive docking can provide information about the location of binding regions which might be unavailable otherwise, and which can be tested experimentally.

Table 1-2 - Results of the Antibody to Haemagluttinin Docking Challenge

	Model Generation and / or Search Method	Scoring Method	Reference	Number of models submitted	Best Model Submitted			Worst RMSD[i] / Å	Weighted Mean RMSD[ii] / Å
					RMSD[i] / Å	% correct residue-residue contacts	% correct haemagluttinin epitope residues		
1	Matching of graphs described by surface grooves and ridges.	Van der Waals and electrostatic interaction energies.	Webster and Rees, 1993	2	30.6	0	0	33.4	32.3
2	Grid representation of molecules. Even sampling of all relevant search space by Fourier correlation.	Shape complementarity.	Katchalski-Katzir et al., 1992	1	9.5	0	23	9.5	9.5
3	As per method 2.	Shape and electrostatic complementarity, followed by refinement (Inc. side-chain flexibility) using a microscopic treatment of thermodynamics.	Gabb et al., 1997, Jackson et al., 1998	8	8.5	14	32	30.9	20.2
4	Matching of surface grooves and ridges (method of Shoichet and Kuntz, 1991)	Solvation free energy from empirical parameters, + interface residue optimisation by van der Waals and electrostatic energy minimisation.	Weng et al., 1996	2	15.1	8	50	19.1	18.3

i. RMSD of the antibody Cα atoms that are within 8Å of the interface in the real complex structure.

ii. RMSD of all models submitted by a group, weighted by the confidence value assigned to the model by the group and divided by the number of models.

1.4.4 Binding Site Prediction

Docking algorithms are often able to predict correctly the structure of a protein-protein complex when some information about the location of the binding sites is known. Experimental biochemical information is not always available, and so several groups have looked at ways of predicting interfaces from sequence and three-dimensional structure alone.

Lichtarge et al., 1996, base their approach on the assumption that the functional sites of proteins from a particular family (i.e. a group of proteins with the same fold and function) will have a common location, and that their constituent residues will have lower mutation rates when compared to other surface regions. Furthermore, where mutations have occurred they indicate functional divergence. The method looks for residue conservation in multiple alignments, and then maps the results onto a representative three-dimensional structure. It was successful at identifying the ligand binding sites of SH2 and SH3 domains, and of DNA binding domains of nuclear hormone receptors, but its general usefulness is limited by two possible problems: lack of multiple sequence data, and the potential for mutation of the interface residues of both components. The second of these problems has been addressed by Pazos et al., 1997.

Pazos et al., 1997, developed a method that uses the assumption that the requirement for specific residue-residue contacts in an interface will be reflected in the sequences of the two interacting proteins, and that evolutionary changes of the interface residues of one protein will be compensated by changes in the other. The detection of such correlated mutations demonstrated i) that in general more highly correlated pairs of positions were spatially closer in the three-dimensional structure, and ii) that by using this knowledge, correct structures of two interacting domains could be distinguished from randomly generated alternatives, and often from structures close to the real one. This method is promising because it can be used to provide information about interface residues in the absence of any three-dimensional structure. However, it is difficult to test on non-covalently bound protein-protein complexes (the main topic of this thesis) because of the lack of known complex structures where the sequences from many different species are also available.

Jones and Thornton, 1997b use patch analysis to predict the locations of protein-protein interaction sites. An accompanying paper (Jones and Thornton, 1997a) analysed interface sites and similar sized patches elsewhere on the surface, looking at solvation potential, residue propensity, hydrophobicity, planarity, protrusion and accessible surface area. The results have been discussed in more detail already (see section 1.2), but in general it was found that these properties were different in interface and non-interface patches. The prediction algorithm uses these observations to assign probabilities of being part of the interface to areas of the protein surface, with the exact combination of properties dependent on the system (homo-dimer, hetero-complex or antibody-antigen complex). Two-thirds of the predictions were considered to be correct, with the other third mostly accounted for by the presence of multiple binding sites. These results are tempered somewhat by the use of the same data set in both the analysis and prediction stages, although this is offset by the patches in the predictions being generated afresh, with their sizes determined from average interface sizes seen in the analysis.

Russell et al., 1998, analysed the binding sites of groups of proteins with common folds that, because of very low sequence similarity, were assumed to be a result of convergent evolution. They were able to detect nine such groups of analogues where the location of the binding site was conserved across all members of the group, though an estimated 40% of such groups were thought to show no common binding site. Related work (Russell, 1998) looked for conserved three-dimensional patterns of side-chains, and identified new ones as well as those previously known, such as the Ser-His-Asp catalytic triad of serine proteases.

1.4.5 Conclusions

The reports on the two blind trials of predictive protein-protein docking (Strynadka et al., 1996, and Dixon, 1997) do not report in detail the biochemical knowledge used by each group to filter their results. As was seen earlier in this introduction, such knowledge can drastically improve the results of predictions, and therefore it is difficult to compare fairly the various docking algorithms. Also, table 1-1 and table 1-2 show that there is no general approach that is clearly better than the others. Use of methods for predicting the location of binding sites in future blind trials may well increase the accuracy of structures submitted.

1.5 Thesis Outline

The aims of the work presented in this thesis were two-fold. Firstly, a protein-protein docking algorithm previously produced by this laboratory (Walls and Sternberg, 1992) was investigated and developed. This development took the form of a re-implementation of the algorithm in a more available form, and a detailed analysis of its behaviour. The analysis led to changes in the scoring function, and elaborations such as side-chain truncation and a treatment of electrostatic complementarity. This work is presented in Chapter Two. The problems that were encountered in allowing for conformational change, and the lack of a general analysis of this in the literature, were the motivating factor behind the work presented in the next three chapters. Methods used for measuring structural differences are given in Chapter Three, along with the structures to which they were applied. Chapter Four gives the results of the methods when applied to independently solved structures of identical proteins. These data act as controls for Chapter Five, in which structures of proteins in unbound forms are compared with their structures when complexed. Chapter Six discusses the implications that this work has for modelling, and concluding remarks are made in Chapter Seven.

Chapter Two

Development of a Protein-Protein Docking Algorithm

2.1 Introduction

There are more than seven thousand protein structures currently available in the Brookhaven Protein Data Bank (the PDB), of which less than two hundred are of protein-protein complexes. Therefore the prediction of the structure of protein-protein complexes from the structures of their unbound components is one of the major goals of molecular modelling. This ‘docking problem’ is usually defined in the following way: given the three-dimensional structure of two proteins that are known to associate, can the structure of their complex be predicted? For a solution to this problem to guarantee that a structure close to the real structure of the complex is generated, many different structures, evenly spaced over the whole of the relevant conformational space must be produced. The problem then becomes one of picking the correct structure from the list.

This chapter presents the development and refinement of a specific docking algorithm known as DAPMatch, originated by a previous student in the laboratory (Walls and Sternberg, 1992). The use of surface complementarity to evaluate potential complexes is investigated.

The DAPMatch algorithm is described in the next section. It was intended to be the first stage in a complete docking procedure, reducing the set of millions of candidate structures down to just a few hundred. These few hundred would then be analysed by more sophisticated methods (such as a continuum model of the thermodynamic processes involved (Jackson and Sternberg, 1995), or a multi-copy method of side-chain optimisation (Jackson et al., 1998)). These are more able to pick out the correct structure, but are too computationally intensive to be used on the large initial set.

DAPMatch was originally developed on antibody - protein antigen complexes, and to exploit the specialised parallel architecture of a computer that can perform thousands of operations simultaneously. However, a predictive docking method for other types of complex is required, and the parallel computer is not widely available. It was intended that the algorithm would be converted to run on serial architecture machines when increases in their power made this practical, and the results of such a modification are described in this chapter, along with developments intended to improve the results. The applicability to other biological systems (specifically enzyme-inhibitor complexes) is investigated.

2.2 Original Algorithm

2.2.1 Methods

DAPMatch was designed to run on a 64 × 64 processor parallel architecture machine (an AMT DAP). It is described in detail by Walls and Sternberg, 1992, and an outline is given below, with particular emphasis on the details needed to explain subsequent work.

Structural Data

The algorithm was applied to three antibody - protein antigen complexes (table 2-1), and developed using the HyHel10 system. Docking simulations were performed using the structures of the antibodies in their bound form, plus one modelled structure of antibody D1.3, and that of the antigen (lysozyme in each case) in an unbound form.

Table 2-1 - Protein Structure Data Used in the Original DAPMatch Algorithm

Protein	PDB Code	Resolution / Å
Antibody HyHel-10 - Lysozyme Complex	3hfm (Padlan et al., 1989)	3.0
Antibody HyHel-5 - Lysozyme Complex	2hfl (Sheriff et al., 1987)	2.5
Antibody D1.3- Lysozyme Complex	(from Dr. S. Phillips)	2.8
Lysozyme	6lyz (Diamond, 1974)	2.0
Antibody D1.3 (model)	(from Dr. A. Lesk)	N/A

Summary of Algorithm

The algorithm is summarised in figure 2-1. The procedure starts with the structures of the components of the complex in an unbound form. These structures are treated as rigid-bodies, which means that no internal degrees of freedom are considered. Thus the number of degrees of freedom is reduced from thousands to just six. These six are sampled in the following way:

a) The assumption is made that the protein is roughly spherical, and that an even division of the surface of a sphere will give an even division of the surface of the protein. The surface of the sphere is divided by regular tessellation of an icosahedron, to produce 432 uniformly distributed points.

b) The coordinate centres of the tessellated icosahedron and of the first protein are superposed.

c) Both are rotated together so that each point in turn is uppermost in the z-axis. For each point, a 32 × 32Å slice of the protein is taken. These slices are centred on the relevant point, and are in a plane that is perpendicular to the z-axis. Each slice is divided into 64 × 64 half Angstrom squares. For each of these elements the maximum height to the van der Waals surface of the protein is taken. The heights are then discretised into 64 blocks of 0.25Å each, and smoothed to reduce the effects of small conformational changes caused by complex formation. The maximum height is therefore 16Å, and anything 16Å or more below this is set to zero.

d) This slicing process is repeated for the second protein, which completes the sampling of four of the six degrees of freedom.

e) The fifth is sampled by, for one protein only, rotating about each surface point in 8° steps, and slicing the surface as before. This gives 45 slices for each surface point.

f) Then, for all possible pairs of slices of the first and second protein, the slice of the second protein is turned upside down (by inversion in the z-axis), and both slices are brought together along the z-axis so that they are just touching. The surface complemetarity is scored as described in the next section, and then the slices are moved together in eleven 0.5Å steps (giving a maximum overlap of 5Å), with surface complementarity scored at each step. Thus the sixth and final degree of freedom is sampled. An additional level of sampling was performed using small shifts in the plane of the slice at each separation, because the DAP provides routines that can do this quickly. Any match which had less than 2000 pairs of elements that both contain non-zero heights was discarded. This corresponds to an overlap of 500Å^2, and so the number of improbable matches is reduced. For each pair of slices, only one match is kept. This is the one with the separation and in-plane translation that give the best score.

A full search therefore involves evaluating 2,099,520,000 different possible structures of the complex:

= 432 slices of first protein

× 432 slices of second protein

× 45 rotations

× 5 in-plane translations in the x-axis

× 5 in-plane translations in the y-axis

× 11 separations in the z-axis

In the original work this was reduced to 342,199,000 by only making 64 slices of the antibodies. These were centred around the complementarity determining region.

Finally, the resulting list is reduced in size by several methods. Clustering discards orientations that are similar to one with a better score. Matches with good electrostatic complementarity are picked out by a simple function - a single sphere represents each side-chain, and a value of +1, 0, or -1 on each sphere represents the charge on the residue. The orientations are then scored by a residue-residue interaction energy ('+1' with '-1' is good, for example), summed over all interactions. Also, orientations that do not match known binding regions, or do not allow known and specific residue-residue interactions, are removed.

Scoring Function

A softened Lennard-Jones potential 'V_{soft}' (see figure 2-2), which allows for unfavourable surface matching caused by differences between bound and unbound structures, is used.

$$V_{soft}(x) = \begin{cases} 256x^4 & & x < 0Å \\ 4x^2 & \text{when} & 0A \le x \le 4Å \\ 64 & & x > 4Å \end{cases}$$

where x is the distance between surfaces, and is negative when surfaces overlap, zero when they just touch, and positive when they are separated. The potential is summed over all pairs of surface elements.

Every slice contains areas with no surface mapped, particularly at the edges. These areas need to be corrected for, otherwise matches of slices covering more surface area would have worse scores simply because more elements contribute to the score. This is corrected for as below:

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

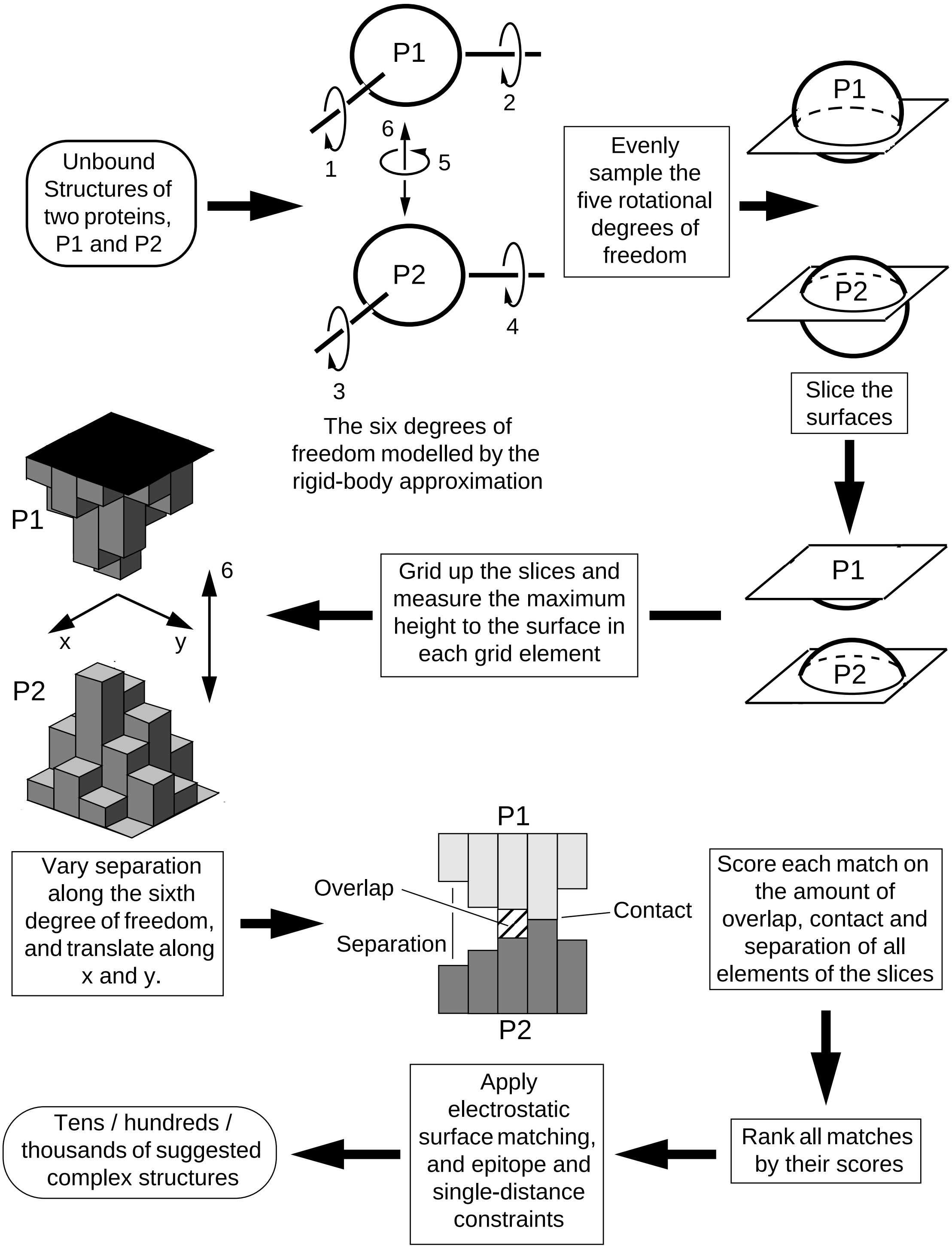

Unbound Structures of two proteins, P1 and P2
P1
P2
1
2
3
4
5
6
The six degrees of freedom modelled by the rigid-body approximation
Evenly sample the five rotational degrees of freedom
P1
P2
Slice the surfaces
P1
P2
Grid up the slices and measure the maximum height to the surface in each grid element
P1
P2
x
y
6
Vary separation along the sixth degree of freedom, and translate along x and y.
P1
Overlap
Separation
Contact
P2
Score each match on the amount of overlap, contact and separation of all elements of the slices
Rank all matches by their scores
Apply electrostatic surface matching, and epitope and single-distance constraints
Tens / hundreds / thousands of suggested complex structures

Figure 2-1 - Summary of the DAPMatch Algorithm

where N_{overlap} is the number of pairs of elements in the match whose elements both contain surface information. The algorithm therefore favours burial of large amounts of surface.

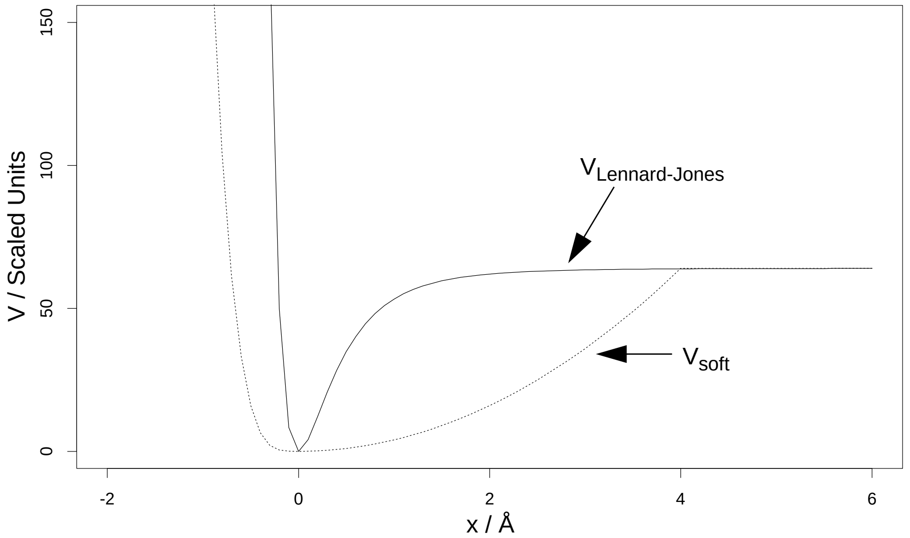


Figure 2-2 - The Original Soft Potential Compared to a Lennard-Jones Type Potential

$$V_{\text{Lennard-Jones}} = 4\varepsilon\left\{\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right\} \qquad V_{\text{soft}} = \begin{cases} 256x^4 & \text{when} \quad x < 0\text{Å} \\ 4x^2 & \qquad 0\text{Å} \le x \le 4\text{Å} \\ 64 & \qquad x > 4\text{Å} \end{cases}$$

where ε = the well depth, r = the distance between atom centres, σ = the distance at which V = 0, x = the surface separation, and r = x - 2. The Lennard-Jones potential has been scaled to the range of the soft potential by setting ε to the well depth of the soft potential (64) and by adding ε.
The soft potential is more tolerant of overlap of surfaces or spaces between them.

2.2.2 Results

The results were varied for the different complexes. D1.3 was predicted well, with the best solution found fifth in a list of twenty-five structures that remained after filtering. This structure had a Cα RMSD of the lysozyme equal to 1.7Å. The best solution for HyHel10 came third in a list of eighteen, but with a slightly poorer lysozyme Cα RMSD of 3.4Å. HyHel5 and the model of D1.3 performed the worst, with, respectively, the best solution thirtieth out of forty and with a Cα RMSD of 7.5Å, and ninth out of fifteen with a Cα RMSD of 11.4Å. However, these structures were reasonable in the interface, with Cα

RMSD's of 3.5Å and 4.8Å. All four predictions showed a tendency for more separation in the interface than in the real complexes, which the original authors (Walls and Sternberg, 1992) said indicates allowance of side-chain movement. This is true if the movements are towards the interface. If they are away from the interface (i.e. if the side-chains would clash if they did not move), then it indicates that the scoring function is not sufficiently soft.

2.3 Program Development

2.3.1 Methods

Conversion to Serial Architecture

The DAP is a specialised parallel architecture machine, and as such it is not available to the majority of people who might be interested in using a docking program. Because of this, it was decided to rewrite DAPMatch to run completely on serial architecture machines, which are in much wider use. The initial conversion left the algorithm essentially unchanged from that described above, the differences being in its implementation. However, clustering, electrostatic scoring, and epitope and single-distance constraints were not applied at first, as the primary requirement was to investigate how well surface complementarity was measured.

Structural Data

The original program was developed on and applied to three antibody - protein antigen complexes (table 2-1). To test and to improve the performance with other systems, the data set was extended to include three enzyme-inhibitor complexes (table 2-2). No attempt was made to dock the modelled structure of antibody D1.3 because of the poor performance of the original algorithm on this system. The structure of the D1.3-lysozyme complex was from the PDB rather than from Dr. S. Phillips. To reduce the amount of unnecessary computation, only the variable domains (chosen by eye) of the antibodies were used. This is reasonable because these domains contain the complementarity determining regions (CDR’s) where all known antigens bind. Even if the structures of the real complexes were unknown, it would be assumed that the antigens bind to the CDR’s.

Table 2-2 - Structural Data Used in the Development of DAPMatch

Structure	State	PDB Code	Resolution / Å
Antibody-Antigen Complexes			
D1.3 Fab - Lysozyme	Complex	1fdl (Fischmann et al., 1991)	2.5
HyHel5 Fab - Lysozyme	Complex	2hfl (Sheriff et al., 1987)	2.5
HyHel10 Fab - Lysozyme	Complex	3hfm (Padlan et al., 1989)	3.0
Lysozyme	Unbound	6lyz (Diamond, 1974)	2.0
Enzyme-Inhibitor Complexes			
Subtilisin - Chymotrypsin Inhibitor	Complex	2sni (McPhalen and James, 1988)	2.1
Subtilisin	Unbound	1sbc (Neidhart and Petsko, 1988)	2.5
Chymotrypsin Inhibitor	Unbound	2ci2 (McPhalen and James, 1987)	2.0
Chymotrypsin - Ovomucoid	Complex	1cho (Fujinaga et al., 1987)	1.8
Chymotrypsin	Unbound	5cha (Blevins and Tulinsky, 1985)	1.7
Ovomucoid	Unbound	2ovo (Bode et al., 1985)	1.5
Trypsin - Pancreatic Trypsin Inhibitor	Complex	2ptc (Marquart et al., 1983)	1.9
Trypsin	Unbound	2ptn (Walter et al., 1982)	1.6
Pancreatic Trypsin Inhibitor	Unbound	4pti (Marquart et al., 1983)	1.5

Root Mean Square Deviation

Differences between predicted structures and real structures were measured by calculating the Root Mean Square Deviation (RMSD) of their Cα atoms. The RMSD between a set of *N* atoms from structure *a* and *N* equivalent atoms from structure *b* is given by the following equation:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i^b - x_i^a)^2 + (y_i^b - y_i^a)^2 + (z_i^b - z_i^a)^2}{N}}$$

where x, y, and z are the coordinates of the atoms.

Structural Superposition

Pairs of proteins were superposed on their Cα atoms by the least squares fitting algorithm of McLachlan, 1979. This algorithm takes two equivalent sets of atoms, *a* and *b*, and calculates the transformation matrix that minimises the RMSD between them (see “Root

Mean Square Deviation” on page 55). The implementation of Suhail Islam (personal communication) was used, and pairs of proteins were superposed on their Cα atoms only.

The Best Possible Results

There are differences between the structures in their complexed and unbound forms. This means that it is impossible for any rigid-body docking algorithm to predict a structure for the complex that is exactly the same as the real one. The minimum Cα RMSD that can be achieved is given by a superposition of the structures of the unbound proteins on to those in the complex (see “Structural Superposition” on page 55). The values for the data set used in this work are given in table 2-3.

When the complex structures are reconstructed from the transformations given by the DAPMatch algorithm, the largest protein is held in a fixed position (the position given by its superposition) and the smaller protein is oriented in a position relative to this. Therefore the measure used to assess the quality of the structures generated is the Cα RMSD between the smaller protein and its real structure. In the six complexes used (table 2-2), the smallest protein is the antigen or the inhibitor.

Table 2-3 - The Best Possible Answers that can be Expected From Rigid-body Docking

Complex	Cα RMSD between unbound and complexed structures / Å	
	Antibody[i]	Antigen
D1.3 - Lysozyme	0.00	0.46
HyHel5 - Lysozyme	0.00	0.48
HyHel10 - Lysozyme	0.00	0.58
	Enzyme	Inhibitor
Subtilisin - Chymotrypsin Inhibitor	0.54	0.46
Chymotrypsin - Ovomucoid	0.47	1.16
Trypsin - Pancreatic Trypsin Inhibitor	0.34	1.10

[i] No structures of the antibodies in an unbound form were available.

Restriction of Search Space

To allow the search space to be limited to regions known to be important, a program was written to produce a cap of the tessellated icosahedron around a specified residue and to a specified size (figure 2-3). This allows some epitope information to be included from the beginning of the procedure rather than as a filter at the end. The program therefore runs more quickly, which is beneficial both for prediction and for development, because the effect of program changes on the results can be seen more quickly.

The program takes the protein structure and the tessellated icosahedron points, superposed on their coordinate centres. A vector from the centre to the specified atom is calculated. The size of the cap is specified by a cone angle, which gives the maximum angle allowed between this vector and an equivalent vector to each of the icosahedron points. The cap contains only those points that are within the cone angle from the atom. These points are then used in the main algorithm to locate surface slices.

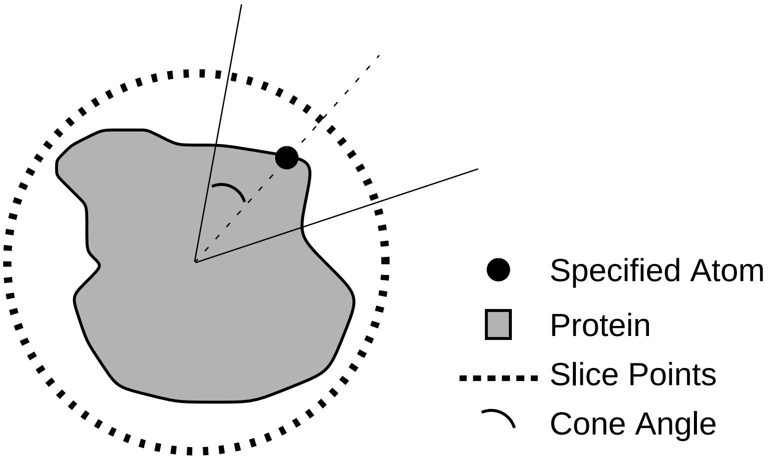


Figure 2-3 - Restricting the Search Space
The program takes the protein structure and the tessellated icosahedron points, and produces a cap containing only those points that are within a specified cone angle from a specified atom. These points are then used in the main algorithm to locate surface slices.

In a real docking prediction, the atoms used to restrict the search space would be ones at the centre of a known binding region. In development, the atoms used were chosen by taking the unbound components superposed on to their positions in the complex. A line that connects their coordinate centres was drawn, and then the atom in each that was closest to this line was found. The cone angle was selected by starting with a small value, generating the appropriate structures and finding the one with the lowest Cα RMSD. The angle was then gradually increased until no structure with a better Cα RMSD was generated. The same cone angle (20°) was used for all the structures. The sizes of the caps produced, measured by the number of icosahedron points that they contain, can be different because of the differing positions of the specified atoms with respect to the icosahedron points.

Truncation of Side-chains.

There are differences between the structures of a protein in a complex and in its unbound form. These differences mean that the surfaces of the unbound components are not as complementary to each other as they are in the complex, and this can cause problems for predictive docking.

The structures of the three enzyme-inhibitor complexes in table 2-2 and the structures proposed by DAPMatch were examined visually. It was seen that the structure with the lowest Cα RMSD had more surface clash than both the real complex and the structure with the best score. There are two ways of dealing with this without explicitly modelling flexibility: use a soft scoring function, or truncate the offending side-chains. Both these methods reduce the detrimental effect that surface overlap has on the scores. However, side-chain truncation can be applied to specific side-chains. Therefore it can be equivalent to having a residue specific scoring function, which varies according to how likely it is that a particular residue has different conformations in the unbound and complexed structures.

Implementing this idea requires a decision as to which side-chains need truncating, and to what level. As a first test, a visual inspection of the structure with the lowest RMS identified side-chains that clash, and these side-chains were truncated to Cβ. However, this method of side-chain selection is obviously not one that could be used in predictive docking as it requires knowledge of the structure of the complex. The first systematic approach tried was to prune all side-chains down to their Cβ atoms, and then only down to Cγ.

Use of Molecular Surface

Much of the van der Waals surface of a protein is buried in the interior, and it models atoms as spheres without considering the interactions between them. A better representation of molecular surface has been given by Richards, 1977, who defines it in two parts (figure 2-4). The first part is any portion of the van der Waals surface which touches a probe sphere rolled across it, and is known as the contact surface. The second part is called the re-entrant surface. It is produced when the probe sphere simultaneously touches the van der Waals surface of more than one atom, and is that part of the probe sphere bounded by these contacts. The algorithm of Connolly, 1983, with a probe radius of 1.4Å, was used to calculate the molecular surface.

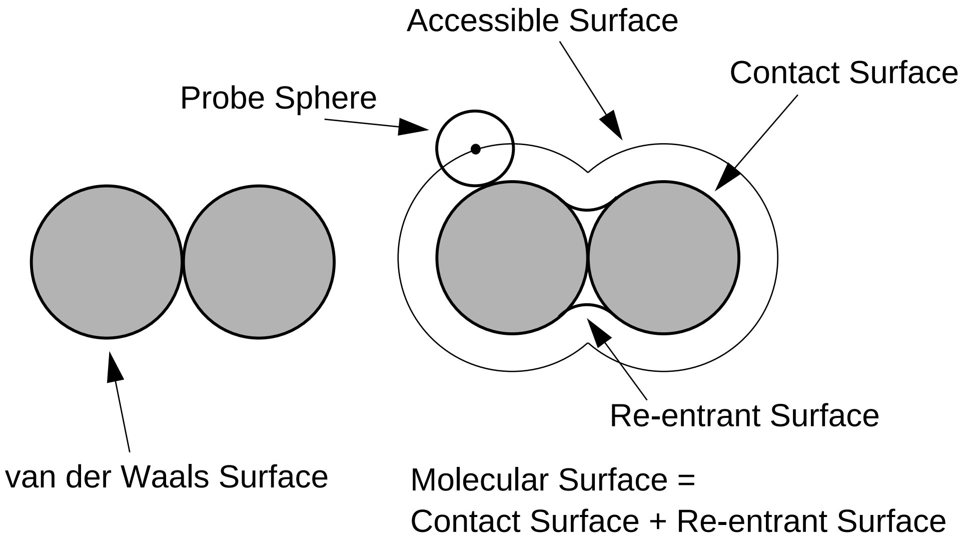


Figure 2-4 - Definition of Different Surfaces
Accessible surface area was defined by Lee and Richards, 1971, and molecular surface by Richards, 1977.

DAPMatch was altered to produce slices of the molecular surface. However, the precision of the slices is such that the differences between a molecular surface slice and a van der Waals surface slice are likely to be minimal.

Truncation Based on Side-chain Exposure

A surface related to the molecular surface, known as the solvent accessible surface, was defined by Lee and Richards, 1971. It is the area mapped out by the centroid of the probe sphere as it rolls over the van der Waals surface (figure 2-4).

The exposure of each residue was measured by the relative accessible surface area (ASA) of its side-chain. The ASA was calculated using the implementation of Suhail Islam (personal communication) and a probe radius of 1.4Å. Relative ASA is the ASA compared to that of the residue in an extended form. The ASA of the extended form is defined by Miller et al., 1987. All residues, except prolines, with a relative side-chain ASA of 80% or more were truncated to Cβ.

Further Restriction of Search Space

To speed up investigation and development of the scoring function and of side-chain truncation schemes, it was decided to restrict the search space still further. Thus for each complex, the first four rotational degrees of freedom (see figure 2-1) were fixed at angles that include the best structure. These angles were chosen from the results of the work described above and in the results section. With these angles fixed, only the rotation and separation on the z-axis and the in-planes translations were varied. The number of orientations analysed was therefore 16,500

= 1 slice of first protein

× 1 slice of second protein

× 60 rotations

× 5 in-plane translations in the x-axis

× 5 in-plane translations in the y-axis

× 11 separations in the z-axis

, and the result of every orientation was stored.

Analysis of Scoring Function

The scoring function was analysed by, for each complex, comparing the match that had the best score with that which represented the structure with the lowest Cα RMSD. For each match, a count of the number of pairs of height elements at a certain distance apart was made. This count was done for every distance represented. In this way it can be seen whether the best scored match has, for example, less elements that overlap than is the case with the structure closest to the real complex. If this were true then it would imply that the scoring function was not sufficiently soft.

The results were used to suggest new scoring functions that addressed the differences between the best scored and the best RMSD matches.

Scoring Electrostatic Complementarity

Point charges on atoms do not model the propagation of charge through the protein and solvent environments, or the effect that the shape of the protein has on the electrostatic surface (Honig and Nicholls, 1995). Electrostatic potentials do not suffer from these limitations, and so the use of a measure of the complementarity of electrostatic surfaces in predictive docking was evaluated. A more correct method would be to measure the force that the charges of one protein experience in the electrostatic field of the other. However, studies of crystal structures of complexes have shown that they involve complementary electrostatic surfaces (Honig and Nicholls, 1995). This observation, combined with the sensitivity of point charges to local conformational changes, justifies the approach outlined here.

Electrostatic potentials for all atoms of each protein were calculated using the program ‘Delphi’ (Nicholls and Honig, 1991), which gives a numerical solution to the Poisson-Boltzmann equation:

$$\nabla \bullet [\varepsilon(r) \nabla \phi(r)] - \varepsilon(r) \kappa^2(r) \sinh[\phi(r)] + 4\pi\rho(r) = 0$$

∇ is the derivative with respect to spatial coordinates. $\varepsilon(r)$ is the dielectric constant at point r, and is a macroscopic property that represents the shielding of charges by the medium in which they sit. $\phi(r)$ is the electrostatic potential at point r in units of kT/q, where k = the Boltzmann constant, T = the absolute temperature, and q = the charge on a proton. $\rho(r)$ is the charge density at point r. κ is the Debye-Hückel parameter, where $\kappa^2 = 8\pi q^2 I/ekT$ and I is the ionic strength.

The first term in the equation represents the electrostatic potential when there are no free charges present and the dielectric constant is different at different positions in space. Water is a highly polarisable medium, and therefore it has a high shielding effect on charges. The interior of proteins have a low shielding effect. Consequently, calculations were performed with $\varepsilon = 80$ for the exterior of the protein (i.e the solvent) and $\varepsilon = 2$ for the interior, as per Nicholls and Honig, 1991. The second term represents the presence of mobile ions and their screening effect on the electrostatic potential, and the third term represents the presence of charges.

