REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2005 Name of Author MOONTI, Q.

Computational Modelling of Protein/Protein and Protein/DNA Docking

GIDON MOONT

2005

Biomolecular Modelling Laboratory
Cancer Research UK
London Research Institute
44 Lincoln's Inn Fields, London, WC2A 3PX

and

Department of Biochemistry and Molecular Biology
University College London
Gower Street, London, WC2E 6BT

Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy in Biochemistry of the University of London.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

The docking problem is to start with unbound conformations for the components of a complex, and computationally model a near-native structure for the complex. This thesis describes work in developing computer programs to tackle both protein/protein and protein/DNA docking.

Empirical pair potential functions are generated from datasets of residue/residue interactions. A scoring function was parameterised and then used to screen possible complexes, generated by the global search computer algorithm FTDOCK using shape complementarity and electrostatics, for 9 systems. A correct docking (RMSD $\leq$ 2.5Å) is placed within the top 12% of the pair potential score ranked complexes for all systems.

The computer software FTDOCK is modified for the docking of proteins to DNA, starting from the unbound protein and DNA coordinates modelled computationally. Complexes are then ranked by protein/DNA pair potentials derived from a database of 20 protein/DNA complexes. A correct docking (at least 65% of correct contacts) was identified at rank $\leq$ 4 for 3 of the 8 complexes. This improved to 4 out of 8 when the complexes were filtered using experimental data defining the DNA footprint.

The FTDOCK program was rewritten, and improved pair potential functions were developed from a set of non-homologous protein/protein interfaces. The algorithms were tested on a non-homologous set of 18 protein/protein complexes, starting with unbound conformations. Using cross-validated pair potential functions and the energy minimisation software MULTIDOCK, a correct docking (RMSD of C_α interface $\leq$ 7Å and $\geq$ 25% correct contacts) is found in the top 10 ranks in 6 out of 18 systems.

The current best computational docking algorithms are discussed, and strategies for improvement are suggested.

Acknowledgements

I would like to thank the following colleagues with whom I worked: Dr Henry Gabb (FTDOCK); Dr Richard Jackson (MULTIDOCK); Dr Patrick Aloy (DNA docking) and Dr Mathew Betts (conformational changes on docking). I would also like to thank Dr Suhail Islam for the graphics generated by PREPI, Dr Marcel Turcotte for programming help, Dr Graham Smith for useful discussion towards the end of this work, and to Professor Michael Sternberg for being a good supervisor with plenty of patience. I would also like to thank the Imperial Cancer Research Fund for funding me.

Abbreviations used

BMM	**Biom**olecular **M**odelling Group
CMM	**C**entre of **M**ass **M**ovement
ICRF	**I**mperial **C**ancer **R**esearch **F**und
MRSA	**M**inimum **R**elative **S**urface **A**rea
MS	**M**olecular **S**urface
PCP	**P**ercentage **C**orrect **P**airs
PDB	**P**rotein **D**ata **B**ank
RMSD	**R**oot **M**ean **S**quare **D**eviation - always measured in Ångstroms
SAS	**S**olvent **A**ccessible**S**urface
SCOP	**S**tructural **C**lassification **o**f **P**roteins
%CC	**%** (percentage) **C**orrect **C**ontacts

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

1.1.1 The need for protein/protein and protein/DNA docking

A full description of many biological processes requires knowledge of the three-dimensional structure of macromolecular complexes. Such structural information provides insights into specificity and so can suggest lead compounds for the development of novel pharmaceutical agents. However, structural studies by crystallography and NMR often follow a divide and conquer approach, so that the coordinates of the component molecules are available but the conformation of the complex is unknown. Indeed in the protein data bank (PDB)[1] there is a large discrepancy between the number of determined protein structures (c. 20 thousand) and the number of protein/protein and protein/DNA complexes (c. 1000). Computer algorithms are therefore needed to predict the structure of macromolecular complexes starting from coordinates of the unbound components. This chapter describes the current computational strategies that can be employed to solve this problem, and this thesis describes the work on one strategy from 1996 to 2001.

The crystal structures of protein/protein and protein/DNA complexes have provided detailed descriptions of the interactions that lead to the specificity that is central to the biological activity of those systems, such as mechanisms that lead to disease. Currently we have details of a variety of protein/protein complexes including enzyme/inhibitor, antibody/antigen, hormone/receptor and cell surface/cell surface proteins.[2–5]

The nature of the inter-protein recognition can be understood in terms of a general shape complementarity, with specificity provided by particular spatial constraints from close packing, hydrogen bonds, and charge-charge ion pairings. Given the full structural information, one could start to alter one of the proteins to affect its recognition properties. For example, the details of a given protein/protein complex might suggest that a particular loop is central to the

interaction. This allows towards the design of lead compounds that could yield novel drugs. Indeed, there are several lead compounds that have been designed based on the structure of a protein receptor interacting with a small molecule ligand (for examples see the 1994 review by Colman[6]). With the increasing number of determined protein structures, the determination of protein/protein complexes by both experiment and modelling should lead to suggestions of new pharmaceutical agents.