These potentials were projected onto the molecular surface. The surface was then sliced and grided in the manner described previously, except that each grid element had an electrostatic potential as well as a height associated with it. The potentials were contoured so that everything below -2kT is classed as negative, everything above +2kT is treated as positive, and everything in between is neutral. These contour values were chosen by visual inspection of GRASP representations of electrostatic surfaces (Nicholls et al., 1991) to identify values which clearly indicated complementarity, and are the same as those used by Honig and Nicholls, 1995. Matches are then scored by a simple function which gives a value of -1 to a match of a positive and a negative element, +1 to a match of negative with negative or positive with positive, and zero for everything else, summed over all elements that are 4Å or closer to each other.

2.3.2 Results

Replication of Results from Original DAPMatch

One run of the initial version of serialised DAPMatch, covering the same search space as the original work (Walls and Sternberg, 1992), would have taken about 20 days on a Silicon Graphics 150MHz R4400 processor, the fastest computer available to us when this work was carried out. A complete search, not restricted to the CDR's of the antibodies, would take over four months. This made it impractical to compare directly the old and new versions of DAPMatch. This is not a problem because the first instance of serial DAPMatch was a simple conversion, using the same parameters and scoring function, and so would give the same results as the original. Also, Walls and Sternberg, 1992 filtered the results after scoring shape complementarity (see section 2.2.1), which was not done here as I wanted to develop the searching and scoring functions. It is therefore important to get results from the simple conversion before making any modifications, so that the effects of these modifications can be assessed properly. The searches were restricted by the information given in table 2-4 (see "Restriction of Search Space" on page 57), and the results are given in table 2-5.

Table 2-4 - Specification of the Search Space Used in Development

Complex[i]	Protein 1		Protein 2		Number of Matches Stored[iv]
	Atom[ii]	Size[iii]	Atom[ii]	Size[iii]	
	Antibody		*Antigen*		
1fdl	2,437	13	210	12	9,360
2hfl	2,032	15	396	12	10,800
3hfm	2,437	12	728	14	10,080
	Enzyme		*Inhibitor*		
2sni	1,526	12	524	12	8,640
1cho	1,370	13	130	12	9,360
2ptc	1,289	16	281	14	13,440

i. The complexes are identified by the PDB code of the structure of the complex (see table 2-2).

ii. The atom used to produce the icosahedron cap, identified by the atom number record in the PDB file of the unbound structure.

iii. The size of the icosahedron cap, given by the number of points, produced by a 20° cone angle centred on the specified atom.

iv. The number of orientations stored by the program = size of cap of protein 1 × number of rotations about z × size of cap of protein 2. The rotations were performed in 6° steps, so the number of rotations = 60.

Table 2-5 - Structure Quality and Selection With and Without Side-chain Pruning

	Complex[i]											
	1fdl		2hfl		3hfm		2sni		1cho		2ptc	
Stage	Cα RMSD[ii] / Å	Rank[iii]	Cα RMSD[ii] / Å	Rank[iii]	Cα RMSD[ii] / Å	Rank[iii]	Cα RMSD[ii] / Å	Rank[iii]	Cα RMSD[ii] / Å	Rank[iii]	Cα RMSD[ii] / Å	Rank[iii]
(Number of Structures Stored)	9,360		10,800		10,080		8,640		9,360		13,440	
Serial DAPMatch	0.9	39	2.5	4,625	2.6	515	4.3	2,684	3.1	4,342	4.0	3,018
Selective Pruning to Cβ	-	-	-	-	-	-	2.5	2,651	2.1	18	1.6	715
All residues Pruned to Cβ	1.7	2,128	1.9	1,128	2.4	3,265	2.2	5,976	1.5	4,976	1.3	6,968
All residues Pruned to Cγ	1.5	1,144	1.4	1,841	1.7	2,332	2.7	7,074	1.9	5,557	1.5	1,226
Molecular Surface Matching	1.3	138	3.3	14	2.9	670	5.6	6,391	4.3	8,268	4.5	3,398

i. The complexes are identified by the PDB code of the structure of the complex (see table 2-2).
ii. The lowest Cα RMSD found between the unbound structure of the component that is mobile in the simulation and its structure in the real complex. The mobile components are the antigens and the inhibitors.
iii. The rank with respect to the scoring function.

For the three antibody-lysozyme complexes, the results are qualitatively similar to those in the original work (Walls and Sternberg, 1992, and see section 2.2.2). 1fdl performs best, with the best structure having a Cα RMSD of 0.9Å and ranked 39th in a list of 9360. 3hfm does slightly worse - the best structure has a Cα RMSD of 2.6Å and is ranked 515th in a list of 10080. 2hfl performs badly, as before, with the best structure (Cα RMSD = 2.5Å) ranked at position 4625 in a list of 10800.

The enzyme-inhibitor systems all perform badly, with the best structures having Cα RMSD's of between 3.1 and 4.3Å, and ranked between a quarter and half way down the lists of all structures stored.

Side-chain Truncation

Table 2-5 shows the lowest Cα RMSD found for each system with and without side-chain truncation. The position of the appropriate structure in the score-ordered list of all structures (the rank) is also given.

In this exploratory study, selective pruning was only performed on the three enzyme-inhibitor systems, as these were not predicted successfully without pruning (see above). The residues that were pruned are given in table 2-6. The new set of structures stored contained one with a better RMS than found previously. The rank of this structure was also improved (table 2-5).

Table 2-6 - Residues that Protrude into the Interface

Complex[i]	Protruding Residues[ii]	
	Enzyme	Inhibitor
2sni	Ser221, His64	Ile56, Thr58, Met59
1cho	-	Met18
2ptc	-	Lys15, Arg17, Arg39

i. Identified by the PDB code of the complex.
ii. Selected by visual inspection of the unbound structures superposed on to the complex.

Pruning of all side-chains was performed on all six systems, as the aim was to develop a generally applicable method. Table 2-5 shows that structures that had lower Cα RMSD's than before were found somewhere in the lists, except for the D1.3 antibody - lysozyme complex (1fdl). However, these structures, with the exception of that for the HyHel5 antibody - lysozyme complex (2hfl), all had worse rank. The differences in performance when pruning to Cβ or Cγ were neither dramatic or consistent across all six systems.

These results indicate the need for a systematic pruning method that can reliably pick out only those side-chains that would be picked by eye. This was attempted by pruning based on side-chain exposure, which was done after molecular surface matching and so the results are given in the section with that heading.

The Lowest Cα RMSD Structure Generated

The matching algorithm tries different separations and in-plane translations for every pair of surface slices, but only stores the one with the best score (see “Summary of Algorithm” on page 48). If the scoring function is not working accurately this structure is not necessarily the one with the lowest Cα RMSD. Keeping all structures revealed some with better Cα RMSD’s, but only by around 0.5Å compared to those in table 2-5. However, the range of the separations and translations give a Cα RMSD of up to 7.5Å between two structures described by the same two maps. This is a substantial amount, and therefore to see the effects of different scoring functions properly, it is necessary to store all the structures that are generated.

In the original algorithm the separations and translations are changed by amounts that are larger than the precision of the maps. Increments at that precision level did not produce any structures with substantially better Cα RMSD’s. In fact it is possible to have sampling that is more coarse than originally while still being able to generate reasonable structures. This gives a simple way to reduce the number of different structures that are analysed, and therefore the run-time of the program, without markedly affecting the quality of the results.

Molecular Surface Matching

When using serialised DAPMatch with no side-chain truncation and searching in the same areas as before, the results for molecular surface matching were worse than those for van der Waals matching (compare the second and last rows of table 2-5). This was particularly the case for the enzyme-inhibitor complexes. The exception was the HyHel5 antibody - lysozyme complex (PDB code: 2hfl). In this case the best Cα RMSD found was slightly worse than before (3.3Å compared to 2.5Å), but the structure with this RMSD was ranked fourteenth out of 10,800, rather than at position 4,625. These differences are probably because the original scoring function was developed for use with van der Waals surfaces.

To investigate and develop the scoring function and side-chain truncation scheme, it was decided to examine a very restricted area around the correct answer (see “Further Restriction of Search Space” on page 60). This allowed the program to run much quicker than before, and it also made it feasible for all orientations (i.e. including all separations and in-plane translations) to be stored. The program was run in this restricted area, using the original potential function and no side-chain truncation, and the scoring function was analysed. Figure 2-5 shows, for each complex, a graph of number of pairs of grid elements against surface separation for both the best scored match and the match that gave the best Cα RMSD structure. The form of the scoring function is different in three regions, which correspond to surface overlap, close contacts, and separation (see figure 2-2). The differences in the counts for each pair of matches in all three regions were examined (figure 2-2). These plots show that the best RMSD structures always have more surface overlap (separation < 0Å), usually have more elements close together (0Å ≤ separation ≤ 4Å), and always have less space > 4Å between surfaces when compared to the structures with the best scores. This implies that the scoring function should have a broader minimum, allowing more clash and not quite as much separation as at present.

Two new functions were developed (figure 2-6) to give improved scoring of the best RMSD matches, based on the results in figure 2-5. Both of these ($V_{soft\#2}$ and $V_{soft\#3}$) allow more overlap of surfaces than the original scoring function ($V_{soft\#1}$). Some of the graphs in figure 2-5 indicate that surface separation in the region 0-4Å is concentrated at the lower end of this range. Hence $V_{soft\#2}$ gives a slightly lower score than $V_{soft\#1}$ at this lower end, but rises more quickly, reaching the maximum score for separation at 3Å. However, some of the graphs indicate the opposite. Also, it may become important to allow some space between surfaces when using structures with truncated side-chains. This would be the case especially if some of the truncated side-chains are not those that have positions which differ between the unbound and complexed structures. Therefore $V_{soft\#3}$ is more lenient for the whole of the range 0-4Å, and in fact does not reach the maximum score for surface separation until the separation equals 5Å.

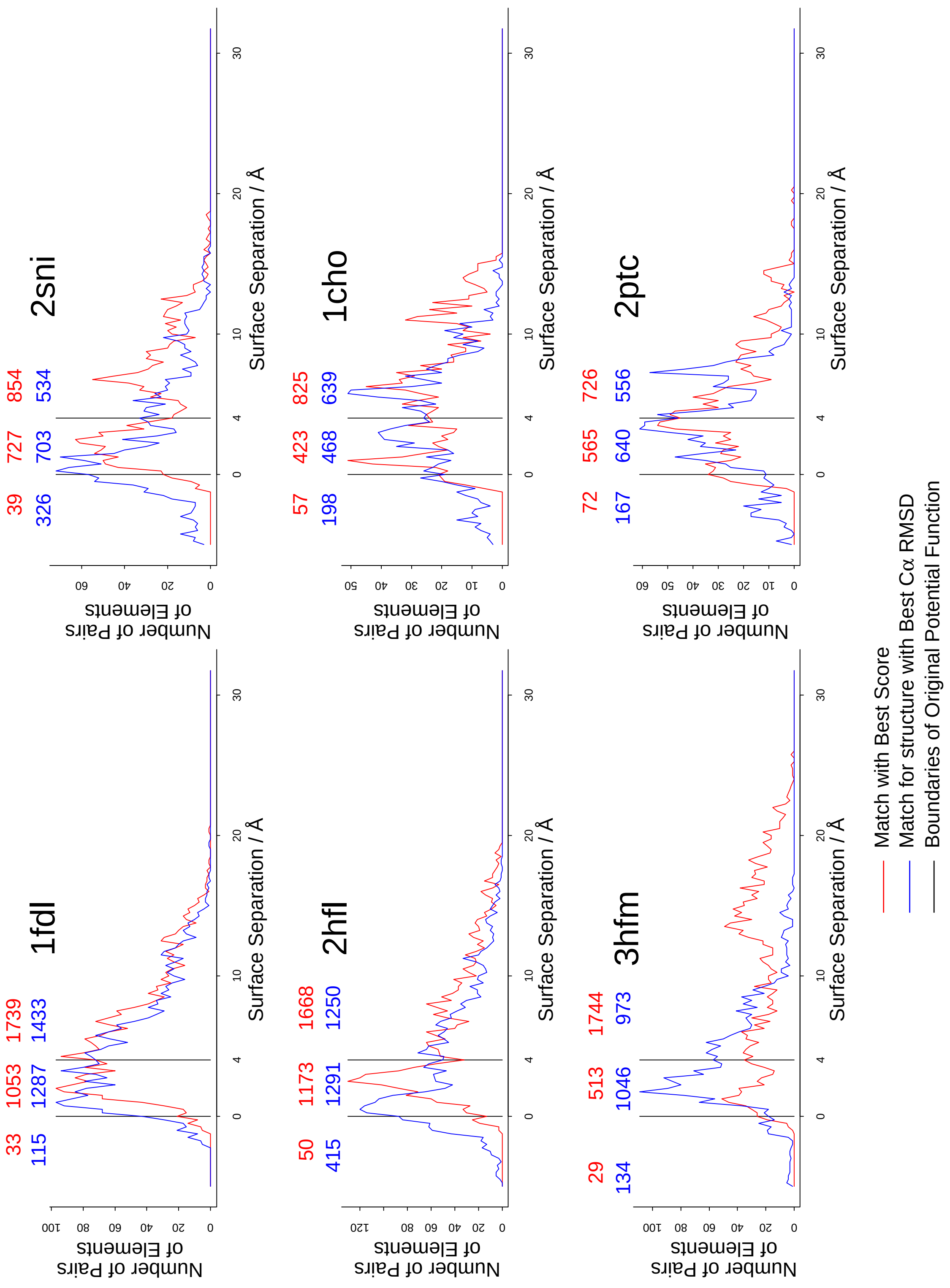


Figure 2-5 - Comparison of the Best Scored and the Best Cα RMSD Matches

The plots are identified by the PDB code of the complex, and the matches are from the search restricted to the area about the correct answer (see “Further Restriction of Search Space” on page 60). The original scoring function has different forms for scoring surface overlap (separation < 0Å), close contacts (0Å ≤ separation ≤ 4Å), and separations greater than 4Å (see figure 2-2). These three sections are divided on the plots by the black vertical lines. The numbers above each section give the total number of pairs of elements with separations in that range. The matches for the best RMSD structures always have more surface overlap, usually have more close contacts, and always have less space > 4Å when compared to the best scored matches.

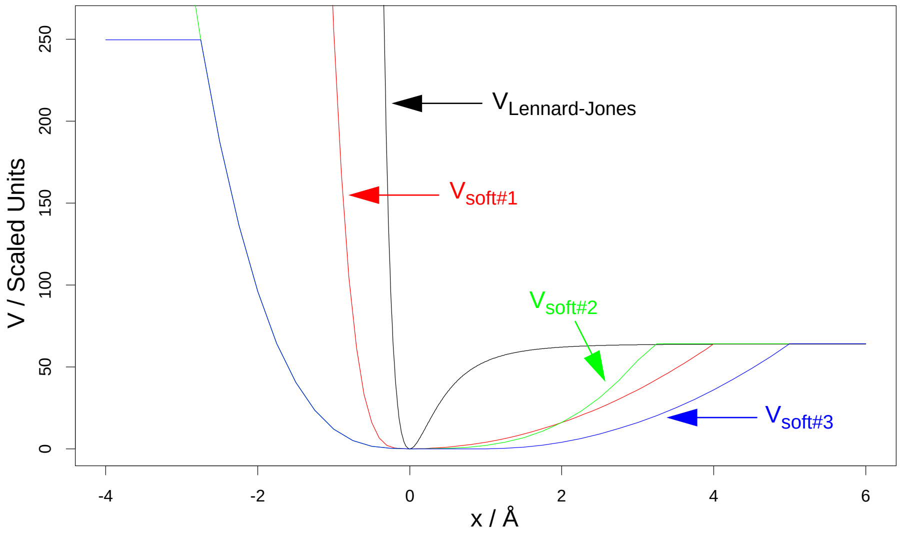


Figure 2-6 - Differences Between the Old and New Soft Potentials

All three soft potentials allow more overlap of surfaces or spaces between them than the Lennard-Jones potential ($V_{Lennard\text{-}Jones}$). $V_{soft\#1}$ is the potential used in the original DAPMatch algorithm (Walls and Sternberg, 1992). $V_{soft\#2}$ and $V_{soft\#3}$ are two new potentials used in an attempt to improve the score of the correct structure in docking predictions. Both allow more overlap of surfaces than $V_{soft\#1}$ does, with $V_{soft\#3}$ the most lenient. $V_{soft\#2}$ has a higher penalty for space between surfaces than $V_{soft\#1}$, and $V_{soft\#3}$ has less.

$$V_{soft\#2}(x) = \begin{cases} 12|x|^3 & \text{when} \quad x < 0 \\ 2x^3 & \quad 0 \le x \le 3 \\ 64 & \quad x > 3 \end{cases}$$

$$V_{soft\#3} = \begin{cases} 12 \times 2.75^3 & \quad x < -2.5 \\ 12|x|^3 & \quad -2.5 \le x < 0 \\ 0 & \text{when} \quad 0 \le x \le 1 \\ 4(x-1)^2 & \quad 1 < x \le 5 \\ 64 & \quad x > 5 \end{cases}$$

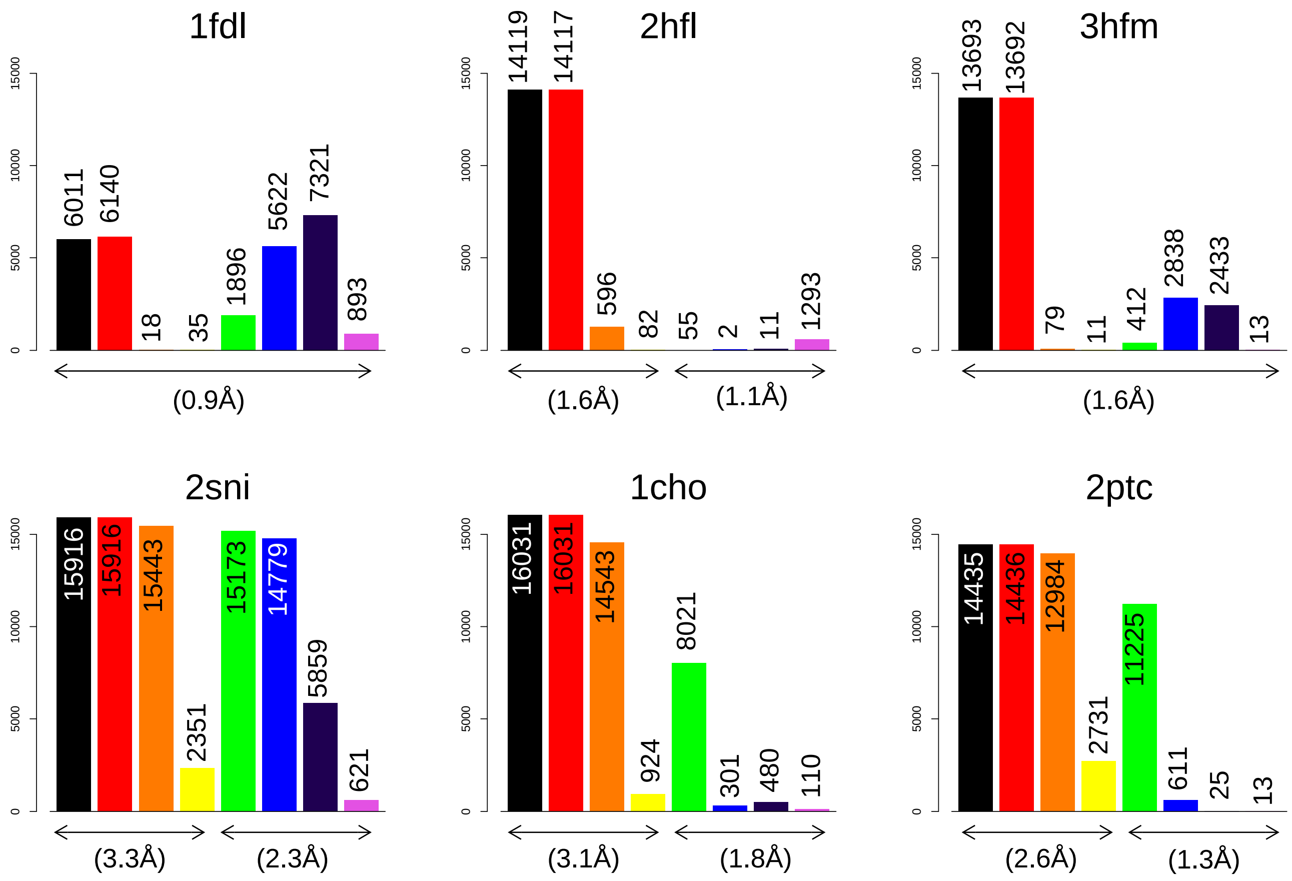
where ε = the well depth, r = the distance between atom centres, σ = the distance at which V = 0, x = the surface separation, and r = x - 2. The Lennard-Jones potential has been scaled to the range of the soft potentials by setting ε to the well depth of the soft potential (64) and by adding ε.

In addition to the modification of the form of the scoring functions, the weight given to the number of overlapping elements was altered. In the original function, 100 × the number of overlapping elements was subtracted from the score (see “Scoring Function” on page 50). This was done to ensure that matches of slices covering only a small amount of surface did not score favourably simply because they are empty. It is equivalent to subtracting 100 from the score of each pair of matched elements. Since the well depth of the scoring function (the difference in the minimum value and the value at infinite separation) is 64, this means that even large separations are scored favourably. For an element that contains at least some height, the minimum height is 0.25Å and the maximum is 16Å (see “Summary of Algorithm” on page 48). Therefore the maximum separation of a pair of elements is 31.5Å. It was decided that such a pair of elements should not score more favourably than a pair where one or both contained no surface. This was done by subtracting 64 × the number of overlapping elements from the score. In effect, the scoring functions now tend to zero at large separations, which is more physically realistic.

Figure 2-7 shows a comparison of the performance of the different scoring functions in the restricted search. The effects of three different side-chain truncation schemes are also shown. These were: no truncation, truncation of all side-chains to Cγ, and truncation of exposed residues to Cβ.

The ranks of the best structures for the antibody-antigen complexes, without truncation, were considerably better when using $V_{soft\#2}$ or $V_{soft\#3}$ instead of the original function. For the enzyme-inhibitor complexes, $V_{soft\#2}$ gave a slight improvement but $V_{soft\#3}$ made a vast difference, with the best structure in the top one to three thousand rather than in the bottom three thousand of sixteen thousand matches. Cγ truncation of all side-chains was more beneficial for the enzyme-inhibitor complexes, which may indicate more induced fit on binding than in the antibody-antigen cases, especially since the antibody structures used were taken directly from the complexes. The approach that worked best overall was side-chain truncation of exposed residues, and scoring using $V_{soft\#3}$. All six complexes had their best structure in the top one thousand, with several performing much better than this. The different weighting of the number of overlapping elements made little difference. This could be because the search was done in a very narrow area about the correct answer, and so the number of overlapping elements does not vary substantially.

To reduce the number of matches analysed, and therefore to increase the speed of the program, coarser levels of sampling of the search space were tried. Rotations, separations and in-plane translations in increments that were twice as big as before were used. The total number of orientations represented in a search restricted to the area around the



■	$V_{soft\#1.1}$	Unpruned
🟥	$V_{soft\#1.2}$	Unpruned
🟧	$V_{soft\#2}$	Unpruned
🟨	$V_{soft\#3}$	Unpruned
🟩	$V_{soft\#1.2}$	Cγ pruned
🟦	$V_{soft\#2}$	Cγ pruned
🟪	$V_{soft\#3}$	Cγ pruned
🟪	$V_{soft\#3}$	asa80 pruned

$V_{soft\#1}$ is the original scoring function (see figure 2-2). $V_{soft\#2}$ and $V_{soft\#3}$ are as described in figure 2-6.

For each match, the number of pairs of overlapping elements is multiplied by 100 and subtracted from the total $V_{soft\#1.1}$. It is weighted by 64 in the other functions.

Side-chains were kept in their entirety ('Unpruned'), or all were truncated to Cγ ('Cγ pruned'), or were truncated to Cβ if their relative accessible surface area was greater than 80% ('asa80 pruned')

Numbers in brackets are the Cα RMSD of the best structure found in the list of 16500. The y-axis gives the position of this structure in the score-ordered list, as do the numbers above each bar.

Each system is identified by the PDB code of the complex (see table 2-2).

Figure 2-7 - The Performance of Different Scoring and Side-chain Truncation Schemes

correct answer was thus 1620. $V_{soft\#3}$ was used, with side-chain truncation based on exposure. For each complex, there was a structure that had a Cα RMSD of 2.5Å or less in the top one hundred scored matches. In four cases this structure was in the top ten. If the same sampling rates were applied to a complete search of both components of the complex, the procedure would take less than two weeks on a Silicon Graphics 150MHz R4400 processor. This compares well with the four months required at the previous levels of sampling, especially considering increases in speed available with more modern computers.

These favourable results encouraged a much wider search of conformational space, using the same level of coarse sampling and the same exposure-based side-chain truncation with the new scoring function. These searches covered the whole of the lysozyme and the complete CDR region of the antibodies, and a similar area for the enzyme-inhibitor systems. They involved the scoring of over 40 million different orientations. Memory and disk limitations made it impractical to store the results for all of these. Therefore the matching program was altered to rank the matches as it went along, and to keep only the top few thousand. This is justified because if the structure closest to the real complex is not in the top one hundred or so, the results are unusable by any subsequent refinement procedures. Five of the six systems had no structure with a Cα RMSD lower than 8Å in the top one thousand best scored structures. The exception was the D1.3 antibody - lysozyme complex (1fdl), where the best Cα RMSD in the top one thousand was 2.0Å, ranked 805th. None of the six systems had a structure in the top one hundred that had a Cα RMSD less than 10Å. In all six cases, the best scored structures had larger and flatter interfaces than those closest to the real complexes (figure 2-8).

Electrostatic Surface Matching

The electrostatic scoring function was developed on the trypsin - BPTI complex (2ptc). This complex was chosen because Honig and Nicholls, 1995, demonstrated by visual inspection that its two components have complementary electrostatic potential surfaces. The same 40 million orientations as above were evaluated. In the best twenty thousand scored structures, none had a Cα RMSD less than 11Å. As with the steric score, the best scored structure had a larger and flatter interface than in that closest to the real complex (figure 2-9). There is considerable surface clash in this structure, which would not have

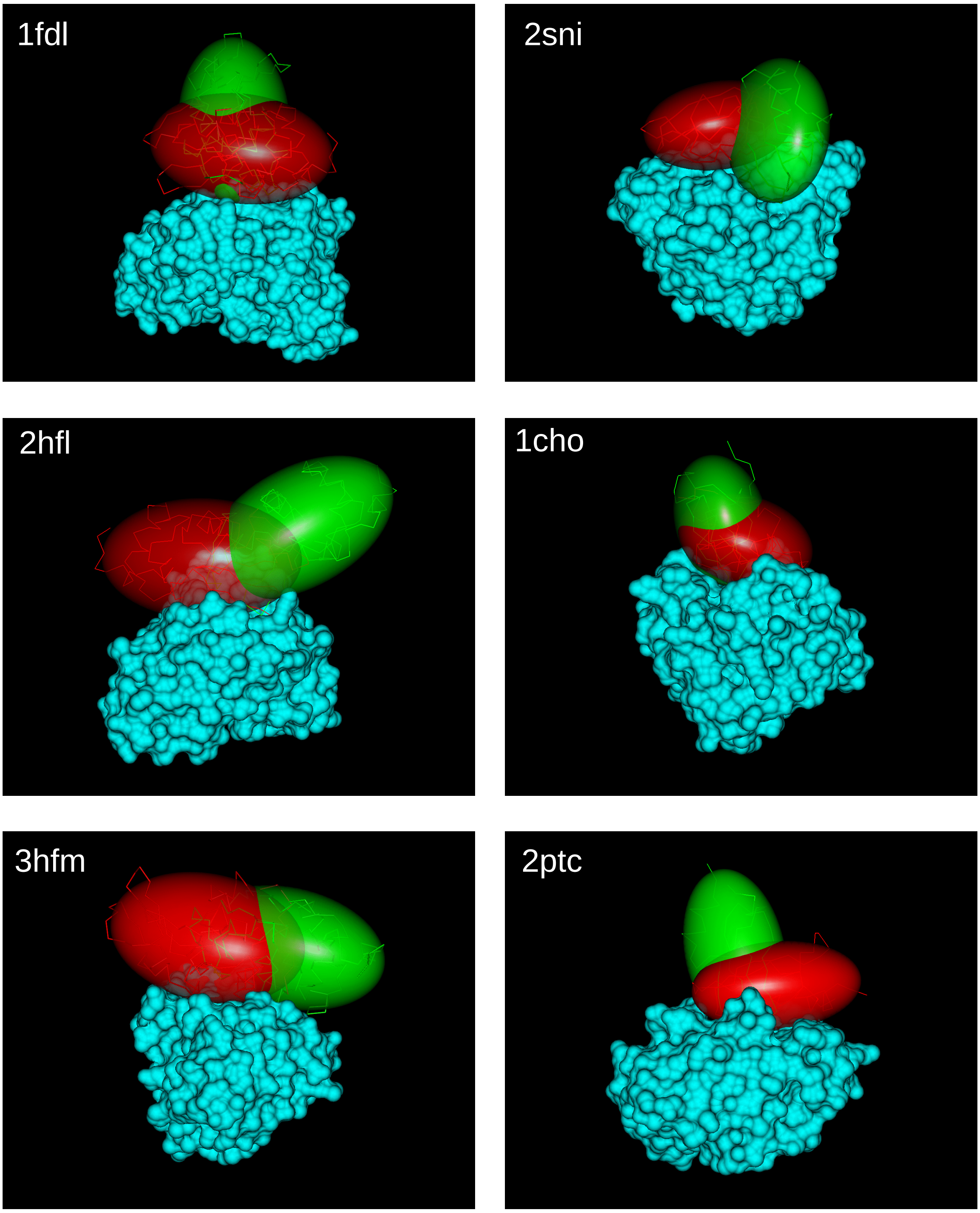


Figure 2-8 - Structures of False Positives and Correct Answers

The structures ranked highest by the shape scoring function (represented by red ellipses) all have larger and flatter interfaces than the structures closest to the real complex (represented by green ellipses). The searches involved the whole of the antigens and inhibitors. For the antibodies, the whole of the CDR regions were covered. Similarly sized sections of the enzymes, centred on the binding sites, were used. The antibodies and enzymes (shown as cyan molecular surfaces) are held in a fixed position in the simulation, hence only one orientation is shown for each. Systems are identified by the PDB code of the complex (table 2-2).

been allowed if the steric scoring function had also been applied. They were not applied together because it is not clear how they should be weighted with respect to each other.

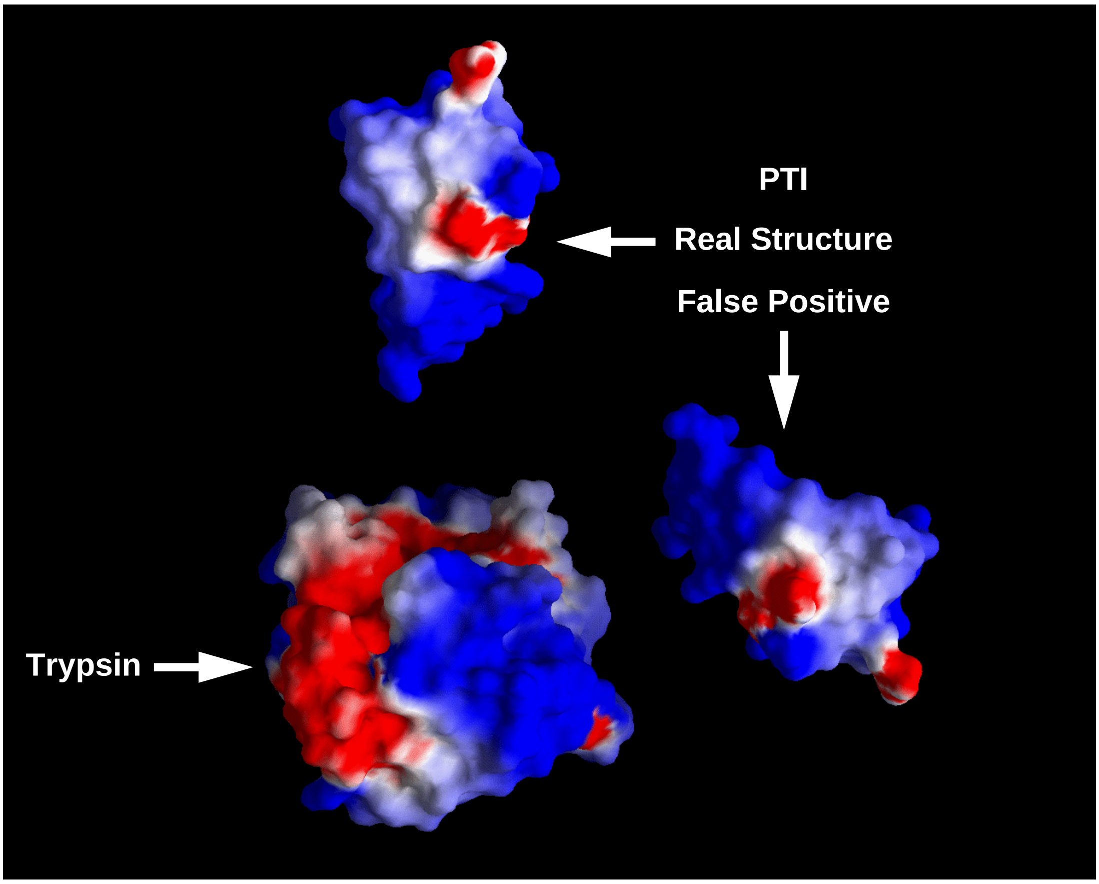


Figure 2-9 - Comparison of the False Positive from Electrostatic Surface Matching of 2ptc with the Real Structure

Surfaces are shown as GRASP representations (Nicholls et al., 1991), coloured by electrostatic potential (red = negative, blue = positive, white = neutral). For both the real structure and the false positive, PTI has been separated from trypsin by translating along the line that connects the two coordinate centres. This has been done to give a better view of the interacting surfaces. The false positive has a larger and flatter interface than the real complex.

2.4 Discussion and Conclusions

The DAPMatch algorithm of Walls and Sternberg, 1992 has been re-written to run on serial architecture computers. Reduced levels of sampling of search space have enabled a complete search to be performed in under two weeks, as opposed to the four months or more required by the first serialised version. Other such reductions of search space, together with the increased speed of modern computers, are likely to reduce the computational time still further.

The program now uses molecular surface (Richards, 1977) rather than van der Waals surface. Exposed side-chains are truncated, and the scoring function has been softened. All three developments improve the results when looking in a narrow region of search-space centred on the correct answer. However, they do not significantly improve the results in a complete search, and it is likely that other information, such as electrostatic complementarity and knowledge of the epitope, will still be necessary to select the correct structure from thousands of possibilities suggested by the program.

A visual analysis of the false positives indicated that their interfaces were larger and flatter than those of the real complexes. This suggests two things: that the representation of the shape of the surface is poor, and / or that shape complementarity is not sufficient to predict the structure of protein-protein complexes. The projection of surfaces onto slices and the scoring of matches of these slices has two problems (figure 2-10). Both are caused by the loss of information in directions perpendicular to that of the projection. Highly-complementary invaginated interfaces score the same as flat interfaces (figure 2-10a), which explains why the false-positives have larger and flatter interfaces. Also, insertions with overhang can show steric clash when there is none (figure 2-10b). This second point may explain some of the beneficial effects of side-chain truncation that were seen, as it is likely that truncation removes such insertions. It may also explain why side-chain truncation does not always help with docking algorithms that do not use surface projection (Gabb et al., 1997).

The results suggest that methods which use surface projection, such as DAPMatch (Walls and Sternberg, 1992) and PUZZLE (Helmer-Citterich and Tramontano, 1994), lose

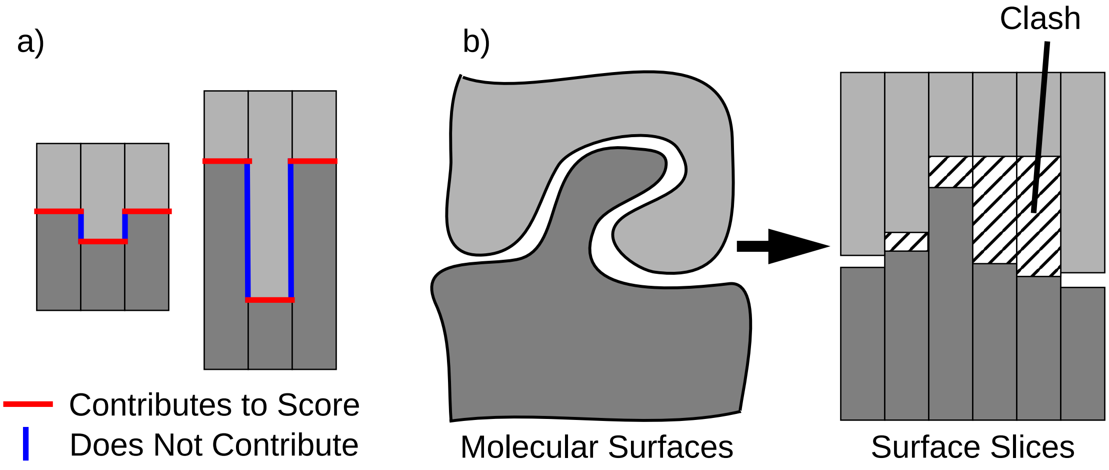


Figure 2-10 - Problems with DAPMatch Surface Representation

a) DAPMatch only scores surfaces in the z-direction, and any contact perpendicular to this is ignored. Therefore the second slice in this diagram would have exactly the same score as the first, despite obviously being a better fit.

b) DAPMatch surface slices are a projection of the molecular surface in the z-direction. Interfaces with invaginated surfaces may be given an unfavourable score because the surface slices falsely indicate clash.

information necessary to evaluate shape complementarity properly. Indeed, the PUZZLE algorithm has been substantially modified (Ausiello et al., 1997), and now uses a different surface representation. Several protein-protein docking algorithms that do not use a projected view of the surface had been published at the time that the work described here was done (see Chapter One). The method of Katchalski-Katzir et al., 1992 performed best in a blind prediction of the binding of β-lactamase inhibitory protein to TEM-1 β-lactamase (Strynadka et al., 1996), and it does not suffer from the problems discussed here. Hence it was developed in this laboratory by Gabb et al., 1997. Side-chain truncation was also investigated, as it proved useful in the development of DAPMatch.

Chapter Three

Data and Methods

3.1 Introduction

In the previous chapter it was demonstrated that predictive docking of proteins by shape complementarity and by using the rigid-body approximation is aided by soft potential functions and side-chain truncation. Both of these approaches allow for conformational changes that occur on association. However, there has been no large scale analysis of the nature of these conformational changes (see Chapter One); the methods are being developed without reference to a general analysis. This was the case because until recently there were few proteins whose structures had been solved in both complexed and unbound forms, from which comparisons could be made. This chapter presents data and methods used to address this problem.

An important additional analysis is that of the extent of conformational difference that exists simply because of experimental differences in structure determination. This gives a measure of the importance of the results from the main investigation. It also gives values for the likely precision with which structural predictions can be made. A data set of pairs of independently solved structures of identical proteins is given, on which this analysis can be performed.

3.2 Structural Data

3.2.1 Structure Quality

All of the structures used were solved by X-ray crystallography and were available in the April 1996 release of the Brookhaven Protein Data Bank (PDB). All had a resolution of 2.8Å or better, and had been refined to an R-factor of approximately 0.2. These conditions were chosen because they mean that the structures are defined to a precision that allows conformational differences between structures to be observed (see Chapter One). Resolution and refinement were identified automatically from PDB files. However, there is no fixed format for their specification in these files, meaning that some structures may have been missed. The large number of structures in the PDB meant that examination of the files by hand was impractical.

Residues identified in a comment in the relevant paper or PDB file as having poor electron density were excluded from calculations of conformational change, as were those residues containing one or more atoms with a B-value greater than or equal to 50Å^2. A B-value of 50Å^2 corresponds to an RMSD of 0.8Å (see figure 1-1), which is approximately half the length of a carbon-carbon bond (Engh and Huber, 1991). The conformation of these residues is expected to differ more than that of others because of uncertainty in their position, or high mobility (see Chapter One).

3.2.2 Use of SCOP Classifications to Identify Identical Proteins

The Structural Classification Of Proteins (SCOP) database (Murzin et al., 1995) classifies proteins on the basis of their structural and evolutionary relationships. The hierarchy of the classification system is as follows:

a) The Class level is based on secondary structure content, and is divided into four sections: all α, all β, mixed α and β, or $\alpha + \beta$.

b) The Fold level clusters proteins with the same topological connections and three-dimensional arrangement of their secondary structure elements.

c) The Superfamily level clusters proteins with low sequence identities, but whose structural and functional features suggest a common evolutionary origin.

d) The Family level clusters proteins whose structures and functions indicate clear evolutionary relationships.

e) The Protein level gives the specific name, and therefore function, of a protein.

f) The Species level indicates the organism in which the protein was found.

In this thesis, proteins were considered to be different if their classifications from the April 1996 release of SCOP differed at any of these levels.

This approach was taken because the information contained in PDB files does not, on its own, allow identical proteins to be easily identified by computer. This is because the PDB format has no fixed way of naming the proteins. The SCOP authors have used a combination of expert knowledge and use of computers where appropriate (such as for comparing sequences), the results of which have greatly simplified the task of identifying identical proteins.

3.2.3 Independently Solved Structures of Identical Proteins

As a control for analysing conformational change, it is necessary to obtain a value for the differences in structure caused by experimental differences in the determination of crystal structures. To this end, pairs of independently solved crystal structures of identical proteins were investigated. A similar analysis has been performed by another group (Flores et al., 1993). Their work was not used here because it was desirable to take advantage of the structures deposited in the database since that work was done, and also because additional information that they did not give was required. For example the differences in the structures of exposed residues, and the differences of individual residues grouped by their amino acid type.

The April 1996 release of the SCOP database (Murzin et al., 1995) was searched for sets of non-complexed structures with 100% identical sequence, and no insertions or deletions. This was done to ensure that any structural differences seen were not due to differences at the sequence level. In addition to the basic structural criteria (see section 3.2.1), only structures with no heteroatoms (except waters) were considered. These will be missed by the sequence checks, and could cause structural differences if present in only one member of a pair. However, pairs of structures with the same heteroatom bound in the same place will also be excluded. It would be difficult to precisely compare the location of heteroatoms in two structures, and small differences could cause potentially large differences in protein conformation.

One SCOP class occasionally gave more than one set of structures that agreed with the conditions above. These corresponded to sets of mutants as well as a set for the wild-type protein. In these cases the native set was chosen, although it could just as easily have been one of the others. If any of these sets contained more than two structures, then the two structures with the best resolution were used. If there were still more than two structures in any set, the two most recently solved structures were chosen. Also, the PDB files of the structures were examined to ensure, as far as possible, that the members of each pair were solved independently.

Twelve pairs were found (table 3-1). Members of each pair were solved in the same space group as each other, except turkey lysozyme (PDB codes 135l and 2lz2). This was also the only pair whose resolutions were not very similar (1.3Å and 2.2Å). Refinement procedures were not always the same for members of each pair. This means that any different systematic errors caused by the different procedures will show up in this analysis. In addition, experimental conditions such as pH and concentration were not always the same. These differences are justified in the context of the comparisons made with pairs of complexed and unbound structures, where the space groups, resolutions, refinement methods and conditions often differ.

Table 3-1 - Pairs of Independently Solved Structures of Identical Proteins

Protein - *Species*	Structure 1				Structure 2			
	PDB Code	Resolution / Å	Space Group	Refinement Method[i]	PDB Code	Resolution / Å	Space Group	Refinement Method
Lysozyme - *Turkey Egg White*	135l	1.3	P 21	X-PLOR	2lz2	2.2	P 61 2 2	PROLSQ
	(Harata, 1993)				(Parsons and Phillips, TBP)			
Basic Fibroblast Growth Factor - *Human*	1bfg	1.6	P 1	PROLSQ	1bas	1.9	P 1	X-PLOR
	(Ago et al., 1991)				(Zhu et al., 1991)			
DNA Polymerase β – *Rat*	1bpb	2.3	P 21 21 2	TNT	1rpl	2.3	P 21 21 2	X-PLOR
	(Sawaya et al., 1994)				(Davies et al., 1994)			
Aspartic Proteinase - *HIV-1*	1hhp	2.7	P 41 21 2	X-PLOR	3phv	2.7	P 41 21 2	X-PLOR
	(Spinelli et al., 1991)				(Lapatto et al., 1989)			
Lysozyme - *Hen Egg White*	1lza	1.6	P 43 21 2	PROLSQ	1lsa	1.7	P 43 21 2	PROLSQ
	(Maenaka et al., 1995)				(Kurinov and Harrison, 1995)			
Interleukin-4 - *Human*	1rcb	2.3	P 41 21 2	X-PLOR + PROLSQ	2int	2.4	P 41 21 2	X-PLOR + PROLSQ
	(Wlodawer et al., 1992)				(Walter et al., 1992)			
Ribonuclease A - *Cow*	1rhb	1.5	P 21	PROLSQ	1rat	1.5	P 21	PROLSQ
	(Kishan et al., 1995)				(Tilton et al., 1992)			
Interleukin-1 β – *Human*	2i1b	2.0	P 43	PROLSQ	4i1b	2.0	P 43	X-PLOR + Restrain
	(Priestle et al., 1989)				(Veerapandian et al., 1992)			

Table 3-1 - Pairs of Independently Solved Structures of Identical Proteins (Continued)

Protein - *Species*	Structure 1				Structure 2			
	PDB Code	Resolution / Å	Space Group	Refinement Method[i]	PDB Code	Resolution / Å	Space Group	Refinement Method
Transforming Growth Factor β – *Human*	2tgi	1.8	P 32 2 1	TNT	1tfg	2.0	P 32 2 1	X-PLOR + TNT
	(Daopin et al., 1992)				(Schlunegger and Gruetter, 1992)			
CD4 - *Human*	3cd4	2.2	C 2	X-PLOR	1cdh	2.3	P 2	X-PLOR + PROLSQ
	(Garrett et al., 1993)				(Ryu et al., 1994)			
Pepsinogen - *Pig*	3psg	1.7	C 2	TNT	2psg	1.8	C 2	PROLSQ
	(Hartsuck et al., 1992)				(Sielecki et al., 1991)			
Chymosin B - *Cow*	4cms	2.2	I 2 2 2	X-PLOR + RESTRAIN	1cms	2.3	I 2 2 2	PROLSQ
	(Newman et al., 1991)				(Gilliland et al., 1990)			

[i] X-PLOR - Brunger et al., 1987. PROLSQ - Konnert and Hendrickson, 1980. TNT - Tronrud et al., 1987. RESTRAIN - Driessen et al., 1989. See section 1.1 for a description of these methods.

3.2.4 Complexed and Unbound Structures

The PDB format has no standard way of specifying that a structure is of a protein-protein complex, and so the following strategy was used to identify such structures. A file containing the sequences of everything in the April 1996 PDB, except DNA / RNA, short chains, or multiple copies of NMR structures, was produced by parsing each coordinate entry (Rob Russell, personal communication). Entries for structures containing only one chain were removed, leaving information from 1447 different PDB files. The percentage identities of different chains of each of these structures were calculated using the program 'multalign' (Barton and Sternberg, 1987). Structures were removed from this file when all their chains had greater than 95% sequence identity to each other. In such cases it is unlikely that the components are able to exist individually. This left 508 structures. Any structures that did not conform to the structural requirements mentioned before (see section 3.2.1) were removed from the list, and the PDB files of those remaining were examined by hand. Theoretical models and structures with only Cα coordinates were removed, together with multi-chain structures that were not complexes, such as insulin, viral coat proteins, proteins cleaved into several chains, and antibodies. After this, ninety-three structures of protein-protein complexes remained. These ninety-three structures represented sixty different complexes. Two complexes were judged to be different when the SCOP classifications of either of their components differed at any level (see section 3.2.2). When more than one structure was available for a particular complex, the one with the best resolution was chosen. If more than one structure had this resolution, the most recently solved was used.

For each component of the complexes, SCOP classifications were used to identify structures of unbound forms with identical classifications (see section 3.2.2). For eight of the complexes the structures of both components in unbound forms were also available. Another twenty-three had one unbound component available, giving a total of thirty-nine proteins whose structures had been solved in both complexed and unbound forms (table 3-2).

Eighteen of the complexes are enzyme-inhibitors, seven are antibody-antigens, and the remaining six are of other types. One of these six is a methylamine dehydrogenase heterotetramer, H_2L_2, bound to two molecules of amicyanin. However, each amicyanin molecule is in contact with the H and L subunits of only one HL dimer (Chen et al., 1992), and so it is justified for us to look at only the interactions between one of these dimers and one amicyanin.

Table 3-2 - Structures of Complexed and Unbound Proteins

Complexed						Unbound					
		Protein 1		Protein 2		Protein 1			Protein 2		
PDB Code	Resolution / Å	Name - *Species*	Chain	Name - *Species*	Chain	PDB Code	Chain	Resolution / Å	PDB Code	Chain	Resolution / Å
				Enzyme - Inhibitor Complexes							
1brb	2.1	Trypsin - *Rat*	e	PTI[i] - *Cow*	i	1bra	-	2.2	1bpi	-	1.1
(Perona et al., 1993)						(Perona et al., 1993)			(Parkin et al., 1996)		
1cgi	2.3	α-chymotrypsinogen - *Cow*	e	PTI[i] - *Human*	i	1chg	-	2.5	1hpt	-	2.3
(Hecht et al., 1991)						(Freer et al., 1970)			(Hecht et al., 1992)		
2kai	2.5	Kallikrein A - *Pig*	a, b	PTI[i] - *Cow*	i	2pka	a, b	2.1	1bpi	-	1.1
(Chen and Bode, 1983)						(Bode et al., 1983)			(Parkin et al., 1996)		
2ptc	1.9	β-trypsin - *Cow*	e	PTI[i] - *Cow*	i	1bty	-	1.5	1bpi	-	1.1
(Marquart et al., 1983)						(Katz et al., 1995)			(Parkin et al., 1996)		
2sic	1.8	Subtilisin - *B*[ii]. *Amyloliquefaciens*	e	Subtilisin I[iii] - *Streptomyces*	i	1sup	-	1.6	2ssi	-	2.6
(Takeuchi et al., 1991)						(Gallagher et al., TBP)			(Satow et al., 1980)		
2sni	2.1	Subtilisin - *B*[ii]. *Amyloliquefaciens*	e	Chymotrypsin I[iii] - *Barley*	i	1sup	-	1.6	2ci2	-	2.0
(McPhalen and James, 1988)						(Gallagher et al., TBP)			(McPhalen and James, 1987)		

Table 3-2 - Structures of Complexed and Unbound Proteins (Continued)

Complexed						Unbound					
		Protein 1		Protein 2		Protein 1			Protein 2		
PDB Code	Resolution / Å	Name - *Species*	Chain	Name - *Species*	Chain	PDB Code	Chain	Resolution / Å	PDB Code	Chain	Resolution / Å
1acb	2.0	α-chymotrypsin - *Cow*	e	Eglin C - *Leech*	i	5cha	a	1.7			
(Frigerio et al., 1992)						(Blevins and Tulinsky, 1985)					
1brc	2.5	Trypsin - *Rat*	e	Amyloid β-protein I[iii] - *Human*	i	1bra	-	2.2			
(Perona et al., 1993)						(Perona et al., 1993)					
1cho	1.8	α-chymotrypsin - *Cow*	e	Ovomucoid - *Turkey*	i	5cha	a	1.7			
(Fujinaga et al., 1987)						(Blevins and Tulinsky, 1985)					
1cse	1.2	Subtilisin Carlsberg - *B*[ii]. *Subtilis*	e	Eglin C - Leech	i	1scd	-	2.3			
(Bode et al., 1987)						(Fitzpatrick et al., 1994)					
1ppe	2.0	Trypsin - *Cow*	e	Trypsin I - *Cucurbita Ficifolia*	i	1bty	-	1.5			
(Bode et al., 1989)						(Katz et al., 1995)					
1sbn	2.1	Subtilisin - *B*[ii]. *Subtilis*	e	Eglin C - *Leech*	i	1sup	-	1.6			
(Heinz et al., 1991)						(Gallagher et al., TBP)					
1stf	2.4	Papain - *Papaya*	e	Stefin B - *Human*	i	1ppn	-	1.6			
(Stubbs et al., 1990)						(Pickersgill et al., 1992)					

Table 3-2 - Structures of Complexed and Unbound Proteins (Continued)

Complexed						Unbound					
		Protein 1		Protein 2		Protein 1			Protein 2		
PDB Code	Resolution / Å	Name - *Species*	Chain	Name - *Species*	Chain	PDB Code	Chain	Resolution / Å	PDB Code	Chain	Resolution / Å
1tab	2.3	Trypsin - *Cow*	e	Bowman-Birk I[iii] - *Adzuki Beans*	i	1bty	-	1.5			
(Tsunogae et al., 1986)						(Katz et al., 1995)					
1tgs	1.8	Trypsinogen - *Cow*	z	PTI[i] - *Pig*	i	1tgt	-	1.5			
(Bolognesi et al., 1982)						(Walter et al., 1982)					
2tec	2.0	Thermitase - *T*[iv]. *Vulgaris*	e	Eglin C - *Leech*	i	1thm	-	1.4			
(Gros et al., 1989)						(Teplyakov et al., 1990)					
4htc	2.3	α-thrombin - *Human*	l, h	Hirudin - *Leech*	i	2hnt	-	2.5			
(Rydel et al., 1991)						(Rydel et al., 1994)					
1udi	2.7	U-DNA Glycosylase - *HSV*	e	I[iii] - *Bacteriophage PBS1*	i	1udh	-	1.8			
(Savva and Pearl, 1995)						(Savva et al., 1995)					

Table 3-2 - Structures of Complexed and Unbound Proteins (Continued)

Complexed						Unbound					
		Protein 1		Protein 2		Protein 1			Protein 2		
PDB Code	Resolution / Å	Name - *Species*	Chain	Name - *Species*	Chain	PDB Code	Chain	Resolution / Å	PDB Code	Chain	Resolution / Å
Antibody - Antigen Complexes											
1mlc	2.1	Antibody Fab D44.1 - *Mouse*	a, b	Lysozyme - *Hen Egg White*	e	1mlb	-	2.1	1lza	-	1.6
(Braden et al., 1994)						(Braden et al., 1994)			(Maenaka et al., 1995)		
1vfb	1.8	Antibody Fv D1.3 - *Mouse*	a, b	Lysozyme - *Hen Egg White*	c	1vfa	a, b	1.8	1lza	-	1.6
(Bhat et al., 1994)						(Bhat et al., 1994)			(Maenaka et al., 1995)		
1nca	2.5	Antibody Fab NC41 - *Mouse*	l, h	Neuraminidase - *Flu Virus*	n				7nn9	-	2.0
(Tulip et al., 1992)									(Varghese et al., 1995)		
1nmb	2.5	Antibody Fab NC10 - *Mouse*	l, h	Neuraminidase - *Flu Virus*	n				7nn9	-	2.0
(Malby et al., 1994)									(Varghese et al., 1995)		
1igc	2.6	Antibody Fab MOPC21 - *Mouse*	l, h	Protein G - *Streptomyces*	a				1igd	-	1.1
(Derrick and Wigley, 1994)									(Derrick and Wigley, 1994)		
1jel	2.8	Antibody Fab JE142 - *Mouse*	l, h	His Containing Protein - *E. Coli*	p				1poh	-	2.0
(Prasad et al., 1993)									(Jia et al., 1993)		

Table 3-2 - Structures of Complexed and Unbound Proteins (Continued)

Complexed						Unbound					
		Protein 1		Protein 2		Protein 1			Protein 2		
PDB Code	Resolution / Å	Name - *Species*	Chain	Name - *Species*	Chain	PDB Code	Chain	Resolution / Å	PDB Code	Chain	Resolution / Å
3hfl	2.7	Antibody Fab HyHel5 - *Mouse*	l, h	Lysozyme - *Hen Egg White*	y				1lza	-	1.6
(Cohen et al., 1996)									(Maenaka et al., 1995)		
				Complexes of Other Types							
1atn	2.8	Deoxyribonuclease I - *Cow*	d	Actin - *Rabbit*	a	3dni	-	2.0			
(Kabsch et al., 1990)						(Oefner and Suck, 1986)					
1gla	2.6	Glycerol Kinase - *E. Coli*	g	GSF^{v} III - *E. Coli*	f				1f3g	-	2.1
(Hurley et al., 1993)									(Worthylake et al., 1991)		
1spb	2.0	Subtilisin - *E. Coli*	s	Subtilisin Prosegment - *E. Coli*	p	1sup	-	1.6			
(Gallagher et al., 1995)						(Gallagher et al., TBP)					
2btf	2.6	Profilin - *Cow*	p	β-actin - *Cow*	a	1pne	-	2.0			
(Schutt et al., 1993)						(Cedergren-Zeppezauer et al., 1994)					
3hhr	2.8	Growth Hormone - *Human*	a	Receptor - *Human*	b, c	1hgu					
(de Vos et al., 1992)						(Chantalat et al., 1995)					

Table 3-2 - Structures of Complexed and Unbound Proteins (Continued)

Complexed						Unbound					
		Protein 1		Protein 2		Protein 1			Protein 2		
PDB Code	Resolution / Å	Name - *Species*	Chain	Name - *Species*	Chain	PDB Code	Chain	Resolution / Å	PDB Code	Chain	Resolution / Å
1mda	2.5	Methylamine Dehydrogenase - *Paracoccus Denitrificans*	l, h	Amicyanin - *Paracoccus Denitrificans*	a				1aan	-	2.0
(Chen et al., 1992)									(Durley et al., 1993)		

i. PTI - Pancreatic Trypsin Inhibitor
ii. B - Bacillus
iii. I - Inhibitor
iv. T - Thermitase
v. GSF - Glucose Specific Factor

3.2.5 Identical Proteins in Different Complexes

We wished to investigate whether different complexed structures of a protein, taken from complexes formed with different proteins, are more similar to each other than to the structure of the protein when unbound. If this were true it could have important implications for docking experiments - starting with a complexed structure could improve the prediction of the structure of a complex with another protein.

The set of bound and unbound proteins (table 3-2) was searched for cases where the same protein was present in different complexes, as well as in an unbound form. SCOP classifications were used to identify identical proteins (see section 3.2.2). Five different proteins were found to have this data available (table 3-3). The lysozyme and neuraminidase complexes were ignored because their partners in the complexes are antibodies. These do not necessarily bind in the same place, and so one would not expect changes in the interface to be common to all the complexes. Three of the five proteins are from the same family (eukaryotic proteases), and two of these are trypsins. All three of these proteins were examined because to ignore them would reduce the size of the data set to an unacceptably small size. However, no attempt was made to distinguish between the movements seen with each of the five proteins based on protein type. Also any conclusions drawn from the five as a whole must be used cautiously, as they will be biased towards the eukaryotic protease family.

Table 3-3 - Structures of Proteins in Several Different Complexes

	Unbound Form		Complexed Forms							
Protein - *Species*			1		2		3		4	
	PDB	Chain	PDB	Chain	PDB	Chain	PDB	Chain	PDB	Chain
Subtilisin - *Bacillus Amyloliquifaciens*	1sup	-	1sbn	e	1spb	s	2sic	e	2sni	e
Trypsin - *Cow*	1bty	-	1tab	e	2ptc	e	1ppe	e		
Trypsin - *Rat*	1bra	-	1brc	e	1brb	e				
Chymotrypsin - *Cow*	5cha	a	1acb	e	1cho	e				
PTI - *Cow*	1bpi	-	2kai	i	2ptc	i	1brb	i		

3.3 Methods

3.3.1 Definitions of Different Regions of Protein Structures

Exposed Residues

In common with Flores et al., 1993, residues were defined as exposed if their total relative side-chain surface area (or total relative main-chain surface area in the case of glycine) was greater than 15%. All others were defined as buried. Surface area was calculated by the algorithm of Lee and Richards, 1971, implemented in the program ‘pdbarea’ (Suhail Islam, personal communication), with a probe radius of 1.4Å. ‘Relative areas’ are relative to that of the particular residue in its extended conformation (Miller et al., 1987). see “Truncation Based on Side-chain Exposure”, Chapter Two, page 60 for a more detailed explanation.

Interface Residues

Interface residues for each component of every complex (table 3-2) were defined as those where an atom centre was 4Å or nearer to the centre of any atom in the other component. This definition was chosen because the maximum separation between the centres of two side-chain substituents whose van der Waals surfaces are just touching is 4Å. This value equals twice the van der Waals radii of a side-chain methyl group, which has the maximum van der Waals radius of any side-chain substituent (Gellatly and Finney, 1982).

Jones and Thornton, 1996, define interface residues as those whose accessible surface area (figure 2-4) decreases by more than 1Å^2 from that when the structure of the component of interest is separated from the structure of the complex to that when the component is not separated from the complex. This definition tends to include slightly more residues at the edges of the interfaces. However, the differences are small and obviously dependent on the cut-offs and van der Waals radii used in both cases.

3.3.2 Structural Superposition

Pairs of proteins were superposed by the least squares fitting algorithm of McLachlan, 1979 (see “Structural Superposition”, Chapter Two, page 55). The pairs of independently solved structures of identical proteins (table 3-1), and the instances of identical proteins in different complexes (table 3-3) were superposed on all equivalent Cα

atoms by the implementation of this algorithm in the Structural Alignment of Multiple Proteins (STAMP) program of Russell and Barton, 1992. The pairs of complexed and unbound proteins were superposed on the Cα atoms of their non-interface residues (see “Interface Residues” on page 94) using the program ‘lsqfit’ (Suhail Islam, personal communication).

3.3.3 Calculations of Conformational Change

Regions in Which the Calculations Were Applied

Conformational differences between pairs of superposed structures (see section 3.3.2), were calculated as described below. For the pairs of independently solved structures of identical proteins (table 3-1), the calculations were performed separately for all residues and for exposed residues (see “Exposed Residues” on page 94). For the pairs of structures of complexed and unbound proteins (table 3-2), the calculations were performed separately for all residues, exposed residues, and interface residues (see “Interface Residues” on page 94). For the structures of identical proteins in different complexes (table 3-3), the calculations were performed only on those residues common to the interfaces of all the complexes. The pairs of complexed and unbound structures did not in general show movements away from the interface that could be attributed to association (see section 5.4), and therefore the only differences in the structures of identical proteins in different complexes will be in the interfaces.

Root Mean Square Deviation

The first measures of conformational change calculated were Root Mean Square Deviations (RMSD’s, described in Chapter Two) of all Cα and side-chain atoms in each region described above. In addition, for each region the Cα and side-chain RMSD’s of individual residues were calculated.

Torsion Angle Change

As well as measuring side-chain RMSD’s as above, changes in side-chain conformations were analysed by measuring the changes in the class of their χ_1 and χ_2 torsion angles. Figure 3-1 shows the definition of these torsion angles and of their different classes. The χ_1 torsion angle is that around the Cα-Cβ bond, and χ_2 is that around the Cβ-Cγ bond.

Both torsion angles have three classes, corresponding to positions of energy minima. These energy minima arise from steric hindrance of overlapping atoms at other positions (Janin et al., 1978), and their idealised positions are given in figure 3-1b and figure 3-1c. Torsion angles were considered to be of a particular class if they were 60° or less from the position of minimum energy of that class (Flores et al., 1993). This implicitly allows for the fact that different residue types have differing patterns of steric hindrance, and therefore have their energy minima in different positions. This is especially important for χ_2 minima when Cγ is trigonal, not tetrahedral as shown in figure 3-1c. Counting changes of class rather than absolute changes of torsion angles ensures that only changes which cross an energy maximum, and are therefore considered to be more important, are counted. χ_2 angles were only examined for change if the related χ_1 angle did not change minima, as changes in χ_1 can be coupled with changes in χ_2 because of alterations in the pattern of steric hindrance (Janin et al., 1978).

Symmetrical and Ambiguously Defined Residues

Figure 3-2 shows that certain residue types (aspartic acid, glutamic acid, phenylalanine, and tyrosine) have portions of their side-chains that are symmetrical, and others (asparagine, glutamine, and histidine) can be considered to have symmetry due to difficulties in distinguishing some atom types in the electron density (Abola et al., 1996). For example, a rotation of 180° of the benzene ring of phenylalanine (around the Cβ-Cγ bond) gives two identical conformations. Differences of this type between all pairs of structures were corrected for by changing the atom labels in one PDB file to match those in the other, to ensure that they did not show up as conformational changes. They were not corrected by simply adding or subtracting 180° to the torsion angle because this would not correct the associated RMSD. Leucine is a special case: it has no such symmetry but has two different conformations, corresponding to a rotation of 180° about χ_2, that are difficult to distinguish in electron density maps (Janin et al., 1978, and see figure 3-3). Re-labelling or rotating by 180° would not make the structures identical, and therefore χ_2 torsion angles of leucines were ignored.

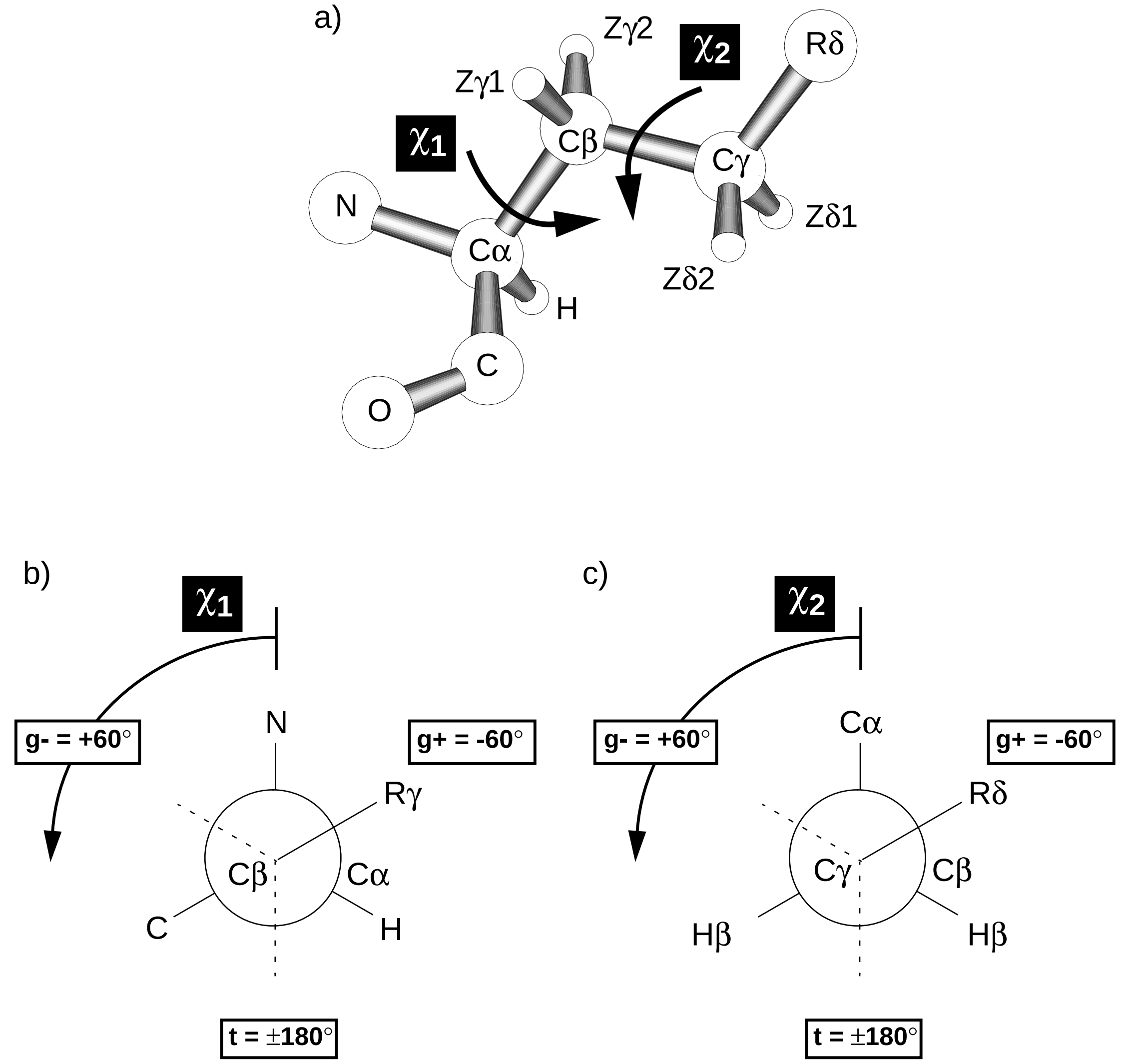


Figure 3-1 - Definition of χ_1 and χ_2 side-chain torsion angles

a) A three-dimensional representation of a side-chain of unspecified type, indicating the bonds that specify the χ_1 and χ_2 torsion angles (N-Cα-Cβ-Cγ and Cα-Cβ-Cγ-Rδ, respectively).

b) A Newman projection down the Cβ-Cα bond, showing the geometry of the χ_1 torsion angle.

c) A Newman projection down the Cγ-Cβ bond, showing the geometry of the χ_2 torsion angle.

'R' indicates the branch with the highest molecular weight at the relevant branch point. 'Z' indicates any other substituent. Idealised positions of the energy minima are shown, together with their class: gauche minus (g-), gauche plus (g+) or trans (t).

Sections b) and c) adapted from Janin et al., 1978.