1.1.2 Overview of the computational approach

This chapter will describe various computational strategies used to predict the structure of protein/protein complexes. The consensus approach to protein/protein docking is described in Figure 1.1. The precise order of implementing the steps can differ, but essentially the key features are:

- Start with the three-dimensional structures of two unbound components. Consider systems that are expected to have limited conformational change on association.
- The rigid-body approximation is then used. This is that one can simulate a docking starting with unbound components given a limited conformational change.
- A search is performed over all of the rotational and translational space that could allow association between the two components. All of this space must be considered (global scan) when there is no biological information about which parts of the molecules interact. If however there are some constraints, these can be used to limit the space of the initial search (targeted scan). Depending on the algorithm used, constraint information may instead be used as a post search filter. The search will sample the space in a discrete manner, and consequently there will be a lower limit on the difference in conformations between two model complexes that determines the resolution of the search procedure.
- A function is developed to score the quality of each model complex, often using simplified terms to evaluate shape and electrostatic complementarity. There are two reasons for simple functions at this stage. First, as a large number of different models are generated in a global scan, the scoring function must be computationally fast to evaluate. Secondly, one requires a soft function that is not unduly sensitive to the conformational differences between the model complexes formed from unbound components and the true complex with its bound components.

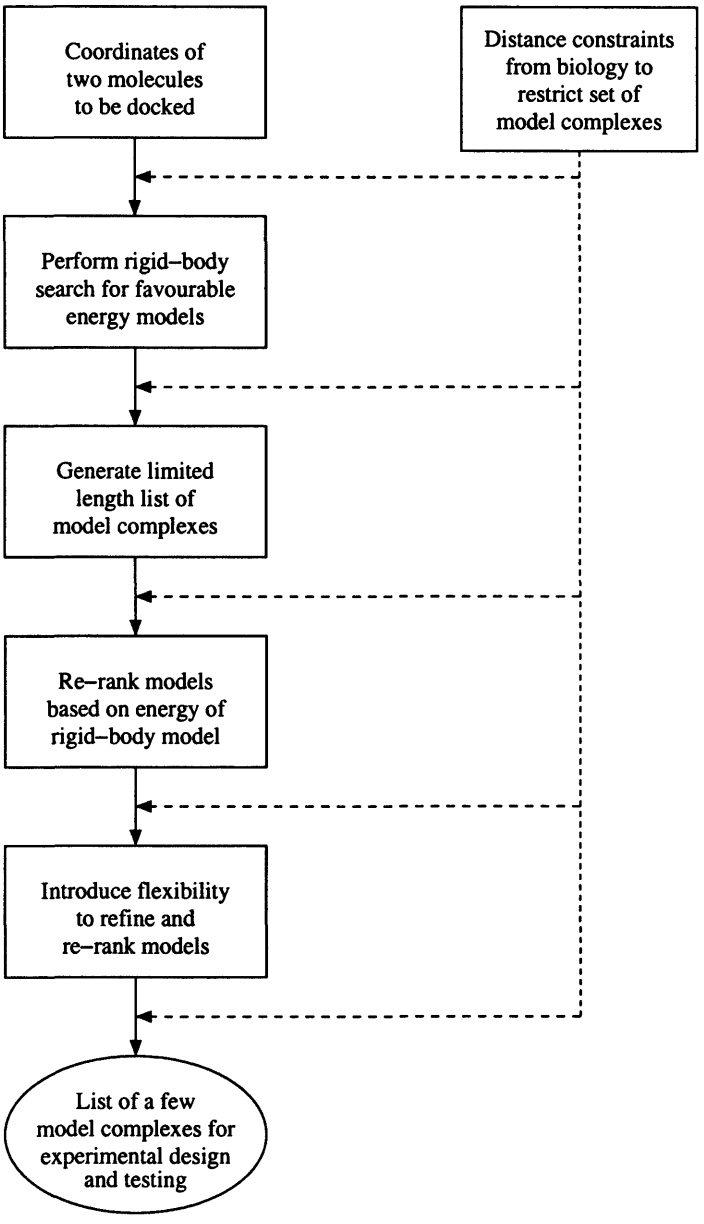


Figure 1.1: Consensus Approach to Computational Docking.

- Ideally, the docking algorithm would thereby identify a single model complex that is a close approximation to the true complex, based on this complex having the best score (*i.e.* the lowest energy model). In practice, the current status of the algorithms is that they generate a limited list of model complexes ranked on the score, and the objective is that one member of a short list should be close to the true complex.
- At this stage, a re-ranking of the model complexes can be done, possibly using more computationally intensive calculations.
- Conformational flexibility is generally introduced into the algorithm when there is only a limited number of models to consider. Perturbations to the structure of the model complex are made and the energy of the resultant conformation is evaluated. The aims are both to improve the structural quality of the model and to improve the power of identifying the best model from the list of false models that have been generated.

1.1.3 Scope of this thesis

This thesis covers work done mainly between 1996 and 2001. This introductory chapter therefore includes reference to work done after this time. This includes the algorithms that have appeared in the wider community, and the Critical Assessment of PRotein Interactions (CAPRI) trials.