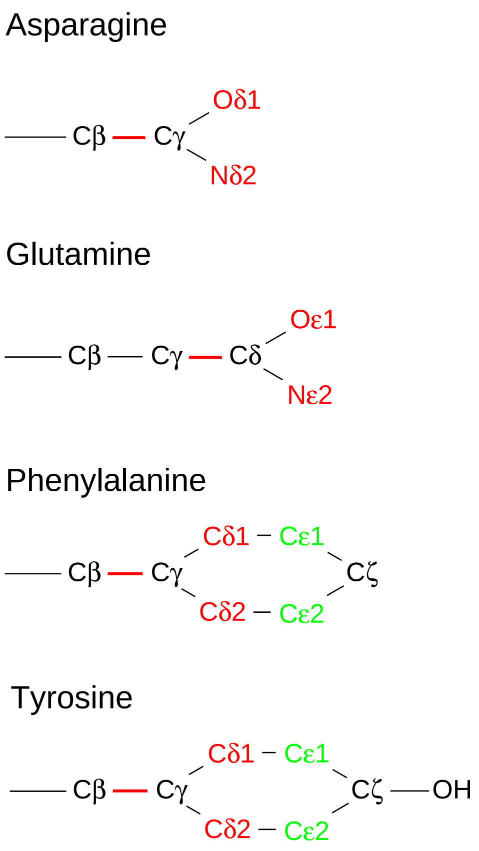


Figure 3-2 - Amino Acid Side-chains that have Symmetry

The symmetry is caused by structurally equivalent positions occupied by atoms of identical types, or of types that are difficult to distinguish in electron density maps (Abola et al., 1996). These are indicated by atom names of the same colour (red or green) as each other. The red bonds are those that the symmetry occurs around. Figure adapted from an earlier and now unavailable version of Abola et al., 1996.

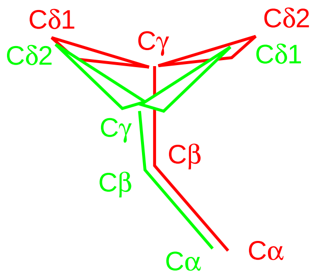


Figure 3-3 - Structural Ambiguity in Leucine Side-chains

A schematic diagram of the two conformations of Leucine side-chains (one in red and one in green) that differ by a rotation of 180° about the Cβ-Cγ bond and are difficult to distinguish in electron density maps (Janin et al., 1978).

3.4 Discussion and Conclusions

This chapter has presented a data set of thirty-nine pairs of complexed and unbound structures of proteins, from which an analysis of the conformational changes that occur on protein-protein association can be made. The structures were selected using criteria that ensured that, as far as possible, the structures were of good quality. A data set of twelve pairs of independently solved structures of identical proteins has also been given. These can be compared to find the extent of conformational differences that are caused by experimental differences in structure, which will be used to assess the importance of differences seen in the complexed-unbound data set. Methods for calculating the conformational change have been detailed, with attention to ensuring that ambiguities caused by the format in which the structures are specified are not carried through into the final measurements.

Chapter Four

Differences of Independently Solved Structures of Identical Proteins

Comparisons of independently solved structures of identical proteins give an indication of the differences in structure that can be expected from differences in their experimental determination. Twelve such pairs of crystal structures (table 3-1) were found in the Brookhaven Protein Data Bank (PDB).

Chapter Five examines conformational changes on protein-protein association, and any conformational changes that have magnitudes that are equal to or smaller than the differences found here cannot be distinguished from differences in the experimental determination of structures. The word ‘control’ is used to refer to the appropriate value.

4.1 Overall Measures

Several measures were used to analyse the overall conformational differences between the members of each pair (see section 3.3.3): Cα root mean square deviation (RMSD), side-chain RMSD, and the percentage of χ_1 and χ_2 angles that occupy different minima. These were calculated separately for both exposed residues and all residues (table 4-1). The data often have non-normal distributions (see figure 4-1), which make means and standard deviations inappropriate measures for comparisons with the other data sets examined in this thesis. Therefore a cut-off was chosen for each measure such that 95% of all the control pairs have values below it.

The differences between the means, maximums and 95% cut-offs are illustrated in figure 4-1. This shows histograms of the number of pairs of structures that have a particular value of conformational change, using three different measures as examples. The all residue Cα RMSD's have a roughly normal distribution, and in this case the 95% cut-off is equal to the mean plus one standard deviation (figure 4-1a). With non-normal distributions, such as shown by exposed residue side-chain RMSD (figure 4-1b) and the percentage of χ_1 angles of exposed residues that change minima (figure 4-1c), the mean plus one standard deviation excludes several pairs of structures. In these cases the 95% cut-offs give a better representation of the amount of conformational change.

The cut-offs are given in the last row of table 4-1, and summarised below. The values for all residues are useful for comparisons with studies by other groups, such as Flores et al., 1993, because they also use this measure. The values for exposed residues are particularly relevant to the work presented in this thesis because the differences in the conformation of interface residues between bound and unbound structures are compared with them. This is because the interface residues are exposed when the components of the complexes are unbound.

The 95% cut-off for RMS deviation of Cα atoms is 0.6Å over exposed residues and 0.4Å over all residues. The Cα RMS deviation over all residues from a similar analysis (Flores et al., 1993) is higher at 1.0Å. This reflects both the differences in the two data sets - this work was done on a newer data set, with a stricter definition of identical proteins (see

Figure 4-1 - Comparisons of Means, Maximums and 95% Cut-offs for Overall Measures of Conformational Change

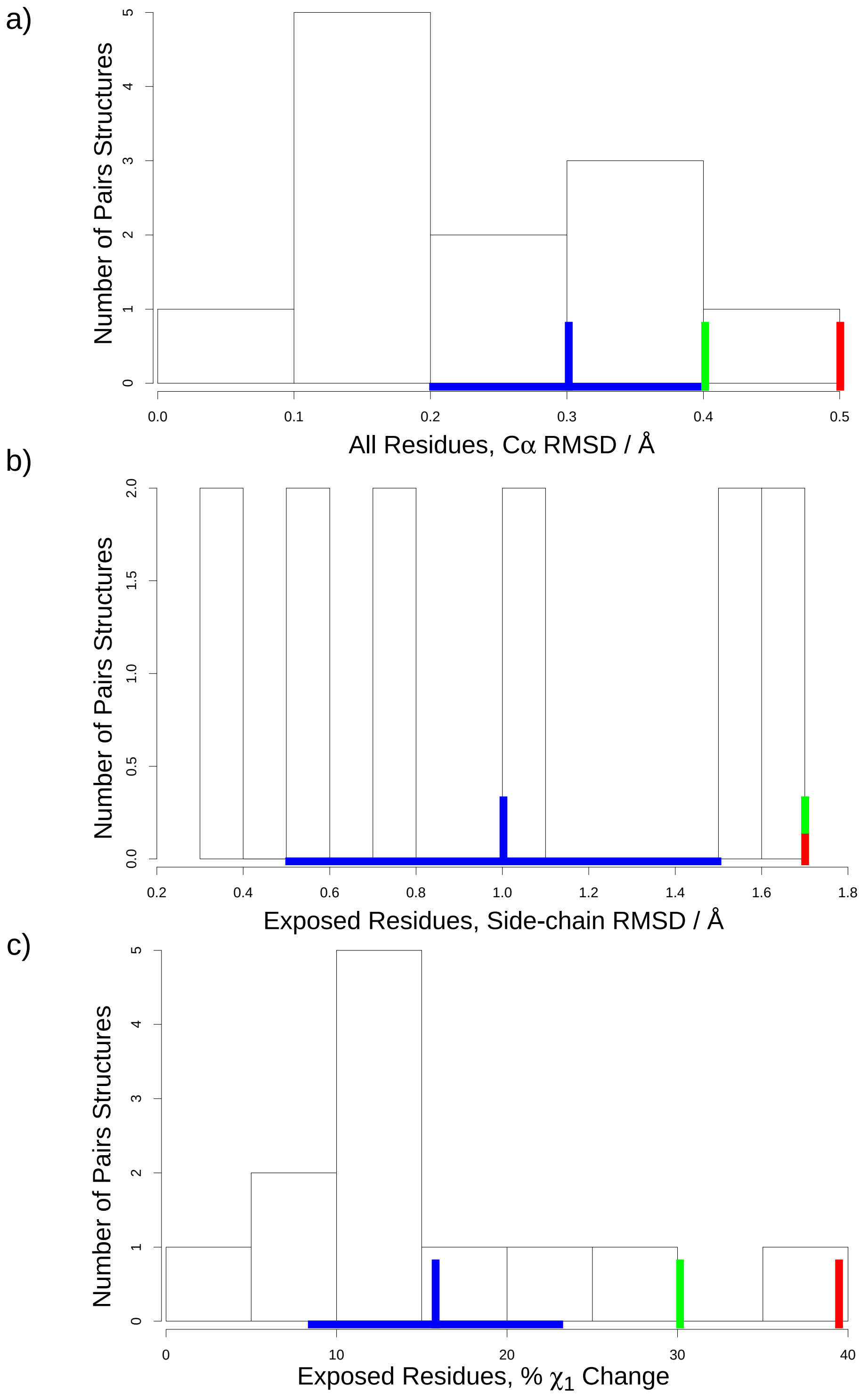
Shown by histograms of the number of pairs of control structures (see table 3-1) that have a particular value of conformational change, with the following three measures as examples:

a) Cα RMSD of all residues.

b) Side-chain RMSD of exposed residues.

c) Percentage of χ_1 angles of exposed residues that change minima.

Red lines - maximum values seen; green line - the 95% cut-off; vertical blue line - mean; horizontal blue line - one standard deviation either side of the mean. See table 4-1.



section 3.2.3) - and the fact that Flores et al., 1993, did not ignore poorly defined residues as was done here (see section 3.2.1). The conformation of these residues is expected to differ more than that of others because of uncertainty in their position, or high mobility. The 95% cut-off for side-chain RMSD is 1.7Å over exposed residues and 1.6Å over all residues. Side-chain RMSD's were not given by Flores et al., 1993.

Changes in side-chain torsion angles were also calculated for exposed residues and for all residues. For structure comparison, a particularly useful measure of torsion angle change is the percentage of side-chain torsion angles that occupy different minima (see "Torsion Angle Change", Chapter Three, page 95). χ_2 angles are only examined for change when their related χ_1 angles does not change. The 95% cut-offs are 31% of χ_1 angles and 23% of χ_2 angles for exposed residues, and 24% and 21% for all residues. Flores et al., 1993 also calculated percentages over all the pairs of structures that they compared, and found that 81.7% of χ_1 angles and 86.7% of χ_2 angles (where χ_1 did not change) occupy the same minima in each structure. Our values are 87.1% for χ_1 and 90.1% for χ_2, suggesting that torsion angle are more conserved in this data set than in that of Flores et al., 1993. For exposed residues, our values are 83.1% for χ_1 and 87.9% for χ_2. Flores et al., 1993 did not calculate values separately for exposed residues.

The two structures of transforming growth factor β (TGF–β) in the data set (table 3-1) have already been compared in detail by Daopin and Davies, 1994, and our results confirm theirs. They also present four different methods for estimating the coordinate errors. Three of these methods require knowledge of the diffraction data, which is not generally available in the public domain. Hence they are not discussed further, except to say that they cannot give a value for systematic differences in the determination of structures; these can be found only by comparing independently solved structures, as presented in this thesis. The fourth method was based on such a comparison, but of only one pair of structures. Another method of estimating coordinate error was given by Tickle et al., 1998, who calculated standard uncertainties for two crystallin structures from full-matrix least-squares refinement. This also requires generally unavailable data and can not quantify systematic differences.

Table 4-1 - Overall Differences in Control Pairs

Structure[i]	Number of Residues[ii]		RMSD of Cα atoms / Å		RMSD of side-chain atoms / Å		% χ_1 change		% χ_2 change[iii]	
1	All	Exposed	All	Exposed	All	Exposed	All	Exposed	All	Exposed
135l	47	16	0.4	0.6	0.9	1.1	14	10	6	16
1bfg	123	82	0.2	0.2	0.7	0.8	10	14	7	9
1bpb	130	69	0.2	0.2	0.8	0.8	14	14	8	9
1hhp	83	57	0.5	0.6	1.6	1.6	35	39	29	23
1lza	129	83	0.4	0.5	1.4	1.7	10	13	15	18
1rcb	90	52	0.3	0.3	1.6	1.7	17	19	15	20
1rhb	104	63	0.2	0.3	0.5	0.6	7	12	4	3
2i1b	79	29	0.2	0.2	0.6	0.6	6	11	0	0
2tgi	77	55	0.1	0.1	0.5	0.4	1	2	0	0
3cd4	109	63	0.4	0.4	1.4	1.6	22	30	21	25
3psg	282	120	0.2	0.2	0.4	0.4	5	6	5	8
4cms	281	141	0.3	0.3	0.9	1.1	14	23	9	14
Mean			0.3	0.3	0.9	1.0	12	16	10	12
Standard Deviation			0.1	0.2	0.4	0.5	9	10	9	9
Maximum			0.5	0.6	1.6	1.7	35	39	29	25
95% Cut-off[iv]			0.4	0.6	1.6	1.7	22	30	21	23

[i] Identified by PDB code of the first structure in the pair (table 3-1).

[ii] This only includes well defined residues (see section 3.2.1) common to both members of the pair.

[iii] Changes in χ_2 minima calculated only when the corresponding χ_1 does not change minima.

[iv] 95% of all the pairs have values less than or equal to this cut-off. In practice this means that one outlier (from twelve pairs) is ignored.

4.1.1 The Effects of Resolution

Resolution is a measure of the global precision of a structure (see section 1.1). A resolution cut-off has already been used in selecting the structures to be examined (section 3.2.1), but it is desirable to see how resolutions better than this cut-off affect the measurements made.

Figure 4-2 shows Cα RMSD's, side-chain RMSD's, and percentages of torsion angles that change minima for all residues and for exposed residues of each pair of structures (table 4-1), plotted against the resolutions of both members of each pair (table 3-1). The general trend seen in each of the plots is for structural differences to decrease with better resolution. The pair of aspartic proteinase structures (identified on figure 4-2 by PDB code 1hhp) have the worst resolutions of any of the structures in table 3-1, at 2.7Å for both. Figure 4-2 shows that they also have the largest conformational differences by any of the measures mentioned, except for the side-chain RMSD of exposed residues (figure 4-2b). In this case the pair of hen egg white lysozyme structures (identified on figure 4-2 by PDB code 1lza) have the largest value. However, at 1.6Å and 1.7Å respectively, they are two of the better resolved structures. They were also solved in the same space group and with the same refinement program as each other, which reduces any differences in their structures caused by systematic differences in the way that they were solved (see section 1.1)

The pair of turkey egg white lysozyme structures (identified on figure 4-2 by PDB code 135l) have resolutions that are quite different from one another. 135l has a resolution of 1.3Å - the best of any of the structures in table 3-1. In contrast, 2lz2 has a resolution of 2.2Å, which is slightly higher than the mean resolution of the structures (2.0Å). The measurements of conformational differences between the two structures sometimes give large values - Cα RMSD's for all residues and for exposed residues are both as high as the 95% cut-offs calculated from all twelve pairs (see table 4-1) - and for all the measures except the percentages of χ_1 angles of exposed residues and of χ_2 angles of all residues that change minima, the value calculated is above the mean (also shown in table 4-1).

The pair of DNA polymerase β structures (identified on figure 4-2 by PDB code 1bpb) are both resolved to 2.3Å, which is towards the poor end of the range seen in table 3-1. However, they have relatively low levels of conformational difference by all the measures of conformational change (see table 4-1 and figure 4-2).

These results show that, in general, overall conformational differences between structures are proportional to their resolutions. However, it is still possible for structures with poor resolutions to have only small differences and for those with good resolutions to have larger differences. These large differences are either genuine or caused by systematic differences in the experimental procedures.

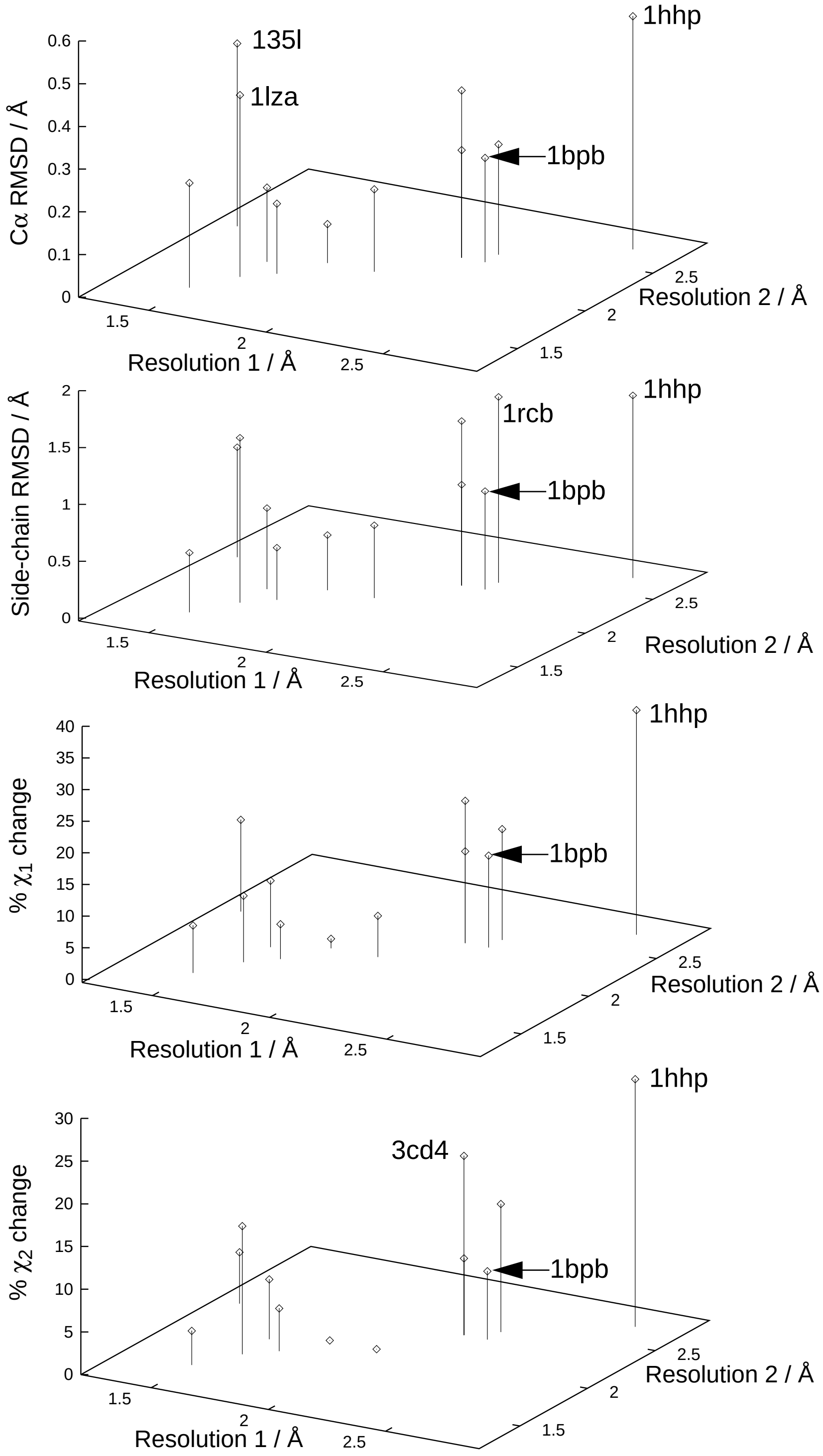
Figure 4-2 - The Relationship Between Resolution and Overall Structural Precision.

a) All residues.

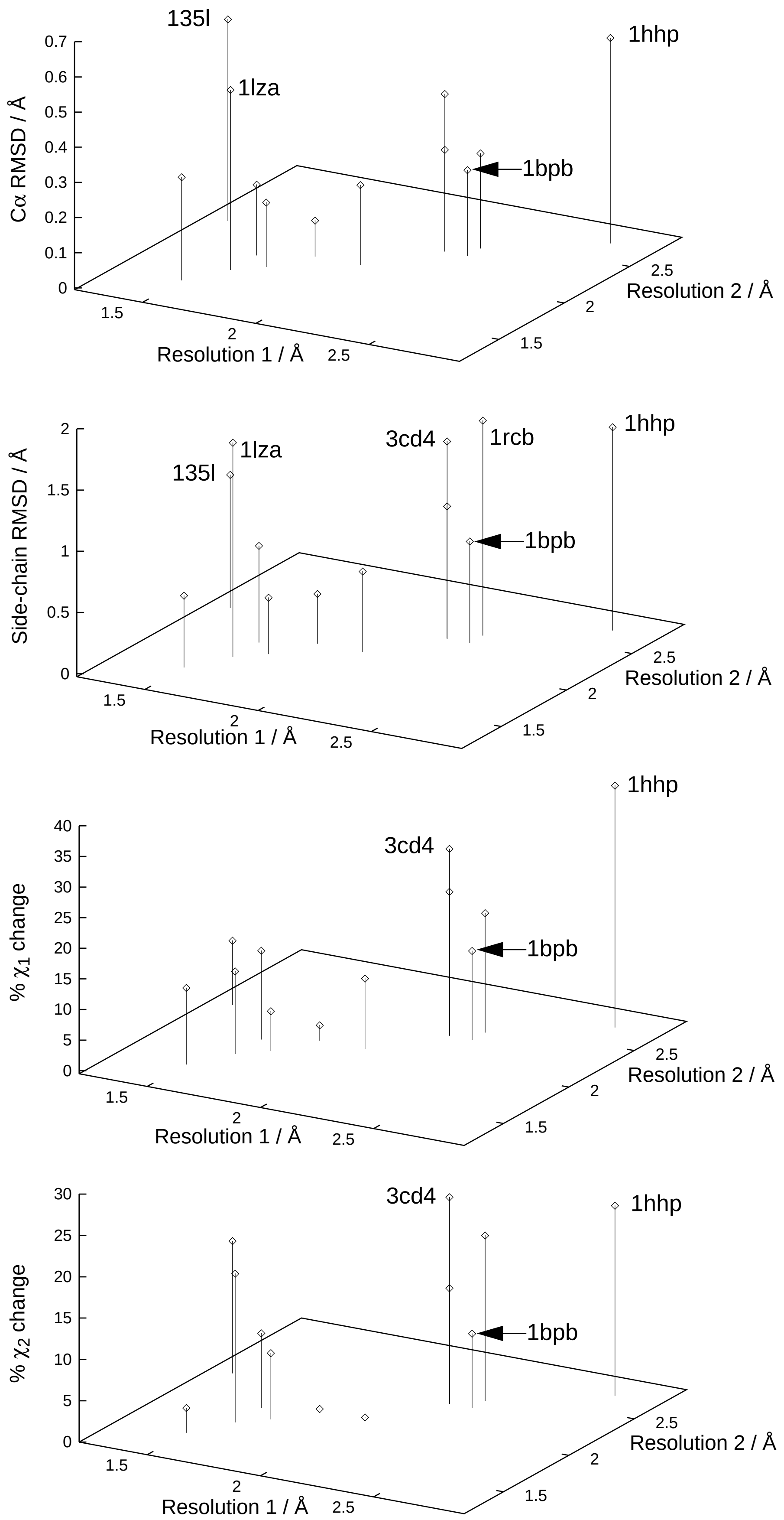
b) Exposed residues.

‘Resolution 1’ and ‘Resolution 2’ refer to the resolutions of the first and second structure in each pair respectively (see table 3-1). Specific points that are discussed in the main text (section 4.1.1) are labelled with the PDB code of the first structure in the pair (table 3-1).

a)



b)


135l
1hhp
1lza
1bpb
Cα RMSD / Å
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Resolution 1 / Å
1.5
2
2.5
Resolution 2 / Å
1.5
2
2.5
1lza
135l
3cd4
1rcb
1hhp
1bpb
Side-chain RMSD / Å
0
0.5
1
1.5
2
Resolution 1 / Å
1.5
2
2.5
Resolution 2 / Å
1.5
2
2.5
1hhp
3cd4
1bpb
% χ_1 change
0
5
10
15
20
25
30
35
40
Resolution 1 / Å
1.5
2
2.5
Resolution 2 / Å
1.5
2
2.5
3cd4
1hhp
1bpb
% χ_2 change
0
5
10
15
20
25
30
Resolution 1 / Å
1.5
2
2.5
Resolution 2 / Å
1.5
2
2.5

4.2 Movements of Individual Residues

For each of the twenty commonly occurring amino acids, the Cα displacements and side-chain RMSD’s of every exposed residue of that type were calculated. As was observed with the overall measures (section 4.1), the data have non-normal distributions. This makes means and standard deviations inappropriate measures of the amount of conformational change that can be expected in other structures. Figure 4-3 demonstrates this, using the side-chain RMSD’s of exposed arginine residues as an example. The distribution is heavy tailed, with seven of the total of forty-nine residues having side-chain RMSD’s that are significantly above (more than one standard deviation) the mean. Therefore the results are given as ‘95% cut-offs’ (table 4-2), rounded to the nearest 0.5Å. 95% of all the measurements have values less than or equal to this cut-off. Figure 4-4 shows that these 95% cut-offs include most residues, but exclude those with large outlying values. These are for N or C terminal residues, which are generally on the surface of proteins and have less constraints on their conformations than other residues, therefore making them more flexible (Thornton and Sibanda, 1983), and for those residues that are adjacent to ones poorly defined in the electron density (the poorly defined ones themselves are excluded from the calculations - see section 3.2.1).

As expected, Cα displacements are largely unaffected by residue type, reflected in equal values of 0.5Å for all types except glycine, where the value is 1.0Å. The larger value for glycine is reasonable when considering that the backbone will be more flexible because of a lack of steric hindrance caused by a side-chain. The values of side-chain RMSD are also sensible. They range from 0.5Å for small residues, such as alanine, and large residues with inflexible rings, such as phenylalanine, through 2.5Å for long and flexible residues, for example lysine and glutamine, up to a maximum of 4.5Å. Only arginine, which has a long and potentially flexible side-chain, has this high a value.

Figure 4-3 - Comparison of the Mean, Maximum and 95% Cut-off for Side-chain RMSD's of Exposed Arginine Residues

Shown by a histogram of the number of pairs of exposed residues from the control structures (see table 3-1) that have a particular value of conformational change. Red line - maximum value seen; green line - 95% cut-off; vertical blue line - mean; horizontal blue line - one standard deviation either side of the mean. See table 4-2.

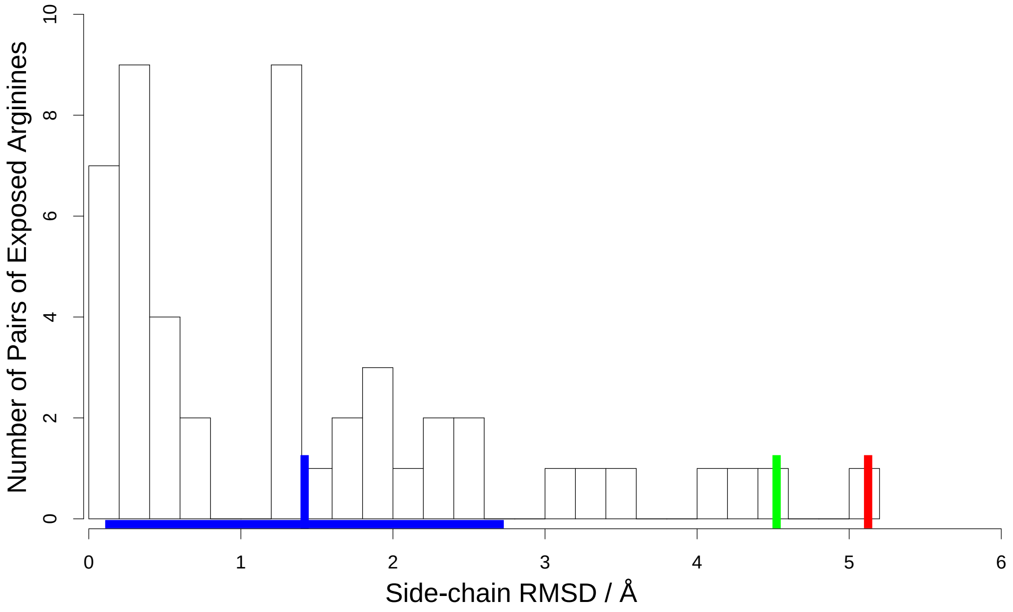
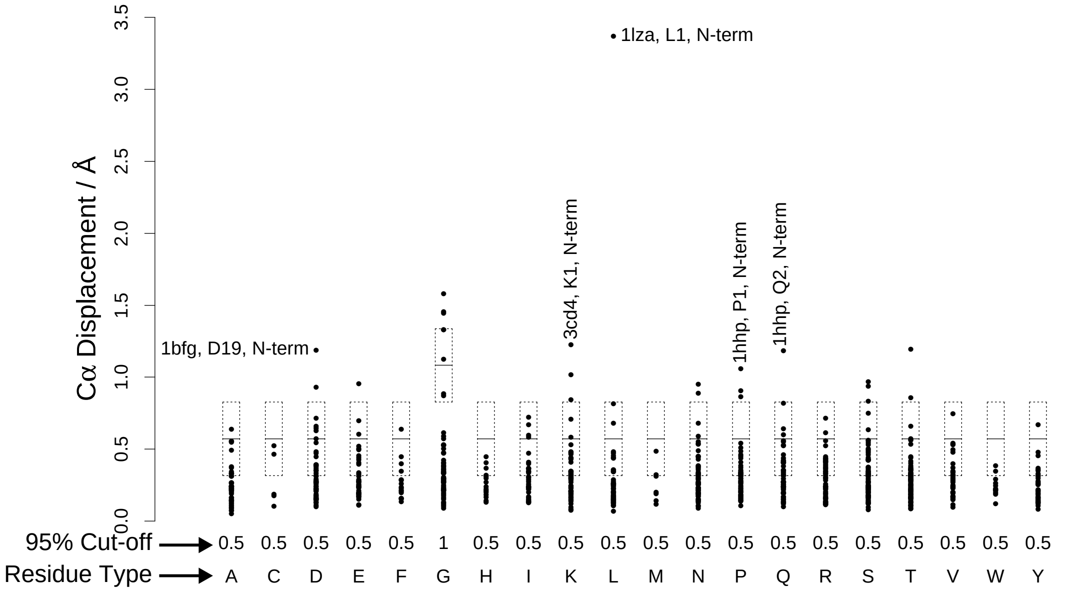


Figure 4-4 - Individual Residues with Cα Displacements and Side-chain RMSD's Above the 95% Cut-offs

Each point represents a pair of equivalent residues from the control structures (table 3-1). Each pair is of a particular amino acid type and has a specific Cα displacement or side-chain RMSD, indicated by the position of the dot along the y-axis. The short solid horizontal lines indicate the 95% cut-off, the value of which is given at the bottom of each plot (also see table 4-2). Anything below the top of the dotted box is deemed to be at or below the cut-off. These boxes are necessary because the cut-offs are given to the nearest 0.5Å, whereas the values for each residue are given to the nearest 0.1Å. Specific residues are identified when their measurements are above the 95% cut-off and the residue is at the N or C terminus (labelled 'N-term' and 'C-term'), or is adjacent to residues poorly defined in the electron density (labelled 'e$^-$')



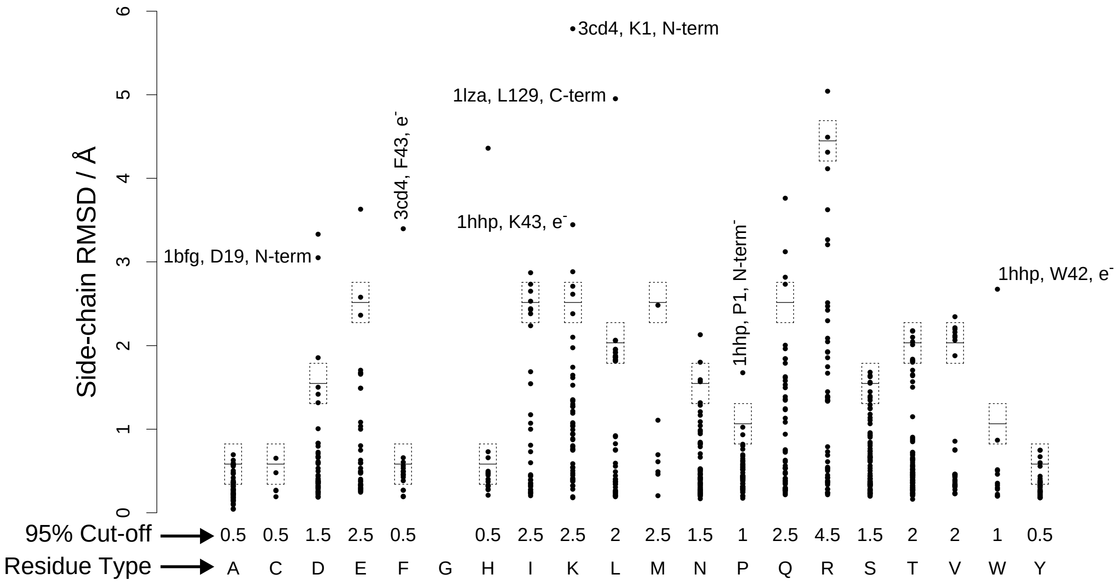


Table 4-2 - Conformational Differences Between Exposed Residues of the Control Pairs, by Residue Type.

Residue Type	Cα Displacement / Å				Side-chain RMSD / Å			
	μ[i]	σ[ii]	Max	95% Cut-off[iii]	μ[i]	σ[ii]	Max	95% Cut-off[iii]
Alanine	0.2	0.1	0.6	0.5	0.3	0.2	0.7	0.5
Cysteine	0.2	0.2	0.5	0.5	0.3	0.2	0.6	0.5
Aspartic Acid	0.3	0.2	1.1	0.5	0.5	0.6	3.3	1.5
Glutamic Acid	0.2	0.7	0.9	0.5	0.8	0.8	3.7	2.5
Phenylalanine	0.2	0.1	0.6	0.5	0.5	0.8	3.4	0.5
Glycine	0.4	0.3	1.5	1.0	N/A	N/A	N/A	N/A
Histidine	0.2	0.1	0.4	0.5	0.6	1.1	4.4	0.5
Isoleucine	0.2	0.2	0.6	0.5	1.0	1.0	2.9	2.5
Lysine	0.3	0.2	1.1	0.5	1.1	1.0	5.9	2.5
Leucine	0.3	0.5	3.2	0.5	1.0	1.0	5.0	2.0
Methionine	0.2	0.1	0.4	0.5	0.8	0.8	2.5	2.5
Asparagine	0.2	0.2	0.9	0.5	0.5	0.4	2.1	1.5
Proline	0.3	0.2	1.0	0.5[iv]	0.4	0.3	1.6	1.0
Glutamine	0.3	0.2	1.1	0.5	0.9	0.9	3.8	2.5
Arginine	0.2	0.1	0.6	0.5	1.4	1.3	5.1	4.5
Serine	0.3	0.2	0.9	0.5	0.6	0.4	1.6	1.5
Threonine	0.3	0.2	1.1	0.5	0.6	0.6	2.2	2.0
Valine	0.2	0.1	0.7	0.5	0.8	0.8	2.3	2.0
Tryptophan	0.2	0.1	0.3	0.5	0.5	0.7	2.7	1.0
Tyrosine	0.2	0.2	0.6	0.5	0.3	0.2	0.7	0.5

i. Mean

ii. Standard deviation

iii. 95% of all residues of the relevant type have values less than or equal to this cut-off. Calculated to the nearest 0.5Å.

iv. 94% cut-off given for proline Cα displacement, to simplify analysis. This includes only one less residue.

4.2.1 The Effects of Temperature Factors

Temperature factors (or ‘B-values’) are measures of the precision of the coordinates of specific atoms in a structure (see section 1.1). A B-value cut-off has already been used in selecting the residues to be included in the calculations (section 3.2.1). The value of this cut-off was chosen so that those residues whose conformational changes were obviously a direct result of large B-values were excluded (see section 3.2.1). However, as with resolution (see section 4.1.1), it is desirable to observe how B-values better than this cut-off relate to the measurements made.

Daopin and Davies, 1994, showed that the displacements between equivalent Cα atoms from the two transforming growth factor β structures given in table 3-1 were correlated with their B-values. This was especially true for those atoms with the largest B-values, i.e. above approximately 50Å^2 - the cut-off used in section 3.2.1. The correlation with B-values below this limit was not as clear.

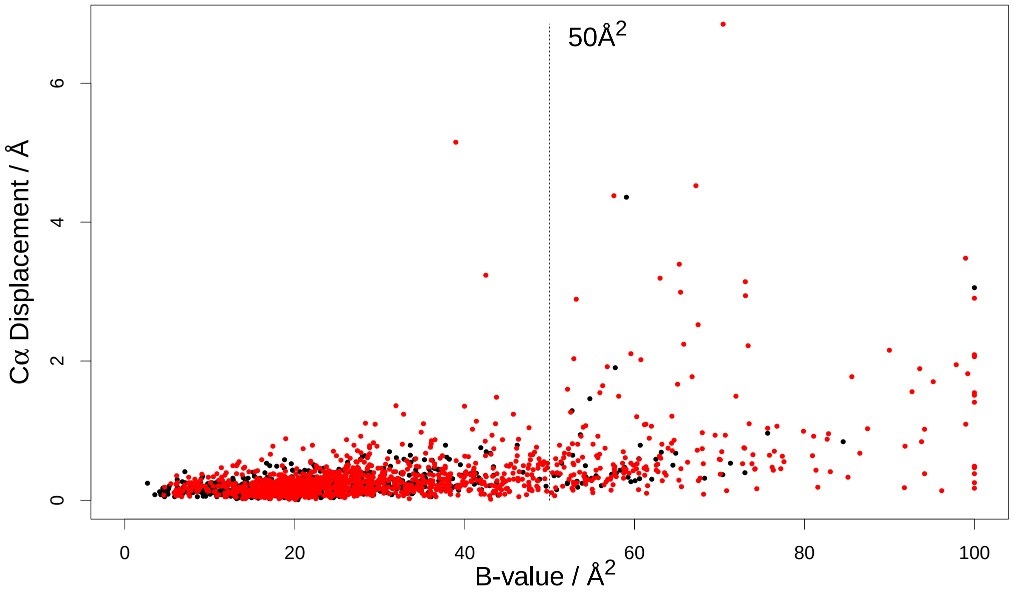
In figure 4-5, Cα displacements and side-chain RMSD’s of equivalent pairs of residues from the structures listed in table 3-1 are plotted against their B-values. The B-value used for each pair of equivalent residues was that which was the highest of any atom in the pair. It would be more intuitive to plot Cα displacement against Cα B-values. In practice, however, it was simpler to use the same B-value as used with the side-chain RMSD, and in fact the Cα displacement does show a marked correlation with the B-value used.

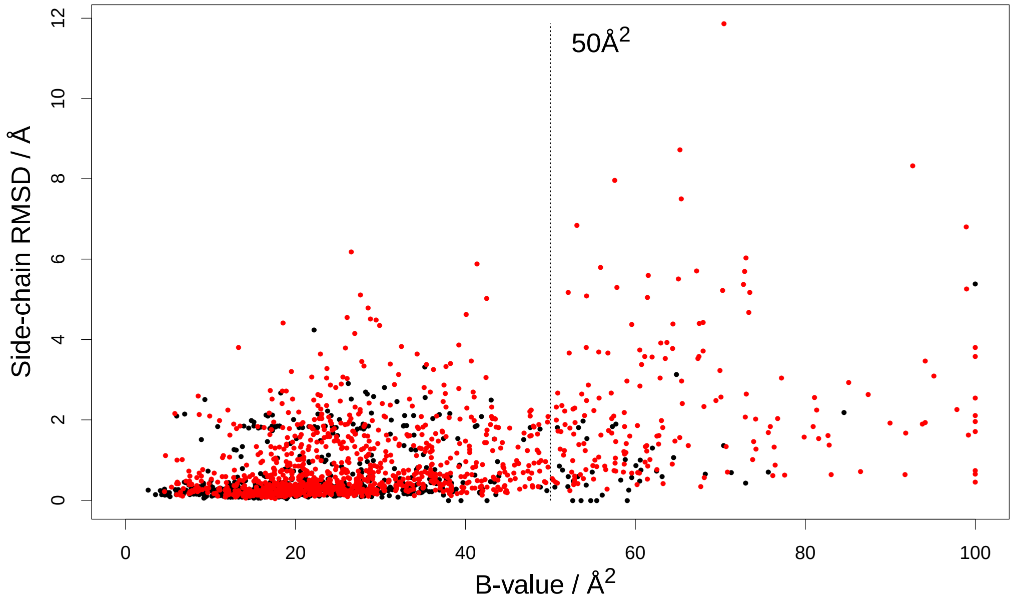
Figure 4-5 shows that the Cα displacements and side-chain RMSD’s increase with increasing B-value. The rate of the increase in Cα displacement is higher at B-values greater than 50Å^2, especially for exposed residues. This provides further justification for the cut-off used in section 3.2.1. This tendency is not as obvious with side-chain RMSD’s, although they are generally higher at B-values above 50Å^2 than they are below.

Thus B-values must be taken into account when examining local conformational differences between proteins. With larger B-values the conformational differences are more likely to be an artifact of the larger potential for movement. This follows from the definition of B-value. However, side-chain RMSD’s in particular can still be large even for residues with low B-values, suggesting that the movements are genuine or are caused by differences in experimental structure determination.

Figure 4-5 - The Relationship Between the Conformational Differences and B-values of Individual Residues

The B-values plotted for each pair of equivalent residues from the pairs of structures that were compared (table 3-1) is that which is the highest of any atom in the pair. The dotted line indicates the cut-off of 50Å^2 - any residues with B-values above this were not included in the main calculations (section 3.2.1). Black dots are for non-exposed (i.e. buried) residues. Red dots are for exposed residues. Red dots were plotted second, and so some black dots are obscured.





4.3 Discussion and Conclusions

In this chapter, the conformational differences in twelve pairs of independently solved structures of identical proteins, presented in table 3-1, have been analysed using the calculations described in Chapter Three. This analysis provides control values with which conformational changes on protein-protein association can be evaluated; only conformational changes above the controls can be said to be substantial. The most important controls for this evaluation are those calculated using exposed residues, as it is exposed residues of unbound structures that form the interfaces when the proteins associate. These controls are an overall Cα RMSD of 0.6Å, overall side-chain RMSD of 1.7Å, and percentages of χ_1 and χ_2 torsion angles that change minima equal to 30% and 23% respectively. Controls were also established for movements of individual residues, with Cα displacements being the same (0.5Å) for all amino acids types, except for glycine which was more flexible (1.0Å). This makes sense because glycine has no side-chain to restrict allowed main-chain conformational space. The side-chain RMSD controls varied by amino acid type, reflecting the differing flexibility of different substituents.

In general, the control values for overall differences were seen to be proportional to the resolutions of the structures being compared: the worse the resolution, the larger the differences. Thus when more structures become available in the future, it will be possible to refine control values and thus better evaluate the conformational difference to be expected at different resolutions. Because the data set is small (only twelve pairs of structures), and because the differences of the pairs were not normally distributed, the controls were calculated such that 95% of the pairs had a measurement at or below the control. With the measurements of individual residue differences, the non-normality of the distributions was even more pronounced, and so the controls were also calculated as 95% controls. These controls tended to exclude residues that were flexible because they were at chain ends or because residues adjacent to them were poorly defined. The individual residue differences were proportional to the temperature factors of the residues, but the temperature factor cut-off employed in selecting those residues to analyse (see section 3.2.1) removed most of the residues with large differences.

Chapter Five

Conformational Changes on Protein-Protein Association

5.1 Introduction

Comparisons of structures of proteins in complexed and unbound forms allow the amount of conformational change on protein-protein association to be quantified. Thirty-nine such pairs of crystal structures (table 3-2), from thirty-one protein-protein complexes, were found in the Brookhaven Protein Data Bank (PDB). Eighteen of the complexes were enzyme-inhibitors, seven were antibody-antigens, and the remaining six were of other types.

Chapter Three presented calculations by which structural differences can be measured, and Chapter Four applied these calculations to twelve pairs of independently solved structures of identical proteins. The values obtained gave the amount of conformational change that can be expected from differences in the experimental determination of structures. This chapter presents the results of the calculations described in Chapter Three when applied to the pairs of complexed and unbound structures. The importance of the values obtained is considered by comparison with the values expected from differences in experimental structure determination. The levels of structural difference in interface and in exposed non-interface regions are compared, as are the levels in the different types of complex. An additional analysis compares the structures of proteins that are available in several different complexes as well as in an unbound form.

5.2 Overall measures

The overall conformational changes in different regions of the protein structures were analysed for all of the pairs of complexed and unbound structures listed in table 3-2. The analysis was applied to three regions of the proteins: all residues, interface residues only and exposed non-interface residues only (see section 3.3.1). The following calculations were performed on these regions: Cα and side-chain RMSD’s over all residues in the regions, and the percentages of χ_1 and χ_2 torsion angles in the regions that change minima (see section 3.3.3). The results are given in figure 5-1, figure 5-2, and table 5-1, and described in the following sections. In each section, the word ‘control’ refers to the amount of conformational change that is expected from experimental differences in the determination of the structures (see Chapter Four, especially table 4-1).

5.2.1 All Residues

Figure 5-1a shows that just over half of all the pairs of structures (twenty of thirty-nine) have all-Cα RMSD’s that are more substantial than the control. This is also shown by red shading in the relevant column of table 5-1. However, for the three measures of conformational change of all side-chains (side-chain RMSD’s, figure 5-1b, and percentages of χ_1’s, figure 5-2a, and of χ_2’s, figure 5-2b, that change minima) more than thirty of the thirty-nine pairs have values that are less than the controls. Nineteen of the thirty-nine pairs have values for all four of these measures that are less than or equal to the controls. These pairs are indicated by yellow shading across the ‘All Residues / Overall Measures’ column of table 5-1.

5.2.2 Interface Residues

In the interface regions, substantial Cα RMSD’s occur in fewer of the pairs than they do when calculated using all residues (table 5-2), with only ten of the thirty-nine pairs having values that are greater than the controls (see figure 5-1c, and the red shading in the relevant column of table 5-1). Substantial movements of side-chains occur more often for interfaces than they do for all residues - more of the pairs of structures have values above the controls for side-chain RMSD (figure 5-1d) and percentages of χ_1’s and χ_2’s that change minima (figure 5-2c and figure 5-2d) than they do in the equivalent figures for all residues (figure 5-1b, figure 5-2a, and figure 5-2b respectively), with changes more common at χ_2 than at χ_1. These differences are seen more clearly in table 5-2. Nineteen of the thirty-nine pairs have values for all four measures that are less than or equal to the controls, shown by yellow shading of the ‘Interface Residues / Overall Measures’ column of table 5-1.

5.2.3 Exposed Non-interface Residues

The exposed non-interface regions show substantial main-chain movement (measured by Cα RMSD and presented in figure 5-1e), more often than is seen with the interface regions (see table 5-2). Table 5-2 shows that all three measures of side-chain conformational change (side-chain RMSD, figure 5-1f, and percentages of χ_1’s and of χ_2’s that change minima, figure 5-2e and figure 5-2f) have similar numbers of pairs of structures with substantial movements as each other. They also have numbers of pairs with substantial movements that are similar to those measured using all residues, but less than is seen with the interface residues (see table 5-2). Twenty-four of the thirty-nine pairs have values for all four measures that are less than or equal to the controls, shown by yellow shading of the ‘Exposed Non-interface Residues / Overall Measures’ column of table 5-1.

5.2.4 Summary

Almost half of the structures do not undergo substantial movement on association. Side-chain movement is seen more often in interface residues than in exposed non-interface residues, and the reverse is true for Cα movement. These results give a general picture of the levels of conformational change and of the differences in different regions. To understand the reasons behind them it is necessary to look at the movements of individual residues.

Figure 5-1 - RMSD’s Between Structures of Complexed and Unbound Proteins

The dotted lines show the controls - values expected from experimental differences in the determination of the structures (table 4-1). Proteins are identified by the PDB code of the complexed structure, followed by the chain identifier(s) of the relevant chain(s).

a) Cα RMSD’s of all residues.

b) Side-chain RMSD’s of all residues.

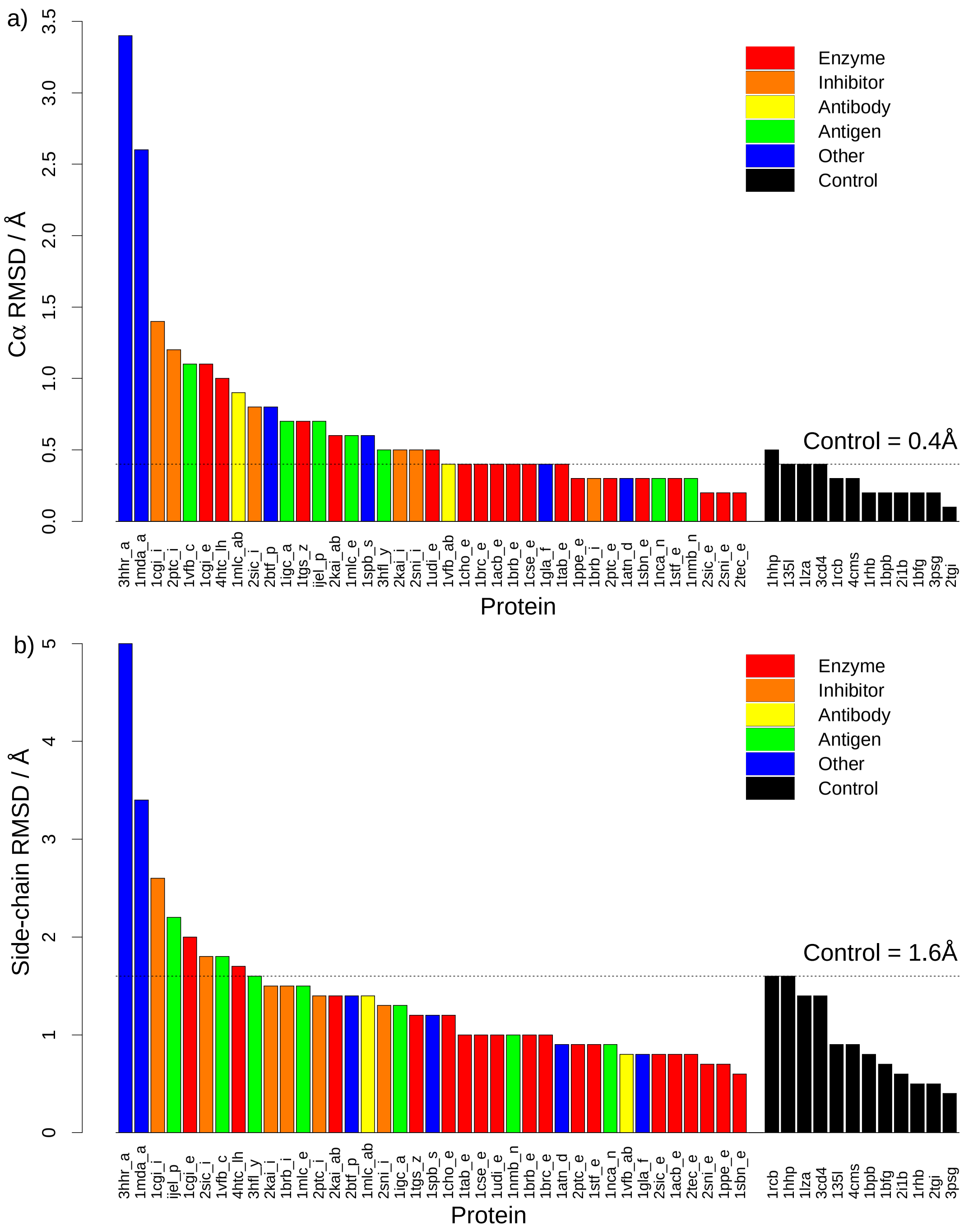
c) Cα RMSD’s of interface residues.

d) Side-chain RMSD’s of interface residues.

e) Cα RMSD’s of exposed non-interface residues.

f) Side-chain RMSD’s of exposed non-interface residues.

The numbers above the bars in c), d), e) and f) are the numbers of residues that have a Cα displacement or side-chain RMSD (as appropriate) that is above the control for that amino acid type (table 4-2). There are no such numbers on a) or b) because individual residue movements were only examined if the residue was exposed or in the interface.



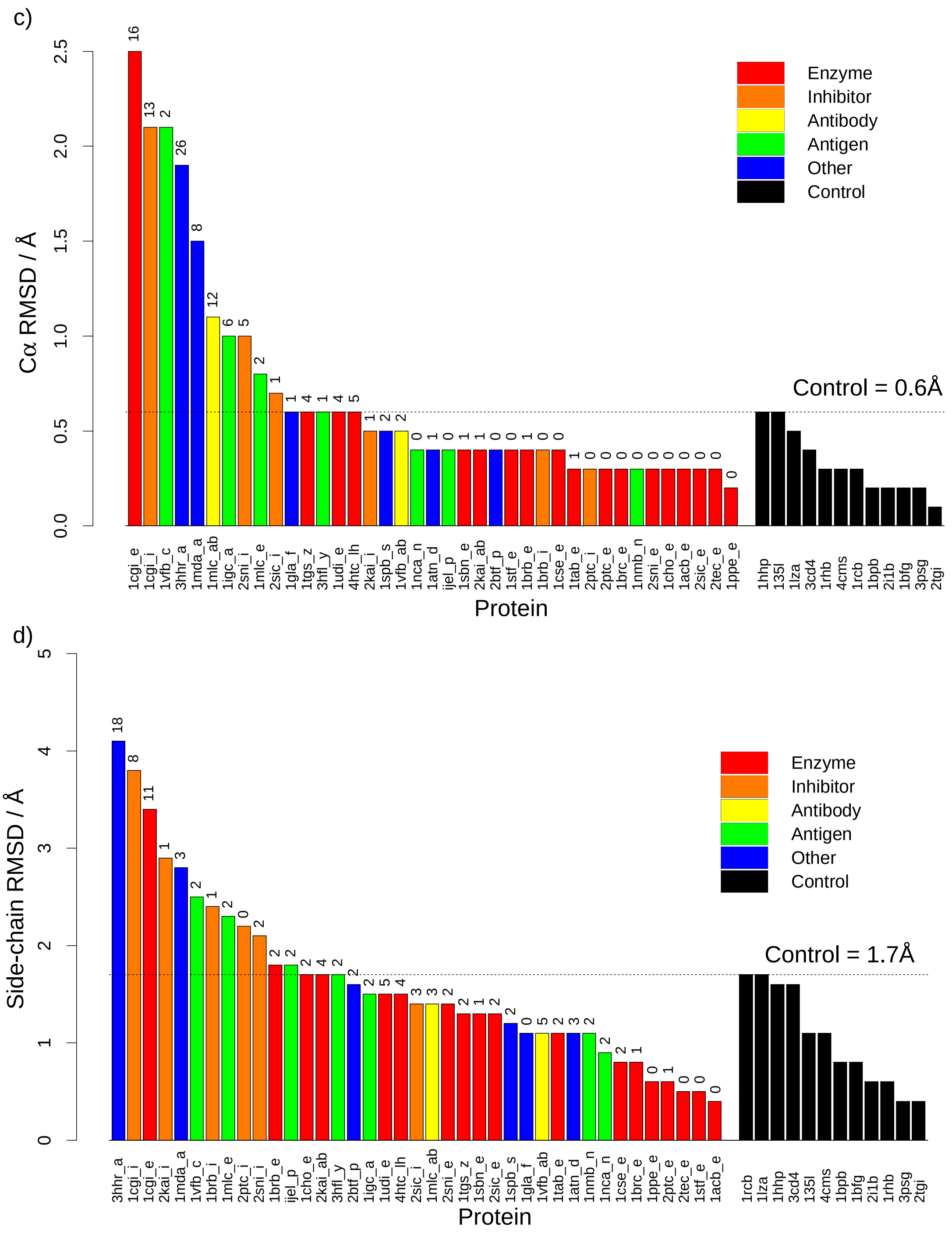


Figure 5-1 (continued)

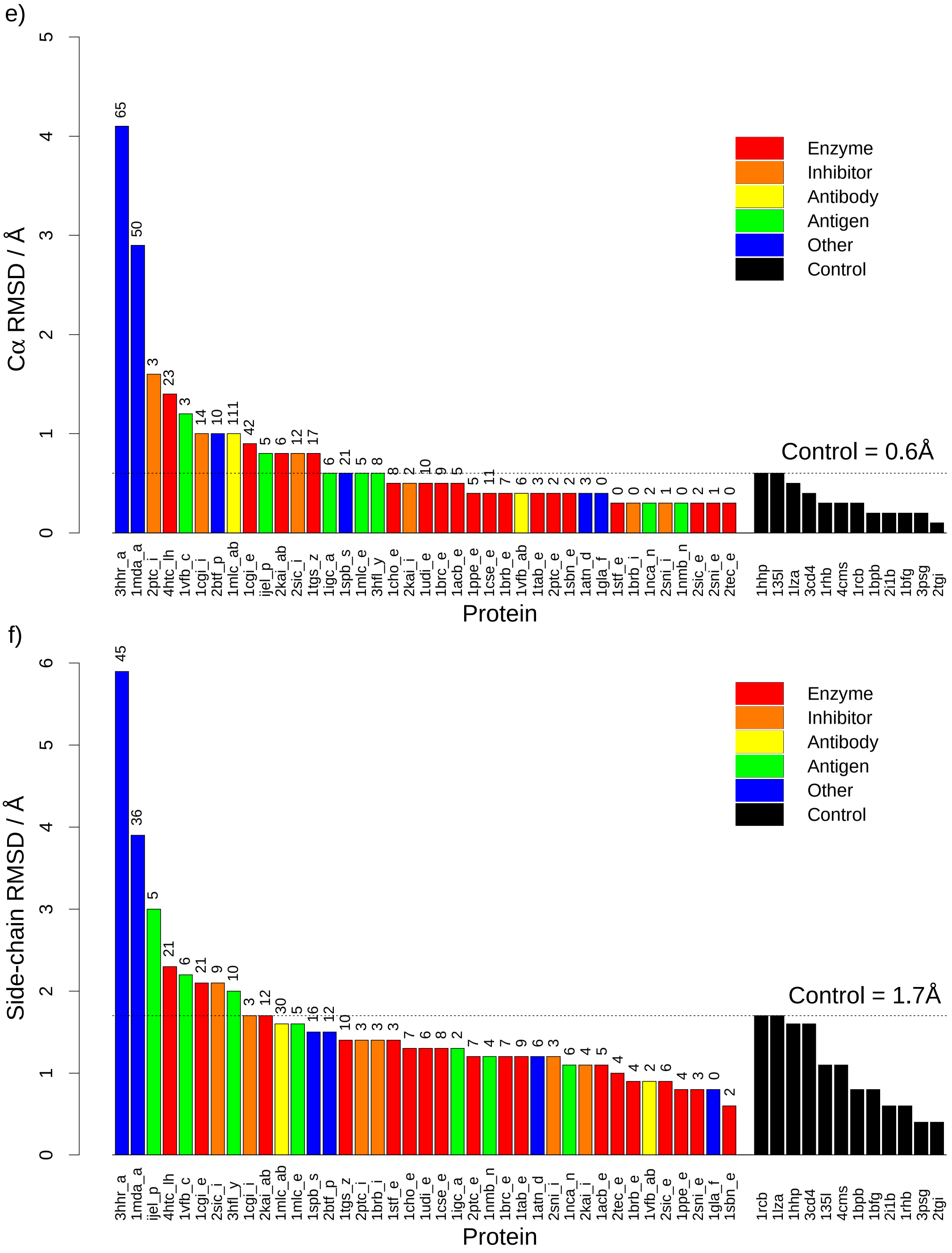


Figure 5-1 (continued)

Figure 5-2 - Torsion Angle Change Between Structures of Complexed and Unbound Proteins

The dotted line show the controls - values expected from experimental differences in the determination of the structures (table 4-1). Proteins are identified by the PDB code of the complexed structure, followed by the chain identifier(s) of the relevant chain(s).

a) Percentages of χ_1's of all residues that change minima.

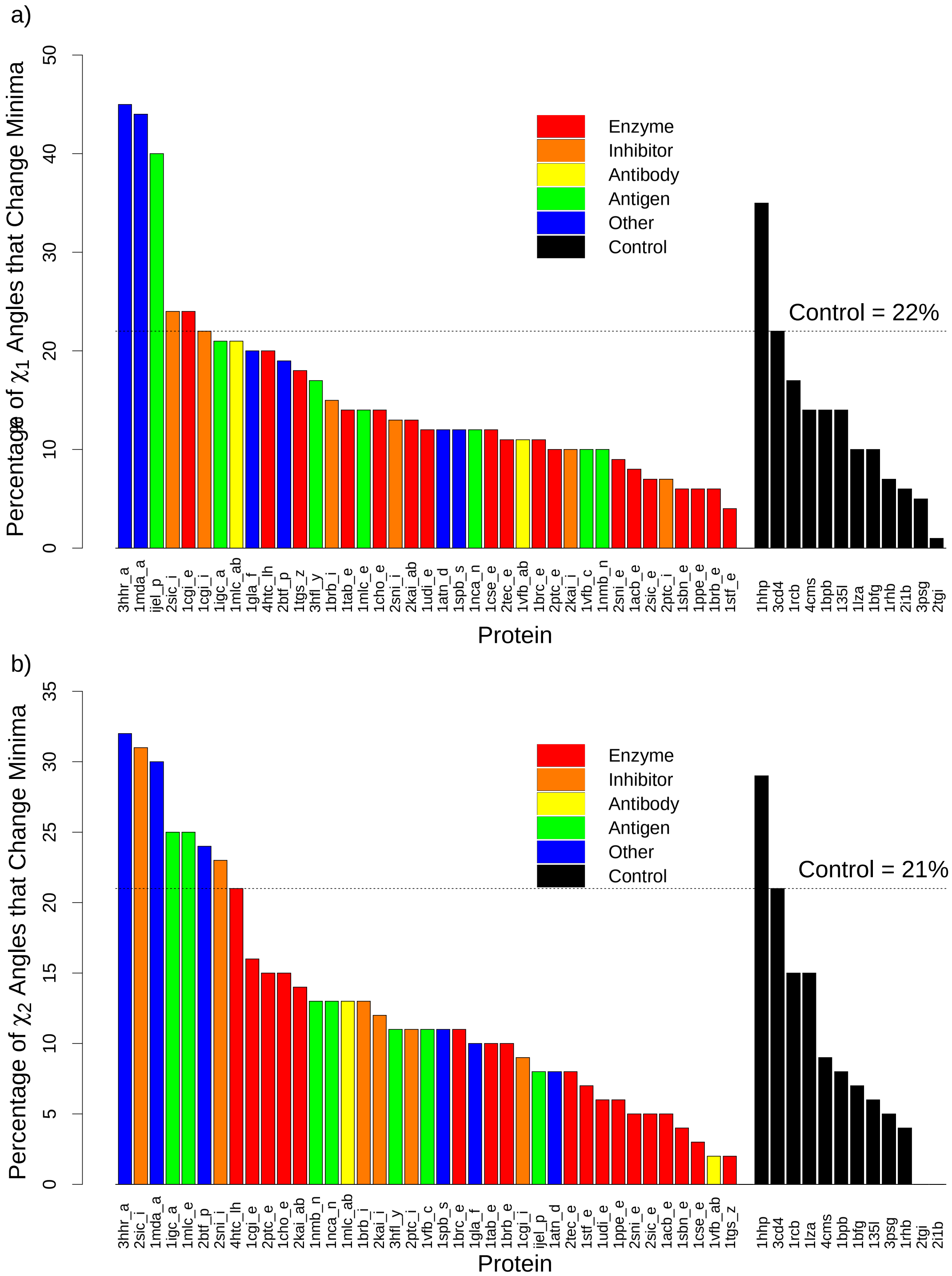
b) Percentages of χ_2's of all residues that change minima.

c) Percentages of χ_1's of interface residues that change minima.

d) Percentages of χ_2's of interface residues that change minima.

e) Percentages of χ_1's of exposed non-interface residues that change minima.

f) Percentages of χ_2's of exposed non-interface residues that change minima.



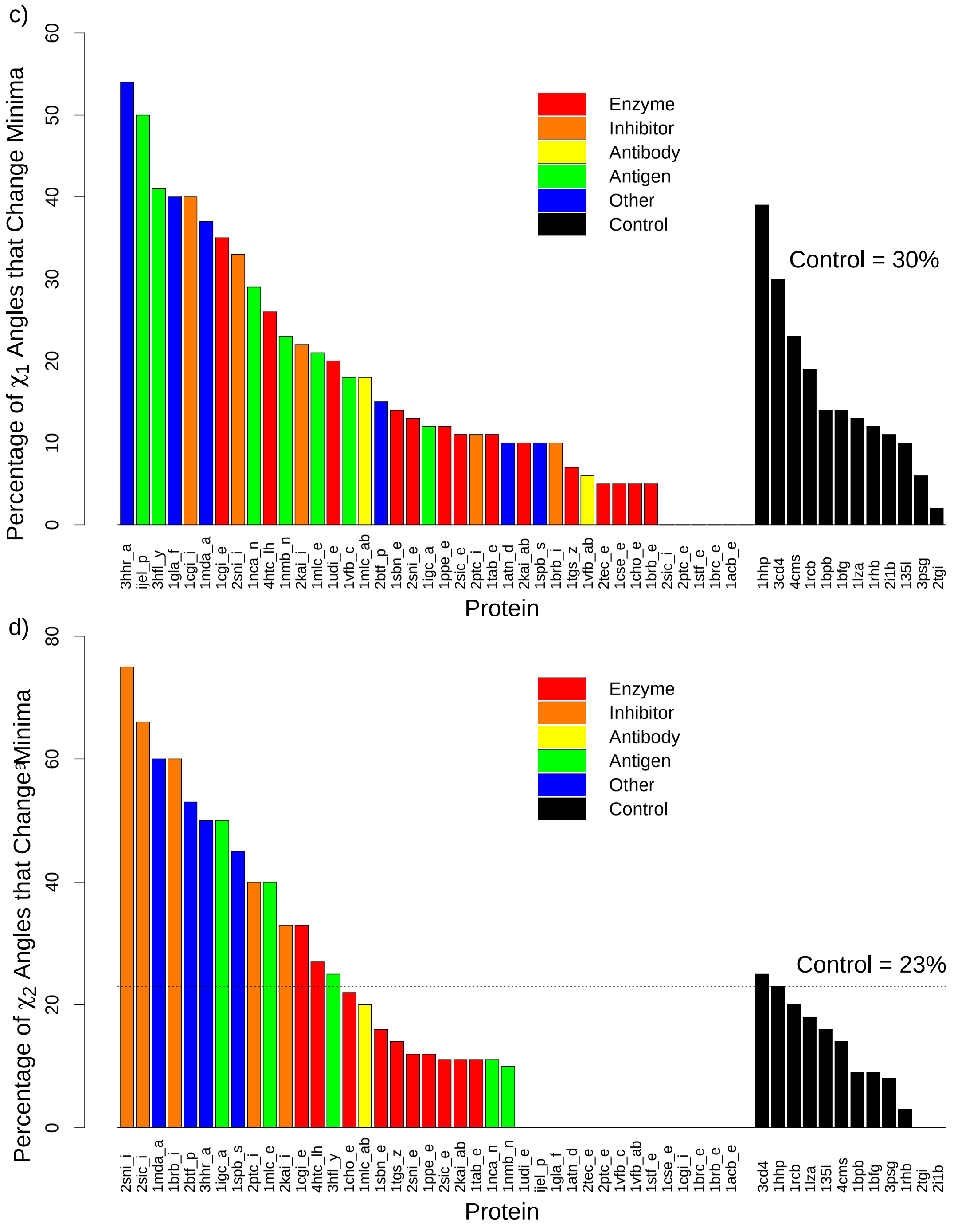


Figure 5-2 (continued)

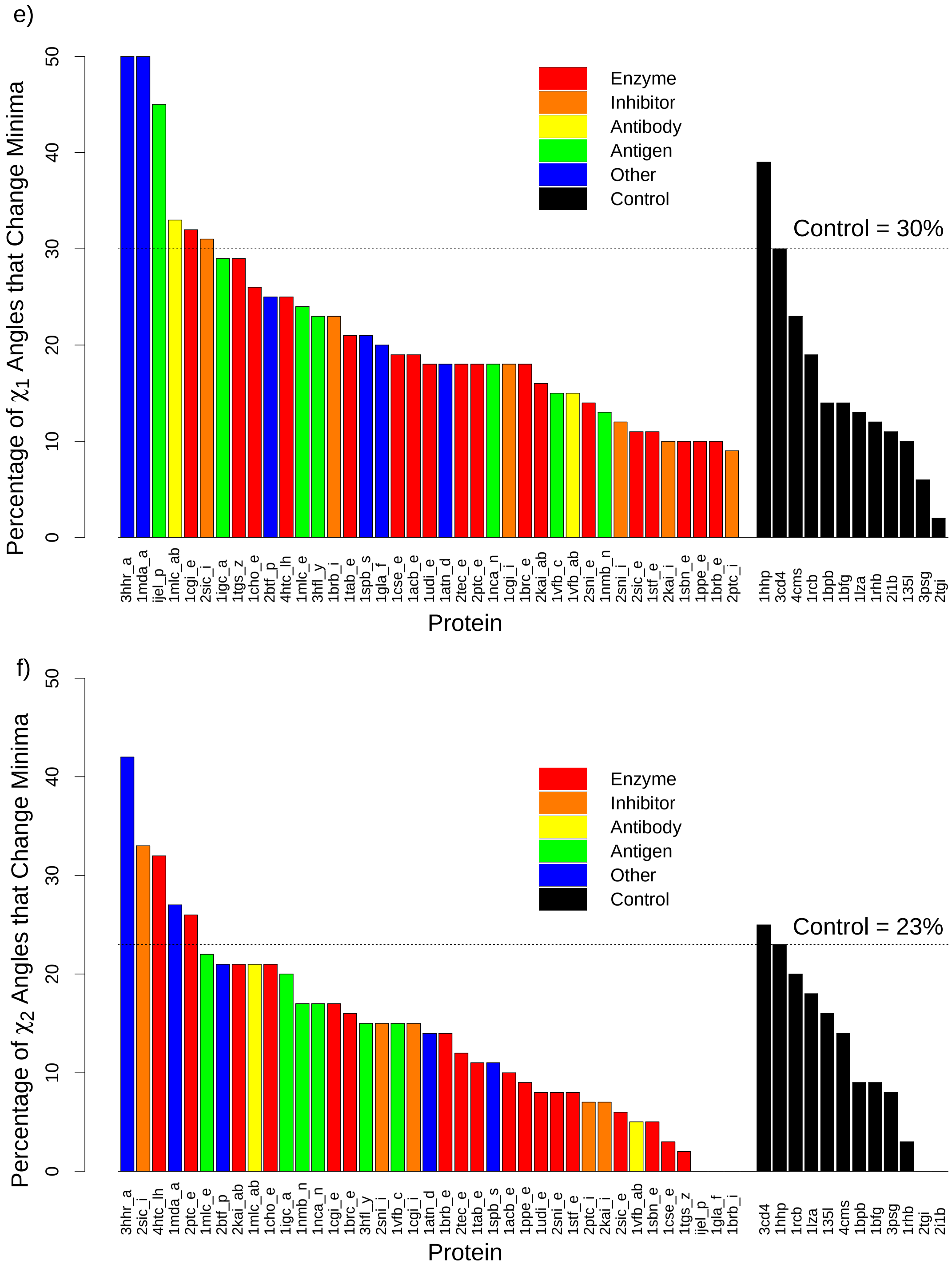


Figure 5-2 (continued)

Table 5-1 - Measurements of Conformational Differences Between Complexed and Unbound Structures

Protein[i]	All Residues					Interface Residues							Exposed Non-interface Residues						
		Overall Measures[iii]					Overall Measures[iii]				Individual Residue Measures			Overall Measures[iii]				Individual Residue Measures	
	N[ii]	RMSD / Å		% torsion change		N[ii]	RMSD / Å		% torsion change		N > Control[iv]		N[ii]	RMSD / Å		% torsion change		N > Control[iv]	
		Cα	Side-chain	χ_1	χ_2		Cα	Side-chain	χ_1	χ_2	δCα[v]	Side-chain RMSD		Cα	Side-chain	χ_1	χ_2	δCα[v]	Side-chain RMSD
1brb_e	217	**0.4**	**1.0**	**6**	**10**	20	0.4	**1.8**	5	0	1	2	105	**0.4**	**0.9**	**10**	**14**	7	4
1brb_i	50	**0.3**	**1.5**	**15**	**13**	13	0.4	**2.4**	10	**60**	0	1	29	**0.3**	**1.4**	**23**	**0**	0	3
1cgi_e	223	**1.1**	**2.0**	**24**	16	23	**2.5**	**3.4**	**35**	**33**	16	11	106	**0.9**	**2.1**	**32**	17	42	21
1cgi_i	51	**1.4**	**2.6**	22	9	16	**2.1**	**3.8**	**40**	0	13	8	25	**1.0**	1.7	18	15	14	3
2kai_ab	200	**0.6**	1.4	13	14	21	**0.4**	**1.7**	**10**	**11**	1	4	90	**0.8**	1.7	16	21	6	12
2kai_i	49	**0.5**	1.5	10	12	13	0.5	**2.9**	22	**33**	1	1	29	**0.5**	**1.1**	**10**	**7**	2	4
2ptc_e	216	**0.3**	**0.9**	**10**	**15**	20	**0.3**	**0.6**	**0**	**0**	0	1	103	0.4	1.2	18	**26**	2	7
2ptc_i	53	**1.2**	1.4	7	11	13	0.3	**2.2**	11	**40**	0	0	29	**1.6**	1.4	9	7	3	3
2sic_e	273	**0.2**	**0.8**	**7**	**5**	23	**0.3**	**1.3**	**11**	**11**	0	2	125	**0.3**	**0.9**	**11**	**6**	2	6
2sic_i	70	**0.8**	1.8	**24**	**31**	7	**0.7**	1.4	0	**66**	1	3	43	**0.8**	**2.1**	**31**	**33**	12	9
2sni_e	269	**0.2**	**0.7**	**9**	**5**	22	**0.3**	**1.4**	**13**	**12**	0	2	123	**0.3**	**0.8**	**14**	**8**	1	3
2sni_i	59	**0.5**	1.3	13	**23**	10	**1.0**	**2.1**	**33**	**75**	5	2	33	**0.3**	**1.2**	**12**	**15**	1	3
1acb_e	164	**0.4**	**0.8**	**8**	**5**	12	**0.3**	**0.4**	**0**	**0**	0	0	68	**0.5**	**1.1**	**19**	**10**	5	5
1brc_e	215	**0.4**	**1.0**	**11**	**11**	16	**0.3**	**0.8**	**0**	**0**	0	1	104	**0.5**	**1.2**	**18**	**16**	9	7
1cho_e	220	**0.4**	**1.2**	**14**	**15**	21	**0.3**	**1.7**	**5**	**22**	0	2	101	**0.5**	**1.3**	**26**	**21**	8	7

Table 5-1 - Measurements of Conformational Differences Between Complexed and Unbound Structures (Continued)

Protein[i]	All Residues					Interface Residues							Exposed Non-interface Residues						
		Overall Measures[iii]					Overall Measures[iii]				Individual Residue Measures			Overall Measures[iii]				Individual Residue Measures	
	N[ii]	RMSD / Å		% torsion change		N[ii]	RMSD / Å		% torsion change		N > Control[iv]		N[ii]	RMSD / Å		% torsion change		N > Control[iv]	
		Cα	Side-chain	χ_1	χ_2		Cα	Side-chain	χ_1	χ_2	δCα[v]	Side-chain RMSD		Cα	Side-chain	χ_1	χ_2	δCα[v]	Side-chain RMSD
1cse_e	274	**0.4**	**1.0**	**12**	**3**	26	**0.4**	**0.8**	**5**	**0**	0	2	123	**0.4**	**1.3**	**19**	**3**	11	8
1ppe_e	203	**0.3**	**0.7**	**6**	**6**	20	**0.2**	**0.6**	**12**	**12**	0	0	90	**0.4**	**0.8**	**10**	**9**	5	4
1sbn_e	228	**0.3**	**0.6**	**6**	**4**	21	**0.4**	**1.3**	**14**	**16**	1	1	84	**0.4**	**0.6**	**10**	**5**	2	2
1stf_e	179	**0.3**	**0.9**	**4**	**7**	20	**0.4**	**0.5**	**0**	**0**	0	0	67	**0.3**	**1.4**	**11**	**8**	0	3
1tab_e	223	**0.4**	**1.0**	**14**	**10**	20	**0.3**	**1.1**	**11**	**11**	1	2	116	**0.4**	**1.2**	**21**	**11**	3	9
1tgs_z	188	**0.7**	1.2	18	2	14	**0.6**	**1.3**	**7**	**14**	4	2	92	**0.8**	1.4	29	2	17	10
2tec_e	273	**0.2**	**0.8**	**11**	**8**	26	**0.3**	**0.5**	**5**	**0**	0	0	116	**0.3**	**1.0**	**18**	**12**	0	4
4htc_lh	252	**1.0**	**1.7**	20	21	25	0.6	1.5	26	**27**	5	4	102	**1.4**	**2.3**	25	**32**	23	21
1udi_e	210	**0.5**	1.0	12	6	19	**0.6**	**1.5**	**20**	**0**	4	5	96	**0.5**	**1.3**	**18**	**8**	10	6
1mlc_ab	422	**0.9**	1.4	21	13	19	**1.1**	1.4	18	20	12	3	238	**1.0**	1.6	**33**	21	111	30
1mlc_e	84	**0.6**	1.5	14	**25**	16	**0.8**	**2.3**	21	**40**	2	2	38	**0.6**	**1.6**	**24**	**22**	5	5
1vfb_ab	199	**0.4**	**0.8**	**11**	**2**	17	**0.5**	**1.1**	**6**	**0**	2	5	103	**0.4**	**0.9**	**15**	**5**	6	2
1vfb_c	103	**1.1**	**1.8**	10	11	14	**2.1**	**2.5**	18	0	2	2	49	**1.2**	**2.2**	15	15	3	6
1nca_n	387	**0.3**	**0.9**	**12**	**13**	19	**0.4**	**0.9**	**29**	**11**	0	2	177	**0.3**	**1.1**	**18**	**17**	2	6
1nmb_n	388	**0.3**	**1.0**	**10**	**13**	19	**0.3**	**1.1**	**23**	**10**	0	2	177	**0.3**	**1.2**	**13**	**17**	0	4

Table 5-1 - Measurements of Conformational Differences Between Complexed and Unbound Structures (Continued)

Protein[i]	All Residues					Interface Residues							Exposed Non-interface Residues						
		Overall Measures[iii]					Overall Measures[iii]				Individual Residue Measures			Overall Measures[iii]				Individual Residue Measures	
	N[ii]	RMSD / Å		% torsion change		N[ii]	RMSD / Å		% torsion change		N > Control[iv]		N[ii]	RMSD / Å		% torsion change		N > Control[iv]	
		Cα	Side-chain	χ_1	χ_2		Cα	Side-chain	χ_1	χ_2	δCα[v]	Side-chain RMSD		Cα	Side-chain	χ_1	χ_2	δCα[v]	Side-chain RMSD
1igc_a	40	**0.7**	1.3	21	**25**	10	**1.0**	1.5	12	**50**	6	2	20	**0.6**	**1.3**	**29**	**20**	6	2
1jel_p	63	**0.7**	**2.2**	**40**	8	9	0.4	**1.8**	**50**	0	0	2	28	**0.8**	**3.0**	**45**	0	5	5
3hfl_y	129	**0.5**	1.6	17	11	14	0.6	1.7	**41**	**25**	1	2	71	**0.6**	**2.0**	**23**	**15**	8	10
1atn_d	251	**0.3**	**0.9**	**12**	**8**	21	**0.4**	**1.1**	**10**	**0**	1	3	111	**0.4**	**1.2**	**18**	**14**	3	6
1gla_g	55	**0.4**	**0.8**	**20**	**10**	5	0.6	1.1	**40**	0	1	0	12	**0.4**	**0.8**	**20**	**0**	0	0
1spb_s	262	**0.6**	1.2	12	11	34	0.5	1.2	10	**45**	2	2	111	**0.6**	**1.5**	**21**	**11**	21	16
2btf_p	139	**0.8**	1.4	19	**24**	21	0.4	1.6	15	**53**	0	2	73	**1**	**1.5**	**25**	**21**	10	12
3hhr_a	164	**3.4**	**5.0**	**45**	**32**	33	**1.9**	**4.1**	**54**	**50**	26	18	70	**4.1**	**5.9**	**50**	**42**	65	45
1mda_lh	103	**2.6**	**3.4**	**44**	**30**	8	**1.5**	**2.8**	**37**	**60**	8	3	68	**2.9**	**3.9**	**50**	**27**	50	36

i. Proteins are identified by the PDB code of the complex, followed by the chain identifier(s) of the relevant component in that complex. For ease of reference, the order of the proteins is the same as that in table 3-2.

ii. N = the number of residues. This number only includes well defined residues (see section 3.2.1) common to both the complexed and the unbound structure.

iii. Red shading indicates a value that is above the control (table 4-1). Yellow shading indicates that all four overall measures for a particular region (all residues, interface residues, or exposed non-interface residues) are less than or equal to the controls (table 4-1).

iv. N > Control = the number of residues that have changes that are greater than the controls for individual residues (table 4-2). This number only includes well defined residues (see section 3.2.1) common to both the complexed and the unbound structure.

v. δCα = Cα displacement.

Table 5-2 - Number of Pairs of Complexed and Unbound Structures with Overall Measurements that are Greater than the Control Values

Region	Number of Pairs > Control Values[i]			
	Cα RMSD / Å	Side-chain RMSD / Å	% χ_1 change	% χ_2 change
All Residues	20	8	5	7
Interface Residues	10	12	8	14
Exposed Non-interface Residues	13	8	6	5

i. Control values given in table 4-1.

5.3 Large Individual Residue Movements

The Cα displacements and side-chain RMSD’s of individual residues were compared to the 95% control values for the relevant amino acid type (table 4-2), and those that had values greater than the controls are described here and in section 5.3.2. Figure 5-1 shows counts of these residues for each pair of complexed and unbound structures, alongside the overall Cα and side-chain RMSD’s of interface and exposed non-interface residues. These counts are also given in table 5-1. They vary widely for pairs with an overall measure above the appropriate control limit. Some of those pairs have counts of zero, for example the interface side-chains of pancreatic trypsin inhibitor in the complex with β-trypsin (2ptci_i on figure 5-1d). This shows that individual residue movements below the individual controls can amount to a substantial measure for the whole region. Other pairs have one or two individual residues with substantial movements, for example the interface side-chains of pancreatic trypsin inhibitor in the complex with kallikrein (2kai_i on figure 5-1d). This demonstrates that movements of a few residues in a region can dominate measures of overall change of those regions, especially when the total number of residues in those regions is small (in both of the examples described there are only thirteen interface residues - see table 3-2). At the other end of the scale are cases such as the interface side-chains of pancreatic trypsin inhibitor in the complex with α-chymotrypsinogen (1cgi_i on figure 5-1d), where a high proportion of the residues have substantial individual movements. The wide variation in the counts indicates that in addition to looking at overall measures, it is important to look at the number and causes of substantial individual movements.

5.3.1 Exposed Non-interface Residues

All of the largest Cα displacements (above 3Å) and side-chain RMSD’s (above 5.6Å) of exposed non-interface residues can be explained by one of the causes given below. These limits are greater than the control limits given in table 4-2. Conformational differences with values between the two sets of limits may be a sign of additional experimental differences, caused by different crystal packing, in the determination of complexed and unbound structures. The causes are listed here, together with examples of movements that can be explained by them. Full lists of such movements are provided in table 5-3 for Cα displacements, and table 5-4 for side-chain RMSD’s.

a) The residue is adjacent to an interface residue that moves, and therefore is part of a loop movement in the interface. For example Aspartic Acid 101 and Asparagine 103 on either side of Glycine 102, which is in the interface of lysozyme complexed to antibody D1.3 (see section 5.3.2 and figure 5-3b). These two residues have Cα displacements of 6.3Å and 4.4Å and side-chain RMSD's of 8.1Å and 7.3Å respectively. In such cases the whole loop has not been classified as interface, because not all the residues that make up the loop have at least one atom 4Å or less from the other component of the complex (see "Interface Residues", Chapter Three, page 94).

b) The residue is at the end of a chain, or only one to three residues away. For example the N-terminal Alanine of β-actin complexed to profilin, which has a Cα displacement of 6.6Å and a side-chain RMSD of 7.7Å, and the C-terminal Glutamine of α-thrombin complexed to hirudin, which has a Cα displacement of 5.9Å and a side-chain RMSD of 10.7Å.

c) The residue is at the end of a cleavage fragment, or only one to three residues away. For example Aspartic Acid 14l of α-thrombin complexed to hirudin, which has a Cα displacement of 10.6Å and a side-chain RMSD of the same size. The unbound structure of this protein used in the comparison was actually γ-thrombin, which is cleaved in several places by autolysis (Rydel et al., 1994).

d) The residue is adjacent to a region missing from or poorly defined in the electron density map. A good example of this is amicyanin complexed with methylamine dehydrogenase. In this protein the first fifteen N-terminal residues form an irregular outer β-strand connected to a loop of six residues that are poorly defined in the electron density (Durley et al., 1993). These fifteen residues have Cα displacements that vary between 3.2 and 8.3Å, and side-chain RMSD's between 3.8 and 10.8Å.

Hence the largest movements of exposed residues that are not in the interface can be explained by either their close proximity to the interface (point 'a' in the list), or by structural disorder and flexibility (points 'b', 'c', and 'd'). Structural disorder and flexibility are also the causes of differences greater than the controls in the pairs of structures from which the controls were calculated (see section 4.2). They are not due to hinge-bending or shear movements between domains as sometimes seen when small molecules bind (Gerstein et al., 1994). An exception to these generalities is human growth hormone complexed with its receptor (and thus table 5-3 and table 5-4 do not contain information for this protein). This is a four helix bundle with two long crossovers connecting the first two and last two helices, and a short loop that connects the middle two. The main changes occur in these connections and involve many interface residues (Chantalat et al., 1995) - see section 5.3.2 - but also extend outside the interface regions.

Table 5-3 - Explanations for all Exposed Non-interface Cα Displacements that are Greater than 3Å

Protein[i]	Residue[ii]	Cα Displacement / Å	Explanation of Difference
4htc_lh	D14L (L)	10.6	Fragment end.
2ptc_i	A58	8.4	C-terminus.
1mda_a	I5	8.3	Connected to poorly defined region.
1mda_a	A3	8.3	Connected to poorly defined region.
1mda_a	S7	8.0	Connected to poorly defined region.
1mda_a	S9	7.6	Connected to poorly defined region.
1mda_a	T4	6.9	Connected to poorly defined region.
2btf_p	A1	6.6	N-terminus.
1vfb_c	D101	6.3	Adjacent to interface mover.
4htc_lh	Q244 (H)	5.9	C-terminus.
1mda_a	E8	5.5	Connected to poorly defined region.
1mda_a	P10	5.1	Connected to poorly defined region.
1mda_a	A13	5.1	Connected to poorly defined region.
2kai_ab	A171 (B)	5.0	Adjacent to region missing from e-density.
4htc_lh	I14K (L)	4.3	Fragment end.
1vfb_c	N103	4.2	Adjacent to interface mover.
2kai_ab	H172 (B)	4.2	Adjacent to region missing from e^- density.
1mlc_ab	E213 (A)	3.9	Adjacent to C-terminus (which is poorly defined).
1mda_a	A14	3.7	Connected to poorly defined region.
1mda_a	F11	3.7	Connected to poorly defined region.
1mda_a	A17	3.7	Connected to poorly defined region.
1cgi_e	E78	3.7	Adjacent to region not located in e^- density.
1mda_a	A20	3.5	Connected to poorly defined region.
1mda_a	M72	3.5	Between two interface movers.
1mlc_ab	N212 (A)	3.4	Two residues away from C-terminus.
1jel_p	L84	3.3	Adjacent to C-terminus (which has high B-factor).
1mda_a	E15	3.2	Connected to poorly defined region.
1mda_a	V16	3.1	Connected to poorly defined region.
1mda_a	A50	3.1	Adjacent to interface mover.

i. Identified by the PDB code of the complex, followed by the chain indentifier(s) of the component. Human growth hormone (3hhr_a) is excluded as many of its exposed non-interface residues move as a direct result of receptor binding (see section 5.3.1).

ii. Identified by one letter amino acid code, number, and insertion code (if any). If the component has more than one chain, the chain identifier for the residue is given in brackets.

Table 5-4 - Explanations for all Exposed Non-interface Side-chain RMSD's that are Greater than 5.6Å

Protein[i]	Residue[ii]	Side-chain RMSD / Å	Explanation
1mda_a	I5	10.8	Connected to poorly defined region.
4htc_lh	Q244 (H)	10.7	C-terminus.
4htc_lh	D14L (L)	10.6	Fragment end.
1mda_a	S7	10.3	Connected to poorly defined region.
2ptc_i	A58	10.0	C-terminus.
1mda_a	S9	9.9	Connected to poorly defined region.
1cgi_e	E78	9.4	Adjacent to region not located in e^- density.
1mda_a	T4	9.1	Connected to poorly defined region.
1mda_a	M72	9.0	Between two interface movers.
2kai_ab	H172 (B)	8.4	Adjacent to region missing from e^- density.
1mda_a	P10	8.3	Connected to poorly defined region.
1vfb_c	D101	8.1	Adjacent to interface mover.
1mda_a	K74	7.8	Adjacent to interface mover.
1mda_a	A3	7.8	Connected to poorly defined region.
2btf_p	A1	7.7	N-terminus.
1nmb_n	R82	7.7	N-terminus.
1mda_a	F11	7.6	Connected to poorly defined region.
1jel_p	L84	7.6	Adjacent to C-terminus (which has high B-factor).
1vfb_c	N103	7.3	Adjacent to interface mover.
4htc_lh	I14K (L)	7.0	Fragment end.
1mda_a	A13	6.8	Connected to poorly defined region.
2kai_ab	A171 (B)	6.7	Adjacent to region missing from e^- density.
1stf_e	R59	6.7	Adjacent to poorly defined region.
1vfb_c	R73	5.7	Adjacent to poorly defined region.
4htc_lh	D243 (H)	5.6	Adjacent to C-terminus.

i. Identified by the PDB code of the complex, followed by the chain indentifier(s) of the component. Human growth hormone (3hhr_a) is excluded as many of its exposed non-interface residues move as a direct result of receptor binding (see section 5.3.1).

ii. Identified by one letter amino acid code, number, and insertion code (if any). If the component has more than one chain, the chain identifier for the residue is given in brackets.

5.3.2 Interface Residues

Changes in interfaces occur for a variety of reasons: to form specific interactions required for the action of the protein, to avoid steric clash, or to improve shape complementarity and allow hydrogen bonding (Janin and Chothia, 1990). The largest changes of interface residues, i.e. those above 3Å Cα displacement and 5.6Å side-chain RMSD (above which movements of exposed non-interface residues could be explained by flexibility or structural disorder - see section 5.3.1) are discussed below.

Changes that allow the formation of specifically required interactions are the largest and most extensive seen in the structures examined. When chymotrypsinogen binds to human pancreatic secretory trypsin inhibitor (PDB code 1cgi), the specificity pocket and oxyanion hole necessary for inhibitor binding are formed by large movements of loops serine 189 - serine 195 and valine 213 - cystine 220 towards the inhibitor (figure 5-3a). This change is the same as occurs when the zymogen is activated by hydrolysis. Smaller Cα shifts of inhibitor loop tyrosine 10 - arginine 21, along with side-chain movements towards the enzyme of some of these residues, alter the pattern of hydrogen bonding and allow binding to chymotrypsinogen. The changes are largely the same as those noted by Hecht et al., 1991 and Hecht et al., 1992.

Specifically required interactions in the interface between human growth hormone and its receptor (PDB code 3hhr) are also formed by large changes. This complex involves one hormone molecule binding to a dimer of receptors, and it is thought that this dimerisation is caused by hormone binding and that it is the mechanism of signal transduction (Chantalat et al., 1995). Large changes are required for different parts of the hormone to bind to structurally identical parts of each receptor molecule. The biggest occur mainly in the long crossover loop between helices one and two and the short loop between helices two and three (figure 5-3e). Tyrosine 103 on the short loop is involved in receptor binding (Chantalat et al., 1995), and moves by a side-chain RMSD of 8.5Å towards the interface. This change is accommodated by large associated movements of glycine 104 - asparagine 109 away from the interface (Cα displacements up to 11.5Å, and side-chain RMSD’s up to 14.7Å). Other smaller but still extensive changes (Cα displacements up to 5.4Å and side-chain RMSD’s up to 7.7Å) occur in the long crossover loop. They improve surface complementarity by moving away from the interface and forming mini-helices, rather than hydrogen bonding to helix four in a position that would clash with the receptor.

Interactions that appear to be less necessary for function, because they simply alleviate minor steric clash or improve hydrogen bonding and van der Waals contacts, are noticeably less extensive. However, they can still involve large changes of a few residues. Figure 5-3b shows changes of this nature that occur when the interface between hen egg white lysozyme and the variable domain of antibody D1.3 (PDB code 1vfb) is formed. Glycine 102 of lysozyme moves with a Cα displacement of 7.5Å, which brings it to within 2.1Å of arginine 99 on the heavy chain of the antibody. Movement of arginine 99 was noted in a comparison of complexed and unbound antibody (Bhat et al., 1994), along with a decrease in its mobility as shown by a decrease in temperature factor. The two residues either side of lysozyme glycine 102 (aspartic acid 101 and asparagine 103) are not classified as interface but also move significantly - they are part of a loop movement. Another large but isolated discrete change occurs with arginine 125 of lysozyme (side-chain RMSD = 6.3Å), with the possible creation of a hydrogen bond to serine 93 on the light chain of the antibody. In other complexes, discrete changes not directly related to function occur to improve electrostatic complementarity, for example the movement of lysine 73 of amicyanin on binding to methylamine dehydrogenase (PDB code 1mda, figure 5-3c), or to positions that would be highly exposed to solvent if adopted in the unbound structure, for example phenylalanine 39 of α-chymotrypsin (PDB code 1cho, figure 5-3d).

Figure 5-3 - Examples of Large Changes in Interfaces.

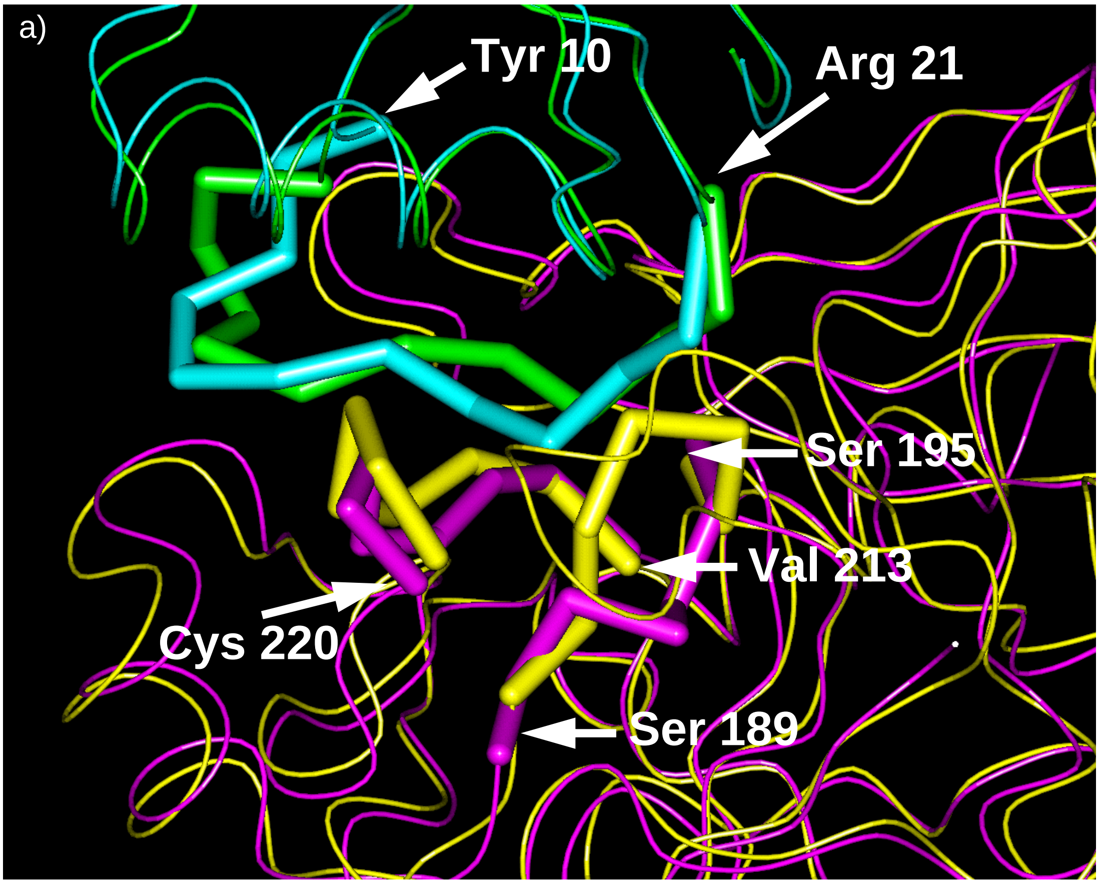
a) Chymotrypsinogen (yellow = complex, mauve = unbound) complexed with human pancreatic trypsin inhibitor (green = complex, cyan = unbound).

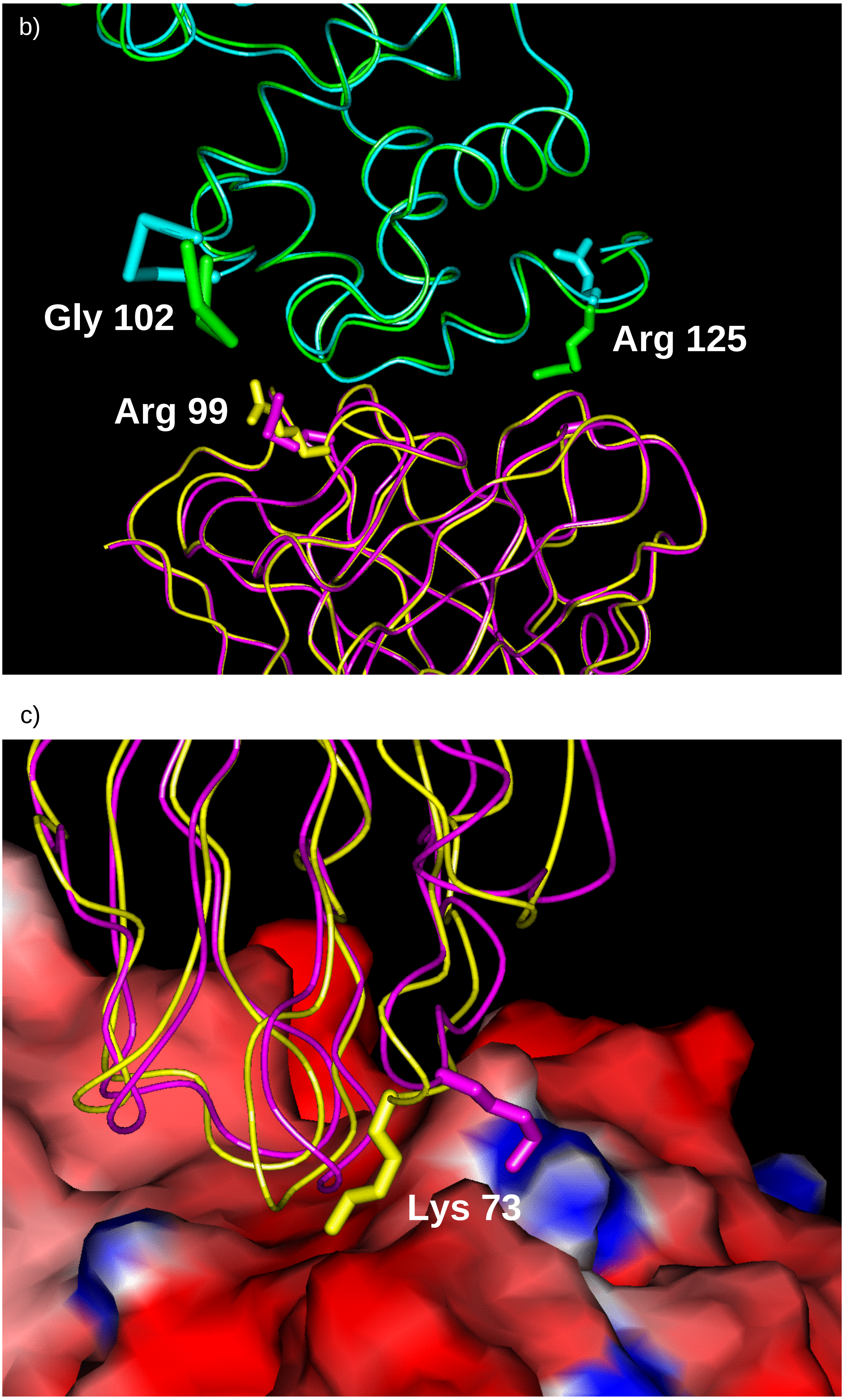
b) Antibody D1.3 (yellow = complex, mauve = unbound) complexed with lysozyme (green = complex, cyan = unbound).

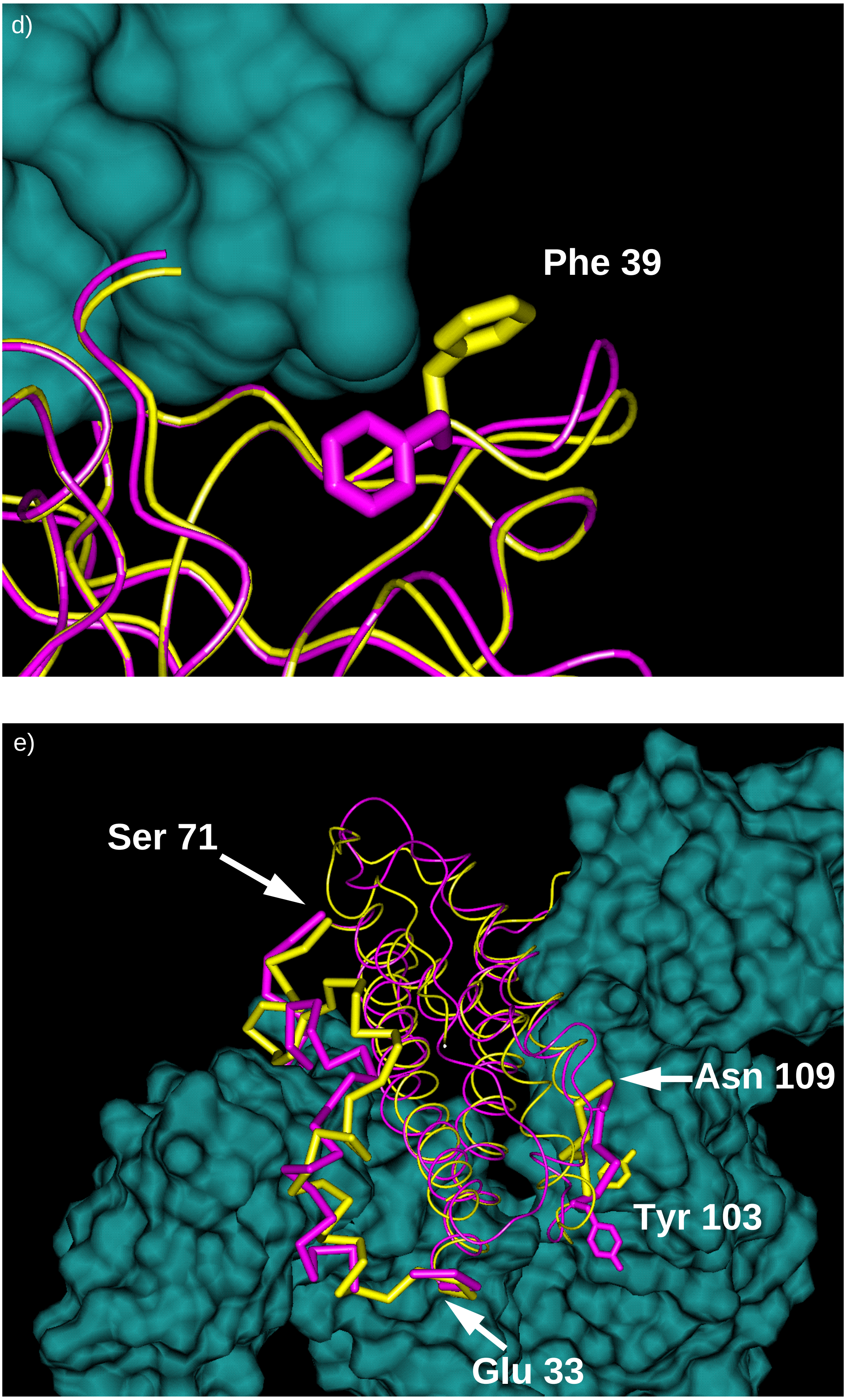
c) Amicyanin (yellow = complex, mauve = unbound) complexed with methylamine dehydrogenase (molecular surface coloured by potential = complex).

d) Chymotrypsin (yellow = complex, mauve = unbound) complexed with ovomucoid (cyan coloured molecular surface = complex).

e) Human growth hormone (yellow = complex, mauve = unbound) complexed with human growth hormone receptor (cyan coloured molecular surface = complex).




b)
Gly 102
Arg 125
Arg 99
c)
Lys 73


d)
Phe 39
e)
Ser 71
Asn 109
Tyr 103
Glu 33

5.4 Do Interface Regions Move More Than Exposed Non-interface Regions?

To answer this question, it is only meaningful to look at those systems where the measurements of movements (defined in section 3.3.3) of the interface and / or the exposed non-interface regions are greater than the movements of exposed residues in the controls (table 4-1). Figure 5-4 shows four plots of interface measurements against exposed non-interface measurements, one for each of Cα RMSD, side-chain RMSD, and percentages of χ_1's and of χ_2's that change minima. On each plot the control value from table 4-1 that is appropriate to the measurement is indicated by two solid lines. One is in the vertical direction, and any point to the right of this line indicates a pair of complexed and unbound structures where the conformation of exposed-non interface residues differs more than the control. The other is horizontal, and any point above it is for a pair of structures where the conformation of interface residues differ more than the control. Thus any point in the bottom-left corner of a plot is for a pair of structures where neither the exposed non-interface residues or the interface residues move more than the control. The line described by y = x is also displayed on each plot. This emphasises those pairs where the differences of their interface residues are larger than the differences of their exposed non-interface residues (plotted above the y = x line) or vice-versa (below the line).

The results presented in figure 5-4 suggest that side-chain movements in interfaces have greater conformational change than other exposed parts of the structures - the plots for the three measurements of side-chain change (side-chain RMSD, figure 5-4b, and percentages of χ_1 's, figure 5-4c, and of χ_2's, figure 5-4d, that change minima) all have more points above the line y = x than below it (see table 5-5). This is probably due to the fact that changes in the interface occur for specific reasons, rather than simply as a result of flexibility or disorder (see section 5.3). This is shown most strongly by the percentages of χ_2's that change minima - of those pairs outside the bottom-left corner, thirteen have greater values for their interface regions than they do for their exposed non-interface regions, and the reverse is true for only one pair. Figure 5-4a and table 5-5 show that eight pairs have greater movement of the main-chain (measured by Cα RMSD) for exposed non-interface regions than they do for interface regions, where as the reverse is true for seven pairs. However, the situation changes if two pairs are removed: human growth

Figure 5-4 - Comparisons of Conformational Changes of Interface Regions with Those of Exposed Non-interface Regions.

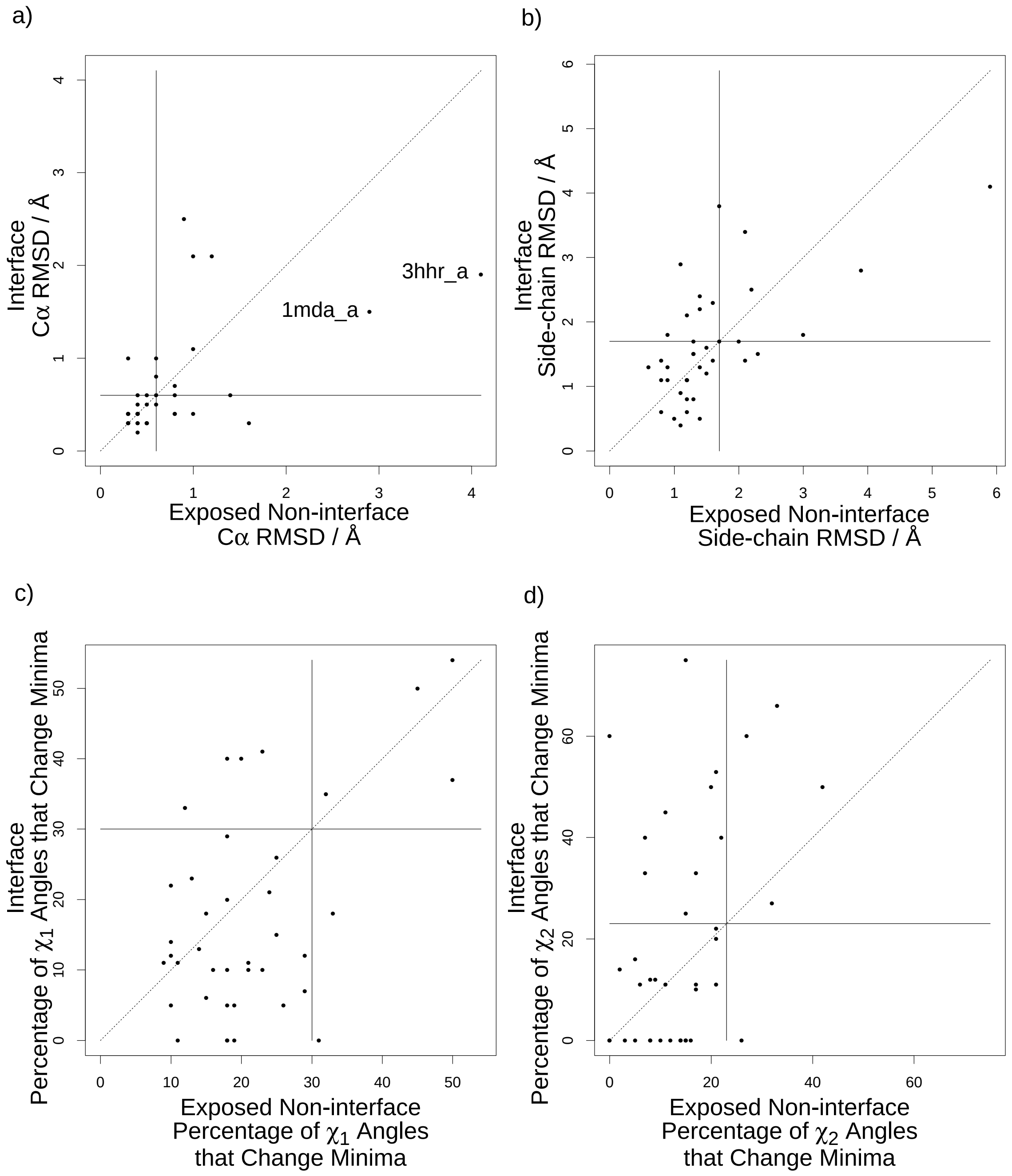
The solid lines show the controls - values expected from differences in the experimental determination of the structures (table 4-1). Therefore differences are only substantial when outside the bottom left section marked out by the solid lines. The dotted lines are for y = x, displayed to clarify the differences between the regions.

a) Cα RMSD. Two points are identified by the PDB code of the complex, followed by the chain identifier of the component considered. These two proteins have much greater differences of their exposed non-interface regions than of their interface regions, and the reasons for this are discussed in the text (section 5.4).

b) Side-chain RMSD.

c) Percentages of χ_1's that change minima.

d) Percentages of χ_2's that change minima.



hormone complexed with its receptor (labelled as 3hhr_a on figure 5-4a), where receptor binding causes changes away from the interface (see section 5.3.1), and amicyanin complexed with methylamine dehydrogenase, where the fifteen N-terminal residues are connected to a region poorly defined in the electron density (also discussed in section 5.3.1).

Table 5-5 - The Numbers of Pairs of Complexed and Unbound Structures Where Conformational Differences of their Interface and / or their Exposed Non-interface Residues are Greater than the Control Values.

Measurement	Number of Pairs of Structures Where:	
	Interface Measurement > Exposed Non-interface Measurement	Exposed Non-interface Measurement > Interface Measurement
Cα RMSD	7	8
Side-chain RMSD	10	6
Percentage of χ_1’s that change minima	7	1
Percentage of χ_1’s that change minima	13	1

5.5 Do Side-chains Move More Than Main-chains?

It would be useful to know if side-chain movements are more substantial than those of main-chains, as this would provide additional justification for the approach of docking procedures that simulate flexibility only in the side-chains of interface residues (for example Weng et al., 1996, and Jackson et al., 1998). Figure 5-5 shows a comparison of the side-chain RMSD's against the Cα RMSD's of the exposed regions of the control systems (plotted as crosses). This plot confirms that the side-chain RMSD's always have the larger values of the two measurements - all the crosses are above the dotted line defined by ‘y = x’, where ‘y’ is side-chain RMSD and ‘x’ is Cα RMSD. This is reasonable because more atoms contribute to side-chain RMSD, and the side-chains are less constrained by local interactions.

Also on figure 5-5, the side-chain RMSD's of the interfaces of the complexed-unbound pairs are plotted against their Cα RMSD's (plotted as dots). As for the exposed residues of the control systems, all have side-chain RMSD's that are greater than their Cα RMSD's. However, some are outliers from the least-squares line calculated from the control pairs - i.e. the ratio between their side-chain and Cα RMSD's is smaller or larger than seen in the controls. The four largest outliers are identified on figure 5-5 and discussed here.

Hen egg white lysozyme (labelled 1vfb_c on figure 5-5) bound to antibody D1.3 deviates most from the least-squares line, with interface side-chain and Cα RMSD almost equal to each other (2.5Å and 2.1Å respectively - see table 5-1). The changes in the interface of this structure have already been examined in section 5.3.2. The largest movement was made by glycine 102, which obviously has no side-chain. The changes in the interface of chymotrypsinogen and PTI bound to each other (1cgi_e and 1cgi_i) were also discussed in section 5.3.2, and also show a ratio of side-chain to Cα RMSD that is less than the ratio seen in the controls. They involve movements of short loops (i.e. main-chain), with accompanying side-chain movements that improve binding. The situation is reversed in the interface of PTI bound to Kallikrein - the ratio of side-chain to Cα RMSD is greater than seen in the controls (the point labelled 2kai_i on figure 5-5 is to the left of the least-squares line). The major change in this structure is made by the side-chain of arginine 17, and avoids steric clash (see figure 5-6).

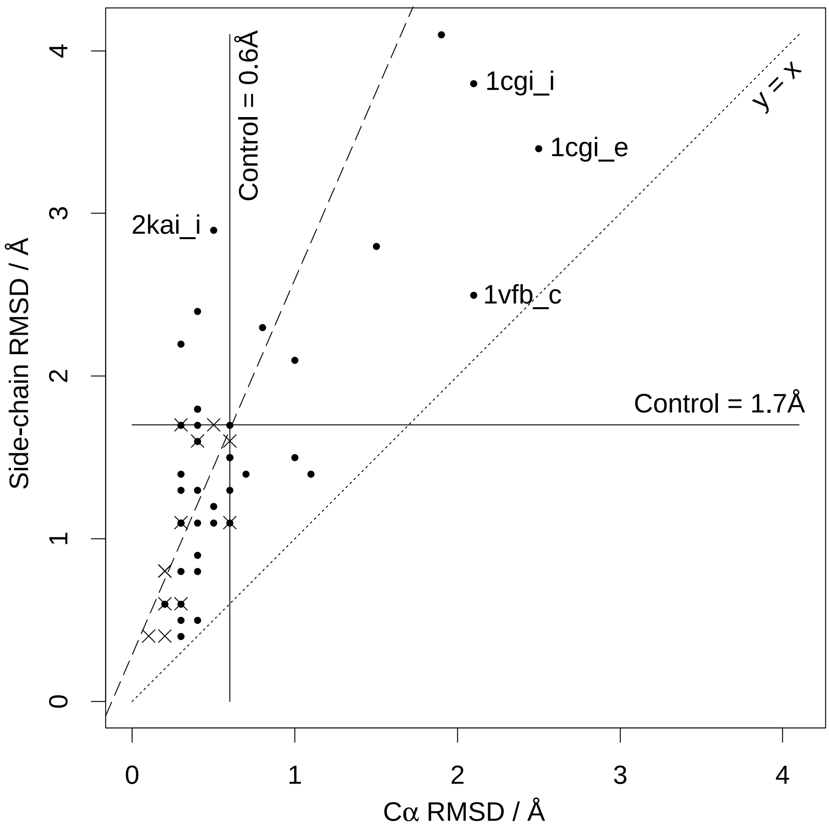
This analysis shows that side-chain RMSD’s are greater than Cα RMSD’s, and so to some extent justifies the simulation of flexibility in side-chains only. However, the side-chain RMSD’s are sometimes caused by main-chain movements, and thus simulation of backbone flexibility is required to satisfactorily model the observed changes. The modelling of side-chain flexibility alone will limit the effectiveness of docking programs.

Figure 5-5 - The Relationship Between Side-chain and Cα RMSD

Crosses: exposed residues of pairs of independently solved structures of identical proteins.

Circles: interface residues of pairs of complexed and unbound structures.

The solid lines show the 95% control values (table 4-1). The dotted line is for y = x, displayed to clarify the differences between the measures. The broken line is a least-squares fit of the data points for exposed residues of pairs of independently solved structures of identical proteins (the crosses). Pairs of structures discussed in the text (section 5.5) are identified by the PDB code of the complex followed by the chain identifier of the relevant component (see table 3-2).



5.6 Differences Between Different Types of Component.

The thirty-nine pairs of complexed and unbound structures in table 3-2 can be separated by their function into five general types. Eighteen are enzymes, six are inhibitors, two are antibodies, seven are antigens, and the remaining six are of other types. In this section different types of components in the same complex are compared. In other words, enzymes are compared with inhibitors and antibodies are compared with antigens. The analysis of the others is presented in section 5.7, in which the different types of complex (enzyme-inhibitor, antibody-antigen, and other) are compared with each other. Only the conformational changes of interface residues were compared, because it has already been seen that the changes of exposed non-interface residues are primarily caused by flexibility and disorder (see section 5.3).

All four measures of overall conformational change of the interfaces were examined. These are Cα RMSD, figure 5-1c, side-chain RMSD, figure 5-1d, and percentages of χ_1's and χ_2's that change minima, figure 5-2c and figure 5-2d. The numbers of pairs of structures of each type that have conformational differences greater than the controls are summarised in table 5-6.

Table 5-6 - The Numbers of Pairs of Complexed and Unbound Structures of Particular Types Where Conformational Differences of their Interface are Greater than the Control Values

Measurement	Number of Pairs of Structures[i] with Interface Conformational Difference > Control[ii]			
	Enzymes (18)	Inhibitors (6)	Antibodies (2)	Antigens (7)
Cα RMSD	1	3	1	2
Side-chain RMSD	2	5	0	3
Percentage of χ_1's that change minima	1	2	0	2
Percentage of χ_1's that change minima	2	5	0	2

i. Total number of pairs of complexed and unbound structures of each type given in brackets.
ii. 95% control values given in Table 4-1.

For each of the measures, almost all (sixteen or seventeen) of the eighteen pairs of enzyme structures, which are denoted by red bars on the figures mentioned, have measurements that are equal to or lower than the control values. Of the six pairs of inhibitor structures (orange bars), between two and five have values that are greater than the controls. This suggests that conformational changes in the interfaces of inhibitors are much more common than in enzymes. The two pairs of antibody structures (yellow bars) do not have values greater than the controls, except for the Cα RMSD of antibody D44.1. Also, the majority (four or five) of the seven pairs of antigen structures in the data set, shown as green bars on the figures, do not have values greater than the controls, suggesting that both antibodies and antigens seldom have substantial interface conformational changes.

However, when comparing different types of components it is better to compare different components from the same complex. This will show whether conformational change in one component is compensated by conformational change in another, or if one component changes to fit a largely motionless partner, or if both are static. It also ensures that any differences seen are not simply because there are more cases of a particular type of component available in both complexed and unbound forms.

In table 3-2 there are eight complexes (six enzyme-inhibitors and two antibody-antigens) which have both of their components solved in an unbound form. A comparison of overall RMSD’s is inappropriate here, because inhibitors and antigens have smaller interfaces than their partners in the complexes (see table 5-1), with between thirty and eighty-four percent of the number of residues. Thus the same number of large side-chain movements will give a bigger overall RMSD in these components than they would in their partners. The number of individual interface residues that have a side-chain RMSD larger than the relevant control is similar for the different components of each complex (table 5-7). The same is true for Cα displacement (table 5-7), except for the enzyme subtilisin complexed with chymotrypsin inhibitor (complex PDB code = 2sni) and antibody D44.1 bound to lysozyme (complex PDB code = 1mlc). This suggests that in many cases the extent of conformational change is the same in the different components.

Table 5-7 - Number of Interface Residues from Different Types of Component that have Conformational Differences Greater than the Controls

PDB Code of Complex	Measurement			
	Cα Displacement		Side-chain RMSD	
	Number of Interface Residues with Measurement > Control			
	Enzyme	Inhibitor	Enzyme	Inhibitor
1brb	1	0	2	1
1cgi	16	13	11	8
2kai	1	1	4	1
2ptc	0	0	1	0
2sic	0	1	3	2
2sni	0	5	2	2
	Antibody	Antigen	Antibody	Antigen
1mlc	12	2	3	2
1vfb	2	2	5	2

5.7 Differences Between Different Types of Complex

A comparison of the amount of conformational change in equivalent components of different types of complexes was also performed. This could aid predictive docking by giving an idea of how much conformational change to expect for any particular system. Enzymes are comparable with antibodies and inhibitors are comparable with antigens in terms of their relative sizes in the complexes. Also, Janin and Chothia, 1990, found that the two types of complexes have similar levels of conformational change.

As in the previous section, only the conformational changes of interface residues are compared because it has already been seen that the changes of exposed non-interface residues are caused simply by flexibility and disorder (see section 5.3). A comparison of the inhibitors and antigens in our data set (table 3-2) is justified as there are six and seven of each, respectively, that have structures of both the complexed and unbound forms available. The numbers of these that have values above the controls suggest that side-chain movement is more common in the interfaces of inhibitors than in that of antigens. This is shown by both side-chain RMSD (figure 5-1d), where five of the six inhibitors but only three of the seven antigens have values greater than the control, and by the percentage of χ_2's that change minima (figure 5-2d): five of the six inhibitors have values greater than the control for this measurement, but this is the case in only two of the seven antigens. Once again, the differences are caused by large changes of a few residues. However, this does not invalidate the results because of the similar number of residues in the interfaces (table 3-2). There are only two antibodies with both components solved in an unbound form, and so a comparison of them with the enzymes is not justified.

The other complexes, that are not enzyme-inhibitor or antibody-antigen, show mixed results and should be considered individually. Table 3-2 shows that profilin (PDB code 2btf, chain p), in complex with β-actin, has a similar number of residues in its interface when compared to inhibitors and antigens (though at the high end of the range), and figure 5-2d shows that a substantial percentage of the χ_2's of these residues change minima. None of the other overall measures of the movement of this interface are above the controls (Cα RMSD, figure 5-1c, side-chain RMSD, figure 5-1d, and the percentage of χ_1's that change minima, figure 5-2c). Amicyanin (1mda_a) complexed with methylamine dehydrogenase and human growth hormone (3hhr_a) complexed with its receptor both have large changes in their interfaces for all four measures - Cα RMSD (figure 5-1c), side-chain RMSD (figure 5-1d), and the percentages of χ_1's and of χ_2's that change minima (figure 5-2c and figure 5-2d). Amicyanin has a small number of interface residues (see table 3-2), so large changes of a few residues have a greater effect on these measures. Human growth hormone has double the number of interface residues that enzymes and inhibitors have (the receptor is a dimer, and the hormone effectively has two interfaces, one with each monomer). Therefore the large values seen for these measures are definitely significant, but there are also large changes of the whole molecule (Chantalat et al., 1995). The number of interface residues in the interface of subtilisin (1spb_s) complexed with subtilisin prosegment is similar to the number in the growth hormone complex, but in this case only the percentage of χ_2's that change minima is above the control (figure 5-2d). Deoxyribonuclease I (1atn_d) complexed with Actin and Glycerol Kinase (1gla_f) complexed with Glucose Specific Factor III (GSF III) both have little substantial movement in their interfaces, except for the percentage of χ_1's of the interface of GSF III that change minima (figure 5-2c). Thus at least two of the six complexes that are not of enzymes and inhibitors or antibodies and antigens show substantial conformational changes. When more structures of such protein-protein complexes become available, it is possible that they might also show substantial conformational change - it may be a requirement for them to carry out their function.

5.8 Differences of Identical Proteins in Different Complexes

Table 3-3 gives information on five proteins that are present in more than one complex in the main data set (table 3-2). The only difference between comparing i) unbound structures with complexed and ii) complexed with complexed is that the interface may be affected. This follows from the observation that the changes of exposed non-interface residues are caused by flexibility and disorder (see section 5.3), rather than by hinge-bending or shear between domains, as sometimes occurs when proteins bind small molecules (Gerstein et al., 1994). Therefore it is appropriate to concentrate just on those residues that are common to the interface of all the complexes of a particular protein. The Cα displacements and side-chain RMSD's of these residues were examined.

Only one of the proteins, bovine pancreatic trypsin inhibitor (PTI), has overall interface side-chain RMSD's between all structures of that protein in a complex and the unbound form that are larger than the control (see figure 5-1d). These structures have only one common interface residue that changes its conformation by more than the control limits. This residue, Arginine 17, has a much more similar conformation in the complexes than it does in the unbound structure (figure 5-6). The change avoids steric hindrance that would occur with the unbound conformation. It is only in this protein that the interfaces of the complexes appear more similar to each other than to the same region in the unbound structure. Arginine 17 in the unbound structure appears to have been placed in the most common conformation by the crystallographers (Parkin et al., 1996), perhaps suggesting that it is mobile and was poorly defined in the electron density map. However, it has a lower temperature factor than in the complexed structures, which implies that it is actually less mobile than when in the complexed structures and therefore that the differences are genuine or a result of crystal contacts in the unbound form.

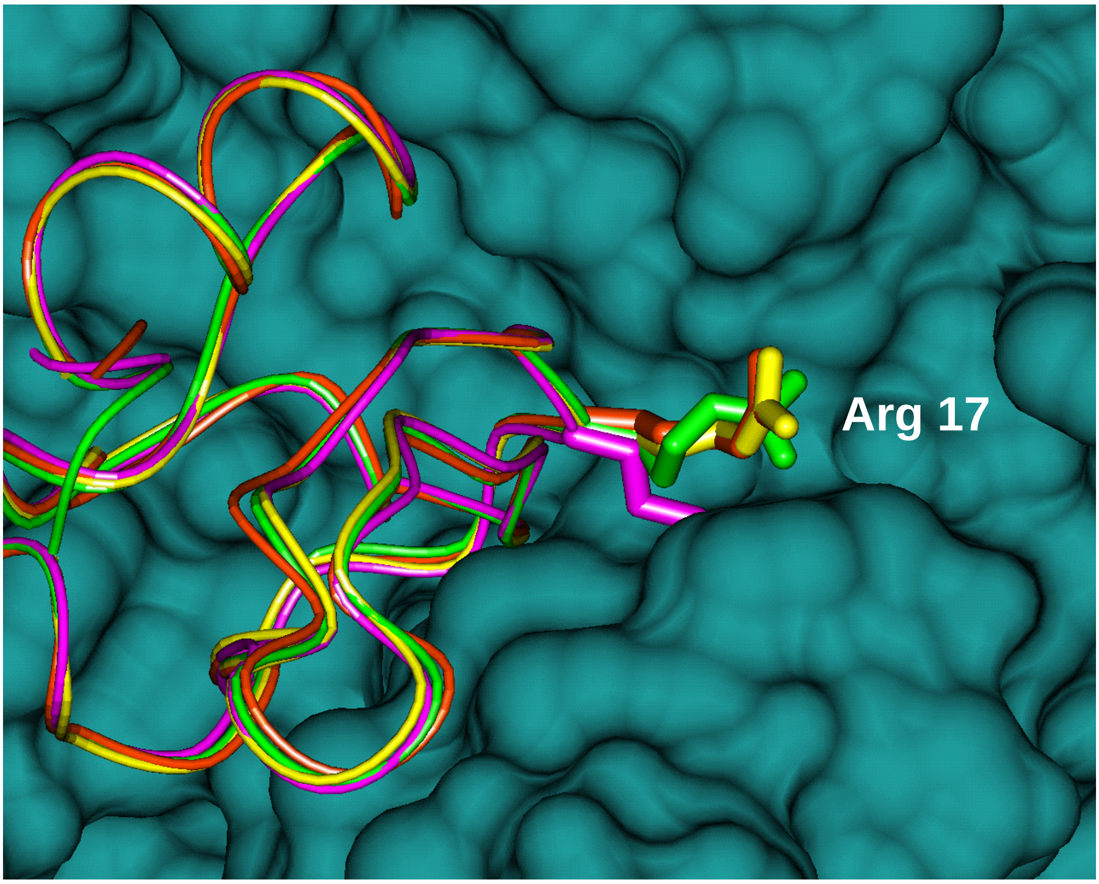
In the subtilisin complexes there are several residues common to the interface that have differences greater than the controls. Histidine 64 in the unbound structure and in the protein bound to subtilisin prosegment has a large side-chain RMSD when compared to the other situations. However, in the unbound structure this residue has two possible positions. The one used in this analysis has an occupancy of 0.8. However, this corresponds to a structure with phenylmethylsulfonate (PMS) bound with an occupancy of 0.7. The 0.2 occupancy structure of histidine 64, with no bound PMS, is much closer to the structures of the complexes with inhibitors, but not to that with prosegment. His64 in the complex with prosegment differs from the others because the bulk of the prosegment binds away from the active site, with only eight residues of the C-terminus extending into the active site. In the other complexes, steric hindrance by the inhibitor, which is different to that caused by PMS, favours the 0.2 occupancy conformation of histidine 64. There are also small differences in the conformations of serine 101 and tyrosine 104, but the conformations in the complexes are not significantly more similar to each other than they are to the unbound conformation. All the other common interface residues have conformations that differ by amounts that are less than the controls.

In all comparisons between the three examples of bovine chymotrypsin (one unbound and two complexed), phenylalanine 39 differs by a large side-chain RMSD (around 5Å). The difference between the two complexed structures is slightly smaller than in comparisons with the unbound, reflecting that the conformational change occurs only after Cβ (i.e. involves a χ_1 rotation), rather than from Cα onwards as is the case in the comparisons with the unbound structure. Tyrosine 146 differs slightly in all comparisons, but is at the end of a chain break. It has already been seen that fragment ends are often more flexible than other parts of structures (section 5.3). Serine 218 is more different in comparisons with one of the complexes than it is in the comparison of the other complex structure with the unbound form. All the other common interface residues have conformations that differ by amounts that are less than the controls.

In the bovine trypsin complexes, the conformations of only one of the common interface residues (tyrosine 39) differ by more than the controls, and in this case the conformations of the complexes are not more similar to each other than they are to that of the unbound. The same residue of rat trypsin differs between the unbound form and the two bound forms, but does not differ between the two bound forms. However, the differences are small (side-chain RMSD's less than 1.1Å).

The data set is limited because it is small and because three of the five proteins are eukaryotic proteases. This means that general conclusions must be made with caution. However, it appears that when the changes in the interface are small, the structures of the interfaces in the complexes are no more similar to each other than they are to the unbound structure. Larger changes are more likely to be common to all complexes, indicating that they may be more important for binding.

Figure 5-6 - A Change Common to Several PTI Complexes
The structure of bovine pancreatic trypsin inhibitor (PTI) in an unbound form (mauve) and in three different complexes (with rat trypsin = yellow, with kallikrein = orange, with bovine β-trypsin = green). The cyan coloured molecular surface is of the kallikrein structure complexed to PTI.



5.9 Conclusions

Conformational changes on complex formation have been evaluated by overall measures of RMSD’s of Cα atoms and of side-chain atoms, and by the percentages of side-chain torsion angles that change minima. In addition, measures of Cα shift and side-chain RMSD’s for individual residues were employed. The main conclusions from this study are given below:

a) A comparison of structural differences between independently solved structures of identical protein provides bench-marks to evaluate conformational change. These bench-marks are an RMSD of 0.6Å and 1.7Å for Cα atoms and for side-chain atoms of exposed residues. Only conformational changes greater than these values were taken as substantial. Shifts for individual residue types were also established. Residues which become part of the interface go from being exposed in the unbound structure to packed, and therefore less mobile, in the complex. Thus using the changes of exposed residues of independently solved structures of identical proteins, which are exposed in both structures, as bench-marks to evaluate the conformational changes of interface residues will overestimate the level above which change should be considered to be substantial. For this reason, protein-protein docking algorithms which are unable to allow for changes up to the level of the bench-marks could well be able to correctly predict the structure of a complex. Movement may also be substantial in more cases than we have suggested. Our analysis is therefore a conservative one.

b) Just over half of the proteins have a substantial shift on complex formation as judged by any of the overall measures. Many of these changes are only just above the benchmark. Thus many heteroprotein complexes are formed without substantial conformational change.

c) Main-chain as well as side-chain atoms can have significant shifts on complex formation.

d) The largest conformational changes in exposed non-interface residues are the consequence of flexibility and disorder rather than a change in conformation caused by, for example, shear or hinge bending between domains on association as occurs on binding small ligands (Gerstein et al., 1994). In contrast, conformational changes in the interface are intimately involved in the complex formation.

e) When account is taken of the different sizes of enzymes and inhibitors, then the extent of conformational change is similar for these two types of components.

f) There are coordinates for bound and unbound forms of both components for eight complexes (six enzyme-inhibitor and two antibody-antigen). All show conformational change in at least one component by at least one of the global measures. In three of the eight complexes (1brb, 2kai, 2ptc), there is only significant global change for the side-chains and no Cα atom moves more than 1.0Å. In the others there are both main-chain and side-chain shifts.

The implications for structure modelling are discussed in the next chapter.

Chapter Six

Implications for Modelling

The aim of this chapter is to address the wider implications of the results shown in the rest of the thesis. Chapter Two presented the development of a protein-protein docking algorithm, and highlighted some of the general problems associated with predicting the structures of complexes. However, this algorithm also had its own peculiarities. Therefore the performance of a more modern docking algorithm (Gabb et al., 1997), which has been tested in a blind trial (Dixon, 1997), is investigated here, with reference to the conformational differences seen in chapters Three, Four, and Five. Before this, these differences are used to evaluate the accuracy of comparative modelling techniques that were also tested in a blind trial (Martin et al., 1997).

6.1 Implications for Structure Modelling

The observed conformational differences between pairs of independently solved structures of identical proteins (table 4-1 and table 4-2) have implications for all attempts at precise modelling of structures, such as comparative modelling and predictive docking. It is unreasonable to expect the models to be accurate to a higher degree than crystal structures. In this chapter, the success of these two modelling techniques is assessed with reference to these control values. The success of predictive docking is also weighed against the amount of conformational change seen between complexed and unbound structures.

Comparative Modelling

Martin et al., 1997 assessed the results of the comparative modelling section of the second Critical Assessment of Structure Prediction (CASP2), held in 1996. An assessment of the importance of any conformational differences was made by comparing with values calculated from three of the targets, whose structures gave two sets of coordinates each. However, these three pairs of structures were not as independently solved as those used in this thesis (see section 3.2.3). Two pairs consisted of different crystal forms solved by the same authors, whilst the other one was made up from two molecules in the asymmetric unit (which were refined independently). These similarities meant that systematic differences in the solution of the structures were likely to be less than in the data set presented in section 3.2.3. However, poorly defined residues were not excluded from the calculations as they were in this thesis (see section 3.2.1). These three pairs each had a Cα RMSD of approximately 0.6Å, which is slightly higher than the value of 0.4Å for all Cα atoms that was presented in table 4-1.

Martin et al., 1997 found that the accuracy of the models submitted to CASP2 was proportional to the similarity of the parent structure to the target structure. With sequence identity of 85% or higher, Cα RMSD’s between the model structure and the target structure were less than 1Å. This means that, overall, these models were only slightly poorer in accuracy than crystal structures. This accuracy decreased at lower identities, though at 26% the Cα RMSD was still as low as 2.2Å for the best models. The major deviations were in loop regions, with local Cα RMSD’s that were 3 to 10Å higher than the global value. These regions also had local sequence identity lower than the global identity. Thus when the sequences were poorly aligned, the more highly conserved ‘core regions’ (Hubbard and Blundell, 1987) were not correctly identified and the whole model suffered as a result.

It was also seen that in those models with Cα RMSD’s less than 1Å, an average of 78.5% of the χ_1 angles were in the correct minima. This shows a lower accuracy than that identified in section 4.1, where 87.1% of χ_1 torsion angles were found to occupy the same minima in pairs of independently solved structures of identical proteins, meaning that accurate side-chain placement was harder than the building of a good structure for the main-chain.

Predictive Docking

The protein-protein docking program FTDOCK (Gabb et al., 1997) was developed and tested on a data set containing five of the complexes analysed in this thesis (table 3-2), using exactly the same structural data for the bound and unbound forms. Thus the effect of the changes identified on FTDOCK’s ability to predict correctly the structure of a protein-protein complex from the unbound structures can be evaluated. The algorithm performs a global rigid-body search of rotational and translational space, and scores each potential structure on shape and electrostatic complementarity. The best 4000 from this search are filtered using distance constraints from biochemical data, and then undergo local refinement scored by shape complementarity, with a higher level of sampling of conformational space than feasible in the global search. A correct structure was defined as one with an interface Cα RMSD of 2.5Å or less when compared to the crystal structure of the complex. The results are given in table 6-1, along with a summary of the conformational changes seen in the interfaces of each component, and are discussed below.

The algorithm performed best on the α-chymotrypsinogen - PTI complex (PDB code 1cgi), with a correct structure (that had a total Cα RMSD of 1.7Å) ranked first out of 133 predictions that remained after local refinement. This is somewhat surprising in the light of our analysis, as the interface regions of the two components show some of the largest Cα and side-chain RMSD’s observed (figure 5-1), and percentages of side-chain angles that change minima that are mostly above the control levels (figure 5-2). This is especially true in the interface. These large values are caused by sizeable movements of several individual interface residues, as discussed in section 5.3.2. However, none of these residues would have caused bad steric clash had they stayed in their unbound conformation. A similar result was given by the Antibody D44.1 - lysozyme complex (PDB code 1mlc), for which a correct structure (that had a total Cα RMSD of 2.0Å) was placed first in a list of 378. The antibody structure has several interface residues that move slightly (side-chain RMSD's ≤ 2.5Å) towards the lysozyme. Arginine 45 of lysozyme moves to avoid clash, with a side-chain RMSD of 5.8Å.

The kallikrein-PTI complex (PDB code 2kai) was predicted less satisfactorily, with a correct structure ranked thirty-third out of 181 that remained after local refinement. Arginine 17 of PTI moves to avoid bad steric clash (see section 5.8 and figure 5-6), with a side-chain RMSD of 5.3Å. Smaller movements of Kallikrein residues Glutamine 41, Tyrosine 99 and Methionine 192 also avoid steric clash in the interface.

A correct structure for the subtilisin-chymotrypsin inhibitor complex (PDB code 2sni) was found second in a list of fifteen possibilities, with only small clash-avoiding conformational changes occurring in the interface.

The final complex, subtilisin - subtilisin inhibitor (PDB code 2sic), had no correct solution in the top 4000 predictions. This is puzzling at first glance. Although both components have some interface residues that show movement above the control, and would cause steric clash if the movements did not occur, these movements are no more severe than those seen in the previous three complexes. However, the unbound structure of subtilisin inhibitor has a region (Ala62 - Met70) where only the approximate path of the main-chain could be traced, with associated uncertainties in the placement of the side-chains (see PDB file for code 2ssi). These residues were therefore excluded from our analysis, but unfortunately some of them are interface residues and would cause substantial steric clash if they remained in their unbound conformations.

These results show that conformational change which does not occur to avoid steric clash can be coped with quite well, even when it is to the level seen in the α-chymotrypsinogen

- PTI complex. There is sufficient shape complementarity to identify the correct complex, despite the large conformational change. Several large clash causing changes are more difficult to deal with.

Table 6-1 - The Effects of Conformational Changes on the Algorithm ‘FTDOCK’

			Conformational Differences in the Interface					
Complex[i]	FTDOCK Results[ii]	Component[iii]	Overall Differences > Controls[iv] ?				Number of Individual Residues with Differences > Controls[v] (Min-Max / Å)	
			Cα RMSD	Side-chain RMSD	$\Delta\chi_1$	$\Delta\chi_2$	ΔCα[vi]	Side-chain RMSD
1cgi	1 / 133	e	✓	✓	✓	✓	16 (0.8-7.3)	11 (1.5-11.5)
		i	✓	✓	✓	✘	13 (0.8-5.0)	8 (0.9-9.7)
2kai	33 / 181	a,b	✘	✘	✘	✘	1 (1.0)	4 (0.8-4.2)
		i	✘	✓	✘	✓	1 (0.8)	1 (5.3)
2sni	2 / 15	e	✘	✘	✘	✘	0	2 (1.4-3.5)
		i	✓	✓	✓	✓	5 (0.9-1.7)	2 (2.3-2.6)
2sic	-	e	✘	✘	✘	✘	0	2 (1.1-3.4)
		i	✓	✘	✓	✓	1 (1.2)	3 (0.9-1.5)
1mlc	1 / 378	a,b	✓	✘	✘	✘	12 (0.8-2.5)	3 (1.4-2.5)
		e	✓	✓	✘	✓	2 (0.9-2.1)	2 (3.0-5.8)

i. Specified by PDB code.

ii. From Gabb et al., 1997. Given by ‘rank / N’, where ‘N’ = the number of predictions after the refinement stage and ‘rank’ = the position of the first correct structure in this list. A correct structure is one where the interface Ca RMSD ≤ 2.5Å.

iii. Specified by the chain identifier(s) in the PDB file of the complex.

iv. See table 4-1 for control values, and table 5-1 for the values for the complexes.

v. See table 4-2 for control values for individual residues.

vi. Cα displacement.

Chapter Seven

Conclusions

This thesis has examined the prediction and analysis of recognition in hetero-protein complexes. Chapter Two presented the development of one particular predictive docking algorithm. This program had problems associated with the representation of surface as a projection onto a plane, with associated loss of information, and restrictions imposed by its intimate ties to a particular type of computer. However, a detailed investigation of its performance highlighted several concerns that will be common to all rigid-body docking methods: measurements of surface complementarity alone were not able to predict correctly the structure of a complex starting from the structures of its components in an unbound form, and conformational differences between the unbound and complexed components complicate matters further. This last effect was reduced by adjustment of the scoring function and by the pruning of side-chains that were likely to be flexible. However, the lack of detailed knowledge of the extent of such conformational differences prompted the work presented in the rest of the thesis.

Chapter Three gave the results of a thorough search of the PDB, and showed that it contained a sufficient number of pairs (39) of good quality complexed and unbound structures from which an analysis of conformational changes on protein-protein association could be made. In addition, twelve pairs of identical protein whose structures were solved independently were found. This was done to provide data on the amount of conformational difference that could be expected from differences in experimental structure determination. Different methods of measuring conformational change were presented, separated into overall change and change of individual residues, with attention to possible ambiguities in the specification of the structures. These methods were applied to the pairs of structures mentioned, and the results given in Chapter Four and Chapter Five.

In Chapter Four the conformational differences in the twelve pairs of independently solved structures of identical proteins were presented. It was seen that exposed regions can be expected to differ by as much as 0.6Å Cα RMSD and 1.7Å simply because of differences in the determination of their structures. Controls were also established for individual residues, based on their amino acid type, and the differences between types were explainable by the differences in their structures. The non-normality of the distributions forced the control values to be higher than may be the case when more structures are available on which the analysis can be performed, though the cut-off used excluded those residues that can be expected to be more flexible than others for reasons such as poor definition in the electron density. These controls were used in Chapter Five to assess the importance of conformational differences between unbound and complexed structures, and it was seen that many heteroprotein complexes are formed without substantial conformational change. In other cases the changes could be in the main-chain as well as the side-chains. Changes of exposed non-interface residues were a consequence of flexibility and disorder rather than domain movements caused by binding.

This thesis confirms the induced-fit model for protein-protein recognition. Often the largest movements are not from the functionally important residues, such as those forming the active sites, but interface regions that are peripheral to these residues. The conformational change can alleviate steric clashes, improve van der Waals packing, or lead to the formation of hydrogen bonds or salt bridges. The program FTDock (Gabb et al., 1997), examined in Chapter Six, was able to predict successfully the structures of complexes that had some of the largest changes seen in Chapter Five. In several of the other systems examined in Chapter Five, the extent of conformational change is not as substantial. For these systems, recognition in shape and charge can, as a first approximation, be treated as a lock and key. Chapter Six also showed that when the sequence identity is high between target and the model, comparative modelling can produce structures accurate almost to the level of the controls.

In the future, the work presented in this thesis could be developed in several ways. The inclusion of more structures with high resolution, as these become available, will improve the measures of conformational change. The cut-offs for structural differences caused by experimental errors will become more robust, and not so dependent on a few structures that may be unusual. In addition, there is still a limited number of protein-protein complexes for which there is information about conformational change. This work would be aided greatly by the availability of the data used to determine the structures, so that disordered and / or flexible regions could be identified more easily. As more structures of complexes and their unbound components are solved, the conclusions from this analysis may need to be revised. In particular the extent of conformational change may vary between the different biological systems. The enzyme-inhibitor complexes that dominate this study may generally exhibit less conformational changes than complex formation involved in other process, such as signalling. The high binding affinity seen in enzyme-inhibitor and antibody-antigen association may rule out large conformational changes, whereas conformational changes of other proteins may be fundamental to their mechanisms. For those systems with limited conformational change, predictive docking should prove a valuable method to obtain structural models from unbound components and thereby provide insights into biological recognition.

References

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM - a new method for protein modelling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.*, 15:488–506.

Abola, E., Bernstein, F., Manning, N., Shea, R., Stampf, D., and Sussman, J. (1996). Protein data bank atomic coordinate entry format description, v2.1. Published online at www.pdb.bnl.gov/pdb-docs/Format.doc/Contents_Guide_21.html.

Ago, H., Kitagawa, Y., Fujishima, A., Matsuura, Y., and Katsube, Y. (1991). Crystal-structure of basic fibroblast growth-factor at 1.6Å resolution. *J. Biochem.*, 110:360–363.

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

Ausiello, G., Cesareni, G., and Helmer-Citterich, M. (1997). ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure. *Proteins*, 28:556–567.

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

Bennett, W. S. and Huber, R. (1984). Structural and functional aspects of domain motions in proteins. *CRC Critical Reviews in Biochemistry*, 15:291–384.

Bhat, T. N., Bentley, G. A., Boulot, G., Green, M. I., Tello, D., Dall'Acqua, W., Souchon, H., Schwarz, F. P., Mariuzza, R. A., and Poljak, R. J. (1994). Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Nat. Acad. Sci.*, 91:1089–1093.

Blevins, R. A. and Tulinsky, A. (1985). The refinement and the structure of the dimer of α-chymotrypsin at 1.67Å resolution. *J. Biol. Chem.*, 260:4264–4275.

Bode, W., Chen, Z. G., Bartels, K., Kutzbach, C., Schmidtkastner, G., and H., B. (1983). Refined 2Å X-ray crystal-structure of porcine pancreatic kallikrein- a, a specific trypsin-like serine proteinase - crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine trypsin. *J. Mol. Biol.*, 164:237–282.

Bode, W., Epp, O., Huber, R., Laskowski junior, M., and Ardelt, W. (1985). The crystal and molecular structure of the third domain of silver pheasant ovomucoid (OMSVP3). *J. Biochem.*, 147:387–395.

Bode, W., Greyling, H. J., Huber, R., Otlewski, J., and Wilusz, T. (1989). The refined 2.0Å X-ray crystal-structure of the complex formed between bovine β-trypsin and CMTI-I, a trypsin-inhibitor from squash seeds (cucurbita-maxima) - topological similarity of the squash seed inhibitors with the carboxypeptidase a inhibitor from potatoes. *FEBS Letters*, 242:285–192.

Bode, W., Papamokos, E., and Musil, D. (1987). The high-resolution X-ray crystal structure of the complex formed between subtilisin carlsberg and eglin-c, an elastase inhibitor from the leech hirudo medicinalis. structural analysis, subtilisin structure and interface geometry. *J. Biochem.*, 166:673–692.

Bolognesi, M., Gatti, G., Menegatti, E., Guarneri, M., Marquart, M., Papamokos, E., and Huber, R. (1982). Three-dimensional structure of the complex between pancreatic secretory inhibitor (kazal type) and trypsinogen at 1.8Å resolution. structure solution, crystallographic refinement and preliminary structural interpretation. *J. Mol. Biol.*, 162:839–868.

Braden, B. C., Souchon, H., Eisele, J. L., Bentley, G. A., Bhat, T. N., Navaza, J., and Poljak, R. J. (1994). Three-dimensional structures of the free and the antigen-complexed Fab form monoclonal anti-lysozyme antibody D44.1. *J. Mol. Biol.*, 243:767–781.

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

Brunger, A. T., Adams, P. D., and Rice, L. M. (1998). Recent developments for the efficient crystallographic refinement of macromolecular structures. *Current Opinion Structural Biol.*, 8:606–611.

Brunger, A. T., Kuriyan, J., and Karplus, M. (1987). Crystallographic R-factor refinement by molecular-dynamics. *Science*, 235:458–460.

Cedergren-Zeppezauer, E. S., Goonesekere, N. C. W., Rozycki, M. D., Myslik, J. C., Dauter, Z., Lindberg, U., and Schutt, C. E. (1994). Crystallization and structure determination of bovine profilin at 2.0Å resolution. *J. Mol. Biol.*, 240:459–475.

Chantalat, L., Jones, N. D., Korber, F., Navaza, J., and Pavlovsky, A. G. (1995). The crystal structure of wild-type growth hormone at 2.5Å resolution. *Protein and*

Peptide Letters, 2:333–340.

Chen, L., Durley, R., Poliks, B. J., Hamada, K., Chen, Z., Mathews, F. S., Davidson, V. L., Satow, Y., Huizinga, E., Vellieux, F. M. D., and Hol, W. G. J. (1992). Crystal structure of an electron-transfer complex between methylamine dehydrogenase and amicyanin. *Biochem.*, 31:4959–4964.

Chen, Z. G. and Bode, W. (1983). Refined 2.5Å X-ray crystal-structure of the complex formed by porcine kallikrein-a and the bovine pancreatic trypsin-inhibitor - crystallization, patterson search, structure determination, refinement, structure and comparison with its components and with the bovine trypsin pancreatic trypsin-inhibitor complex. *J. Mol. Biol.*, 164:283–311.

Cherfils, J., Duquerroy, S., and Janin, J. (1991). Protein-protein recognition analyzed by docking simulation. *Proteins*, 11:271–280.

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

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

Cohen, G. H., Sheriff, S., and Davies, D. R. (1996). The refined structure of the monoclonal antibody HyHel-5 with its antigen hen egg white lysozyme. *Acta Cryst.*, D52:315–326.

Connolly, M. L. (1983). Analytical molecular surface calculation. *J. App. Cryst.*, 16:548–558.

Cruickshank, D. W. J. (1996). Protein precision re-examined: Luzzati plots do not estimate final errors. In *Macromolecular Refinement*, volume DL–CONF–96–001, pages 11–22. CCP4.

Daopin, S. and Davies, D. R. (1994). Comparison of two crystal structures of TGF-β 2: the accuracy of refined protein structures. *Acta Cryst.*, D50:85–92.

Daopin, S., Piez, K. A., Ogawa, Y., and Davies, D. R. (1992). Crystal structure of transforming growth factor-β2: an unusual fold for the superfamily. *Science*, 257:369–373.

Davies, J. F., Almassy, R. J., Hostomska, Z., Ferre, R. A., and Hostomsky, Z. (1994). 2.3Å crystal structure of the catalytic domain of DNA polymerase β. *Cell*, 76:1123–1133.

de Vos, A. M., Ultsch, M., and Kossiakoff, A. A. (1992). Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science*, 255:306–312.

Derrick, J. P. and Wigley, D. B. (1994). The third IGG-binding domain from streptococcal protein G - an analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.*, 243:906–918.

Diamond, R. (1974). Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.*, 82:371–391.

Dixon, J. S. (1997). Evaluation of the CASP2 docking section. *Proteins*, S1:198–204.

Dodson, E. J., Davies, G. J., Lamzin, V. S., Murshudov, G. N., and Wilson, K. S. (1998). Validation tools: can they indicate the information content of macromolecular crystal structures? *Structure*, 6:685–690.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G., and D., M. (1989). RESTRAIN - restrained structure factor least-squares refinement program for macromolecular structures. *J. App. Cryst.*, 22:510–516.

Duncan, B. S. and Olson, A. J. (1993). Shape analysis of molecular surfaces. *Biopolymers*, 33:231–238.

Durley, R., Chen, L., Lim, L. W., Mathews, F. S., and Davidson, V. L. (1993). Crystal structure analysis of amicyanin and apoamicyanin from paracoccus denitrificans. *Prot. Sci.*, 2:739–752.

Engh, R. A. and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst.*, A47:392–400.

Fischmann, T. O., Bentley, G. A., Bhat, T. N., Boulot, G., Mariuzza, R. A., Phillips, S. E. V., Tello, D., and Poljak, R. J. (1991). Crystallographic refinement of the three-dimensional structure of the Fab D1.3-lysozyme complex at 2.5Å resolution. *J. Biol. Chem.*, 266:12915–12920.

Fitzpatrick, P. A., Ringe, D., and Klibanov, A. M. (1994). X-ray crystal structure of cross-linked subtilisn carlsberg in water vs. acetonitrile. *Biochem. Biophys. Res. Comm.*, 198:675–681.

Flores, T. P., Orengo, C. A., Moss, D. S., and Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Prot. Sci.*, 2:1811–1826.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T., and Xuong, N. H. (1970). Chymotrypsinogen, 2.5Å crystal structure, comparison with α-chymotrypsin, and implications for zymogen activation. *Biochem.*, 9:1997–0000.

Frigerio, F., Coda, A., Pugliese, L., Lionetti, C., Menegatti, E., Amiconi, G., Schnebli, H. P., Ascenzi, P., and Bolognesi, M. (1992). Crystal and molecular structure of the bovine α-chymotrypsin-eglin c complex at 2.0Å resolution. *J. Mol. Biol.*, 225:107–123.

Fujinaga, M., Sielecki, A. R., Read, R. J., Ardelt, W., Laskowski, M., and James, M. N. G. (1987). Crystal and molecular structures of the complex of α-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8Å resolution. *J. Mol. Biol.*, 195:397–418.

Gabb, H. A., Jackson, R. M., and Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, 272:106–120.

Gallagher, D. T., Oliver, J. D., Bott, R., Betzel, C., and Gilliland, G. L. (TBP). Subtilisin BPN at 1.6Å resolution: analysis of discrete disorder and comparison of crystal forms. *To be published*, 0:0–0.

Gallagher, T., Gilliland, G., Wang, L., and Bryan, P. (1995). The prosegment-subtilisin BPN complex - crystal-structure of a specific foldase. *Structure*, 3:907–914.

Garrett, T. P. J., Wang, J., Yan, Y., Liu, J., and Harrison, S. C. (1993). Refinement and analysis of the structure of the first two domains of human cd4. *J. Mol. Biol.*, 234:763–778.

Gellatly, B. J. and Finney, J. L. (1982). Calculation of protein volumes: an alternative to the Voronoi procedure. *J. Mol. Biol.*, 161:305–322.

Gerstein, M. and Chothia, C. (1991). Analysis of protein loop closure. Two types of hinge produce one motion in lactate dehydrogenase. *J. Mol. Biol.*, 220:133–149.

Gerstein, M., Lesk, A. M., and Chothia, C. (1994). Structural mechanisms for domain movement in proteins. *Biochem.*, 33:6739–6749.

Gigant, B., Fleury, D., Bizebard, T., J., S. J., and Knossow, M. (1995). Crystallisation and preliminary X-ray diffraction studies of complexes between an influenza hemagluttinin and Fab fragments of two different monoclonal antibodies. *Proteins*, 23:115–117.

Gilliland, G. L., Winborne, E. L., Nachman, J., and Wlodawer, A. (1990). The three-dimensional structure of recombinant bovine chymosin at 2.3Å resolution. *Proteins*, 8:82–101.

Gros, P., Betzel, C., Dauter, Z., Wilson, K. S., and Hol, W. G. J. (1989). Molecular dynamics refinement of a thermitase-eglin-c complex at 1.98Å resolution and comparison of two crystal forms that differ in calcium content. *J. Mol. Biol.*, 210:347–367.

Harata, K. (1993). X-ray structure of monoclinic turkey egg lysozyme at 1.3Å resolution. *Acta Cryst.*, D49:497–504.

Hartsuck, J. A., Koelsch, G., and Remington, S. J. (1992). The high resolution crystal structure of porcine pepsinogen. *Proteins*, 13:1–25.

Hecht, H. J., Szardenings, M., Collins, J., and Schomburg, D. (1991). Three-dimensional structure of the complexes between bovine chymotrypsinogen A and two recombinant variants of human pancreatic secretory trypsin inhibitor (Kazal-type). *J. Mol. Biol.*, 220:711–722.

Hecht, H. J., Szardenings, M., Collins, J., and Schomburg, D. (1992). Three-dimensional structure of a recombinant variant of human pancreatic secretory trypsin inhibitor (Kazal type). *J. Mol. Biol.*, 225:1095–1103.

Heinz, D. W., Priestle, J. P., Rahuel, J., Wilson, K. S., and Grutter, M. G. (1991). Refined crystal-structures of subtilisin novo in complex with wild-type and 2 mutant eglins - comparison with other serine proteinase-inhibitor complexes. *J. Mol. Biol.*, 217:353–371.

Helmer-Citterich, M. and Tramontano, A. (1994). PUZZLE: a new method for automated protein docking based on surface shape complementarity. *J. Mol. Biol.*, 235:1021–1031.

Honig, B. and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149.

Hubbard, T. J. P. and Blundell, T. L. (1987). Comparison of solvent-inaccesible cores of homologous proteins: Definitions useful for protein modelling. *Prot. Eng.*, 1:159–171.

Huber, R. (1979). Conformational flexibility and its functional significance in some protein molecules. *Trends Biochem. Sci.*, 4:271–276.

Hurley, J. H., Faber, H. R., Worthylake, D., Meadow, N. D., Roseman, S., Pettigrew, D. W., and Remington, S. J. (1993). Structure of the regulatory complex of escherichia coli III(GLC) with glycerol kinase. *Science*, 259:673–677.

Jackson, R. M., Gabb, H. A., and Sternberg, M. J. E. (1998). Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J. Mol. Biol.*, 276:265–285.

Jackson, R. M. and Sternberg, M. J. E. (1995). A continuum model for protein-protein interactions: Application to the docking problem. *J. Mol. Biol.*, 250:258–275.

Janin, J. and Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.*, 265:16027–16030.

Janin, J., Miller, S., and Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204:155–164.

Janin, J., Wodak, S., Levitt, M., and Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol.*, 125:357–386.

Janin, J. and Wodak, S. J. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol.*, 42:21–78.

Jia, Z., Quail, J. W., Waygood, E. B., and Delbaere, L. T. J. (1993). The 2.0Å resolution structure of escherichia coli histidine-containing phosphocarrier protein HPR: a redetermination. *J. Biol. Chem.*, 30:22490–22501.

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Nat. Acad. Sci.*, 93:13–20.

Jones, S. and Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, 272:121–132.

Jones, S. and Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, 272:133–143.

Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F., and Holmes, K. C. (1990). Atomic structure of the actin-DNAase I complex. *Nature*, 347:37–44.

Katchalski-Katzir, E., Sharriv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Nat. Acad. Sci.*, 89:2195–2199.

Katz, B. A., Finer-moore, J., Mortezaei, R., Rich, D. H., and Stroud, R. M. (1995). Episelection: Novel ki nanomolar inhibitors of serine proteases selected by binding or chemistry on an enzyme surface. *Biochem.*, 34:8264–8280.

Ke, H. M. (1997). Overview of isomorphous replacement phasing. *Meth. Enz.*, 276:448–461.

Kishan, R. K. V., Chandra, N. R., Sudarsanakumar, C., Suguna, K., and Vijayan, M.

(1995). Water dependent domain motion and flexibility in ribonuclease-a and the invariant features in its hydration shell. an X-ray study of two low humiditycrystal forms of the enzyme. *Acta Cryst.*, D51:703–710.

Konnert, J. H. and Hendrickson, W. A. (1980). A restrained-parameter thermal-factor refinement procedure. *Acta Cryst.*, A36:344–350.

Kurinov, I. and Harrison, R. W. (1995). The influence of temperature on lysozyme crystals. Structure and dynamics of protein and water. *Acta Cryst.*, D51:98–109.

Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J. R., Whittle, P. J., Danley, D. E., Geoghegan, K. F., Hawrylik, S. J., Lee, S. E., Scheld, K. G., and Hobart, P. M. (1989). X-ray analysis of HIV-1 proteinase at 2.7Å resolution confirms structural homology among retroviral enzymes. *Nature*, 342:299–0000.

Lawrence, M. C. and Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J. Mol. Biol.*, 234:946–950.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–400.

Lesk, A. M. and Chothia, C. (1988). Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature*, 335:188–190.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257:342–358.

Maenaka, K., Matsushima, M., Song, H., Sunada, F., Watanabe, K., and Kumagai, I. (1995). Dissection of protein-carbohydrate interactions in mutant hen egg-white lysozyme complexes and their hydrolytic activity. *J. Mol. Biol.*, 247:281–293.

Malby, R. L., Tulip, W. R., Harley, V. R., Mckimm-breschkin, J. L., Laver, W. G., Webster, R. G., and Colman, P. M. (1994). The structure of a complex between the NC10 antibody and influenza virus neuraminidase and comparison with the overlapping binding site of the NC41 antibody. *Structure*, 2:733–746.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W., and Huber, R. (1983). The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Cryst.*, B39:480–490.

Martin, A. C. R., MacArthur, M. W., and Thornton, J. M. (1997). Assessment of comparative modeling in CASP2. *Proteins*, S1:14–28.

McLachlan, A. D. (1979). Gene duplication in the structural evolution of chymotrypsin.

J. Mol. Biol., 128:49–79.

McPhalen, C. A. and James, M. N. G. (1987). Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochem.*, 26:261–269.

McPhalen, C. A. and James, M. N. G. (1988). Structural comparison of two serine proteinase-protein inhibitor complexes. eglin-c-subtilisin carlsberg and CI-2-subtilisin novo. *Biochem.*, 27:6582–6598.

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

Murzin, A., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540.

Neidhart, D. J. and Petsko, G. A. (1988). The refined crystal structure of subtilisin carlsberg at 2.5Å resolution. *Prot. Eng.*, 2:271–276.

Newman, M., Safro, M., Frazao, C., Khan, G., Zdanov, A., Tickle, I. J., and Blundell, T. L. (1991). X-ray analyses of aspartic proteinases. Structure and refinement at 2.2Å resolution of bovine chymosin. *J. Mol. Biol.*, 221:1295–1309.

Nicholls, A. and Honig, B. (1991). A rapid finite difference algorithm, utilizing successive over-relaxation to solve the poisson-boltzmann equation. *J. Comp. Chem.*, 12:435–445.

Nicholls, A., Sharp, K. A., and Honig, B. (1991). Protein folding and association - insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, 11:281–296.

Oefner, C. and Suck, D. (1986). Crystallographic refinement and structure of DNAse-I at 2Å resolution. *J. Mol. Biol.*, 192:605–632.

Ogata, M. (1998). MAD phasing grows up. *Nature Structural Biology*, 5(SS):638–640.

Padlan, E. A., Silverton, E. W., Sheriff, S., Cohen, G. H., Smith-Gill, S. J., and Davies, D. R. (1989). Structure of an antibody-antigen complex. Crystal structure of the HyHel-10 Fab-lysozyme complex. *Proc. Nat. Acad. Sci.*, 86:5938–5942.

Parkin, S., Rupp, B., and Hope, H. (1996). The structure of bovine pancreatic trypsin inhibitor at 125K: definition of carboxyl-terminalresidues Gly57 and Ala58. *Acta Cryst.*, D52:18–29.

Parsons, M. R. and Phillips, S. E. V. (TBP). The three dimensional structure of turkey egg white lysozyme at 2.2Å resolution. *To be published*, 0:0–0.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, 271:511–523.

Perona, J. J., Tsu, C. A., Craik, C. S., and Fletterick, R. J. (1993). Crystal structures of rat anionic trypsin complexed with the protein inhibitors APPI and BPTI. *J. Mol. Biol.*, 230:919–933.

Pickersgill, R. W., Harris, G. W., and Garman, E. (1992). Structure of monoclinic papain at 1.60Å resolution. *Acta Cryst.*, B48:59–67.

Pickett, S. D. and Sternberg, M. J. E. (1993). Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.*, 231:825–839.

Prasad, L., Sharma, S., Vandonselaar, M., Quail, J. W., Lee, J. S., Waygood, E. B., Wilson, K. S., Dauter, Z., and Delbaere, L. T. J. (1993). Evaluation of mutagenesis for epitope mapping: structure of an antibody-protein antigen complex. *J. Biol. Chem.*, 268:10705–10708.

Priestle, J. P., Schaer, H. P., and Gruetter, M. G. (1989). Crystallographic refinement of interleukin-1 β at 2.0Å resolution. *Proc. Nat. Acad. Sci.*, 86:9667–9671.

Richards, F. M. (1977). Areas, volumes, packing and protein structure. *Ann. Rev. Biophys. Bioengineering*, 6:151–176.

Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution. *J. Mol. Biol.*, 279:1211–1227.

Russell, R. B. and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison. *Proteins*, 14:309–323.

Russell, R. B., Sasieni, P. D., and Sternberg, M. J. E. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.*, 282:903–918.

Rydel, T. J., Tulinsky, A., Bode, W., and Huber, R. (1991). The refined structure of the hirudin-thrombin complex. *J. Mol. Biol.*, 221:583–601.

Rydel, T. J., Yin, M., Padmanabhan, K. P., Blakenship, D. T., Cardin, A. D., Correa, P. E., Fenton II, J. W., and Tulinsky, A. (1994). Crystallographic structure of human γ-thrombin. *J. Biol. Chem.*, 269:22000–22006.

Ryu, S. E., Truneh, A., Sweet, R. W., and Hendrickson, W. A. (1994). Structures of an HIV and MHC binding fragment from human cd4 as refined in two crystal lattices. *Structure*, 2:59–74.

Satow, Y., Y., W., and Mitsui, Y. (1980). Solvent accessibility and microenvironment in a bacterial protein proteinase inhibitor SSI (streptomyces subtilisin inhibitor). *J. Biochem.*, 88:1739–0000.

Savage, H. and Wlodawer, A. (1986). Determinination of water structure around biomolecules using X-ray and neutron diffraction methods. *Meth. Enz.*, 127:162–183.

Savva, R., Mcauley-hecht, K., Brown, T., and Pearl, L. H. (1995). The structural basis of specific base excision repair by uracil-DNA glycosylase. *Nature*, 373:487–493.

Savva, R. and Pearl, L. H. (1995). Nucleotide mimicry in the crystal structure of the uracil-DNA glycosylase - uracil glycosylase inhibitor protein complex. *Nature Structural Biology*, 2:752–757.

Sawaya, M. R., Pelletier, H., Kumar, A., Wilson, S. H., and Kraut, J. (1994). Crystal-structure of rat DNA-polymerase-β - evidence for a common polymerase mechanism. *Science*, 264:1930–1935.

Schlunegger, M. P. and Gruetter, M. G. (1992). An unusual feature revealed by the crystal structure at 2.2Å resolution of human transforming growth factor-β2. *Nature*, 358:430–434.

Schutt, C. E., Myslik, J. C., Rozycki, M. D., Goonesekere, N. C. W., and Lindberg, U. (1993). The structure of crystalline profilin-β-actin. *Nature*, 365:810–816.

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C., and Davies, D. R. (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Nat. Acad. Sci.*, 84:8075–8079.

Shoichet, B. K. and Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.*, 221:327–346.

Sielecki, A. R., Fujinaga, M., Read, R. J., and James, M. N. G. (1991). Refined structure of porcine pepsinogen at 1.8Å resolution. *J. Mol. Biol.*, 219:671–692.

Spinelli, S., Liu, Q., Alzari, P. M., Hirel, P. H., and Poljak, R. J. (1991). The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie*, 73:1391–0000.

Stanfield, R. L. and Wilson, I. A. (1994). Antigen-induced conformational changes in antibodies: a problem for structural prediction and design. *Trends Biotech.*, 12:275–279.

Sternberg, M. J. E., Gabb, H. A., and Jackson, R. M. (1998). Predictive docking of protein-protein and protein-DNA complexes. *Current Opinion Structural Biol.*, 8:250–256.

Stowell, M. H. B., Miyazawa, A., and Unwin, N. (1998). Macromolecular structure determination by electron microscopy: New advances and recent results. *Current Opinion Structural Biol.*, 8:595–600.

Strynadka, N. C. J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M., and James, M. N. G. (1996). Molecular docking programs successfully predict the binding of β-lactamase inhibitory protein to tem-1 β-lactamase. *Nature Structural Biology*, 3:233–239.

Stubbs, M. T., Laber, B., Bode, W., Huber, R., Jerala, R., Lenarcic, B., and Turk, V. (1990). The refined 2.4Å X-ray crystal structure of recombinant human stefin b in complex with the cysteine proteinase papain: a novel type of proteinase inhibitor interaction. *EMBO J.*, 9:1939–1947.

Takeuchi, Y., Satow, Y., Nakamura, K. T., and Mitsui, Y. (1991). Refined crystal structure of the complex of subtilisin BPN and streptomyces subtilisin inhibitor at 1.8Å resolution. *J. Mol. Biol.*, 221:309–325.

Teplyakov, A. V., Kuranova, I. P., Harutyunyan, E. H., Vainshtein, B. K., Froemmel, C., Hoehne, W. E., and Wilson, K. S. (1990). Crystal structure of thermitase at 1.4Å resolution. *J. Mol. Biol.*, 214:261–279.

Thornton, J. M. and Sibanda, B. L. (1983). Amino and carboxy-terminal regions in globular proteins. *J. Mol. Biol.*, 167:443–460.

Tickle, I. J., Laskowski, R. A., and Moss, D. S. (1998). Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of γ B and B2 crystallin. *Acta Cryst.*, D54:243–252.

Tilton, R. F., Dewan, J. C., and Petsko, G. A. (1992). Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-a at nine different temperatures from 98 to 320K. *Biochem.*, 31:2469–2481.

Totrov, M. and Abagyan, R. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6Å accuracy. *Nature Structural Biology*, 1:259–263.

Tronrud, D. E., Teneyck, L. F., and Matthews, B. W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Cryst.*, A43:489–501.

Tsunogae, Y., Tanaka, I., Yamane, T., Kikkawa, J. I., Ashida, T., Ishikawa, C., Watanabe, K., Nakamura, S., and Takahashi, K. (1986). Structure of the trypsin-binding domain of bowman-birk type protease inhibitor and its interaction with trypsin. *J. Biochem.*, 100:1637–1646.

Tulip, W. R., Varghese, J. N., Webster, R. G., Laver, W. G., and Colman, P. M. (1992). Crystal-structures of 2 mutant neuraminidase antibody complexes with amino-acid substitutions in the interface. *J. Mol. Biol.*, 227:149–159.

Turkenburg, J. P. and Dodson, E. J. (1996). Modern developments in molecular replacement. *Current Opinion Structural Biol.*, 6:604–610.

Vakser, I. A. (1995). Protein docking for low-resolution structures. *Prot. Eng.*, 8:371–377.

Vakser, I. A. and Aflalo, C. (1994). Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins*, 20:320–329.

Varghese, J. N., Epa, V. C., and Colman, P. M. (1995). The three dimensional structure of the complex of 4-guanidino-neusacen and influenza virus neuraminidase. *Prot. Sci.*, 4:1081–1087.

Veerapandian, B., Gilliland, G. L., Raag, R., Svensson, A. L., Masui, Y., Hirai, Y., and Poulos, T. L. (1992). Functional implications of interleukin-1β based on the three-dimensional structure. *Proteins*, 12:10–23.

Walls, P. H. and Sternberg, M. J. E. (1992). New algorithm to model protein-protein recognition based on shape complementarity. Applications to antibody-antigen docking. *J. Mol. Biol.*, 228:277–297.

Walter, J., Steigemann, W., Singh, T. P., Bartunik, H., Bode, W., and Huber, R. (1982). On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography. *Acta Cryst.*, B38:1462–1472.

Walter, M. R., Cook, W. J., Zhao, B. G., Cameronjunior, R., Ealick, S. E., Walter, R. L., Reichert, P., Nagabhushan, T. L., Trotta, P. P., and Bugg, C. E. (1992). Crystal structure of recombinant human interleukin-4. *J. Biol. Chem.*, 267:20371–20376.

Webster, D. M. and Rees, A. R. (1993). Macromolecular recognition: Antibody-antigen complexes. *Prot. Eng.*, 6(SS):94.

Weng, Z., Vajda, S., and Delisi, C. (1996). Prediction of protein complexes using empirical free energy functions. *Prot. Sci.*, 5:614–626.

Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R. W. W., Vriend, G., Thornton, J. M., Laskowski, R. A., MacArthur, M. W., Dodson, E. J., Murshudov, G., Oldfield, T. J., Kaptein, R., and Rullmann, J. A. C. (1998). Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *J. Mol. Biol.*, 276:417–436.

Wlodawer, A., Pavlovsky, A., and Gustchina, A. (1992). Crystal structure of human recombinant interleukin-4 at 2.25Å resolution. *FEBS Letters*, 309:59–0.

Worthylake, D., Meadow, N. D., Roseman, S., Liao, D. I., Herzberg, O., and Remington, S. J. (1991). 3-dimensional structure of the escherichia-coli phosphocarrier protein-IIIGLC. *Proc. Nat. Acad. Sci.*, 88:10382–10386.

Wuthrich, K. (1995). NMR - this other method for protein and nucleic acid structure determination. *Acta Cryst.*, D51:249–270.

Xu, D., Tsai, C.-J., and Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Prot. Eng.*, 10:999–1012.

Ysern, X., Li, H., and Mariuzza, R. A. (1998). Imperfect interfaces. *Nature Structural Biology*, 5:412–414.

Zhu, X., Komiya, H., Chirino, A., Faham, S., Fox, G. M., Arakawa, T., Hsu, B. T., and Rees, D. C. (1991). Three-dimensional structures of acidic and basic fibroblast growth factors. *Science*, 251:90–93.