Given the reliance of most methods on starting by a rigid-body docking, the next section in this chapter will describe the extent of conformational change on protein/protein association. It will be shown that for many systems, the change is sufficiently limited to suggest that the rigid-body approach is viable.

The following section of this chapter then describes the algorithms and software in the Biomolecular Modelling (BMM) group at the Imperial Cancer Research Fund (ICRF) (now Cancer Research UK) at the end of 1996. The following chapters will describe the subsequent additions and modifications that constitute the work that this thesis is reporting on.

The following section is a review of the various computational strategies that are currently around (mid 2004) and their reported success. This is not an exhaustive list of every protein/protein docking study, but it does cover the broad categories of approach to the problem.

A good test of any predictive method is its success in blind trials, of which there have been several for protein/protein docking. This chapter concludes with a brief report on the entries and results of those trials.

Chapter 2 covers the work which was reported in the 1999 paper **"Use of Pair Potentials Across Protein Interfaces in Screening Predicted Docked Complexes"**.[7] It was the first use of pair potentials to evaluate the quality of

possible models resulting from a docking algorithm.

Chapter 3 covers the work which was reported in the 1998 paper **"Modelling repressor proteins binding to DNA"**.[8] This was the first report of protein/DNA interface pair potentials.

Chapter 4 shows work which integrated the BMM software, and tested it on a significantly larger set of test cases. Parts of it have been reported elsewhere.[9–11] This is the methodology that has been used for the CAPRI competitions.

Chapter 5 contains the conclusions to this thesis.

The appendices include a lists my publications connected to this thesis and a copy of the software manual that is available for the programs I have written and made available.

1.2 Structural studies of protein complexes

The first step in the development of a protein docking algorithm is to examine the known complexes to establish the principles of molecular recognition. Following early work by Chothia,[12] there have been reviews that examined the interactions between hetero-protein complexes.[2,13] These analyses have focused on the static structure of the complexes. A major problem in protein docking is to cope with the conformational flexibility that occurs on complex formation. This was addressed by the study of Betts and Sternberg in 1999.[3] This analysed the conformational changes on complex formation for 39 pairs of proteins, from their unbound states to the formation of protein/protein complexes. The dataset mainly consisted of enzyme/inhibitor and antibody/antigen complexes, but also included other systems such as human growth hormone and its receptor.

The conformational differences were evaluated in terms of root mean square deviations (RMSD), both of C_α positions and of side-chain positions. Residues were identified as exposed when their total relative side-chain accessible area (main-chain for Gly) was greater than 15%. Interface residues were defined as having at least one atom within 4Å of the other component of the complex.

To assess the significance of the differences between bound and unbound structures, it is necessary to identify the differences in coordinates that can occur simply from the crystallographic determination of the structure. Accordingly, 12 pairs of independently solved crystal structures of identical proteins were analysed. 92% of this dataset (11 out of 12) had an RMSD for exposed C_α atoms of < 0.6Å, and for exposed side-chain atoms of < 1.7Å. These values were taken as the control cut-off values.

Four measures were used for overall conformational change

- RMSD of C_α atoms for just the interface atoms
- RMSD of C_α atoms for exposed non-interface atoms

- RMSD of side-chain atoms for just the interface atoms
- RMSD of side-chain atoms for exposed non-interface atoms

It was found that 19 out of the 39 proteins that were involved in complex formation do not have any of their four measures for conformational change above the control determined cut-off values. Many of the other proteins showed changes less than double the cut-off values. Thus for many systems there is limited overall conformational change on protein/protein association. This suggests that the rigid-body approach should be widely applicable.

In addition to overall conformational changes, a few individual residues might show particularly large conformational changes. This could markedly affect the viability of the rigid-body approach. The control cut-off values for the movement of an individual residue are 3.0Å for the C_α and 5.6Å for side-chain atoms. Examination of the complexes showed that all large movements of exposed residues that were not in the interface can be explained either by their close proximity to the interface or by structural disorder. For a few of the systems there are movements of individual or sets of residues in the interface that are above the control cut-offs. These shifts are intimately involved in the complex formation. Thus there are several complexes that have substantial conformational changes for individual or sets of residues, while still having a limited overall structural perturbation.

The general conclusion from this analysis was that protein/protein interactions are described by the induced fit model. That is to say, complex formation is accompanied by conformational changes that benefit the formation. Some examples are illustrated in Figure 1.2. However, for many systems the extent of conformational change is limited, and a lock-and-key model is a valid first approximation. Accordingly, for many systems it is appropriate to develop a protein/protein docking algorithm that starts with the docking of the components as rigid bodies – *i.e.* the rigid-body approach. Only as a refinement is it then necessary to consider the limited conformational changes. The major caveat is that the systems analysed were dominated by enzyme/inhibitor and antibody/antigen complexes. Other biological systems may exhibit a greater degree of conformational change on complex formation.

1.3 BMM protein/protein docking strategy in 1996

In this section, the algorithms developed in the Biomolecular Modelling Laboratory (BMM) at ICRF by the time of CASP2 (December 1996, see 1.5.2) are described. The other chapters describe work carried out since that time. At that time the strategy consisted of:

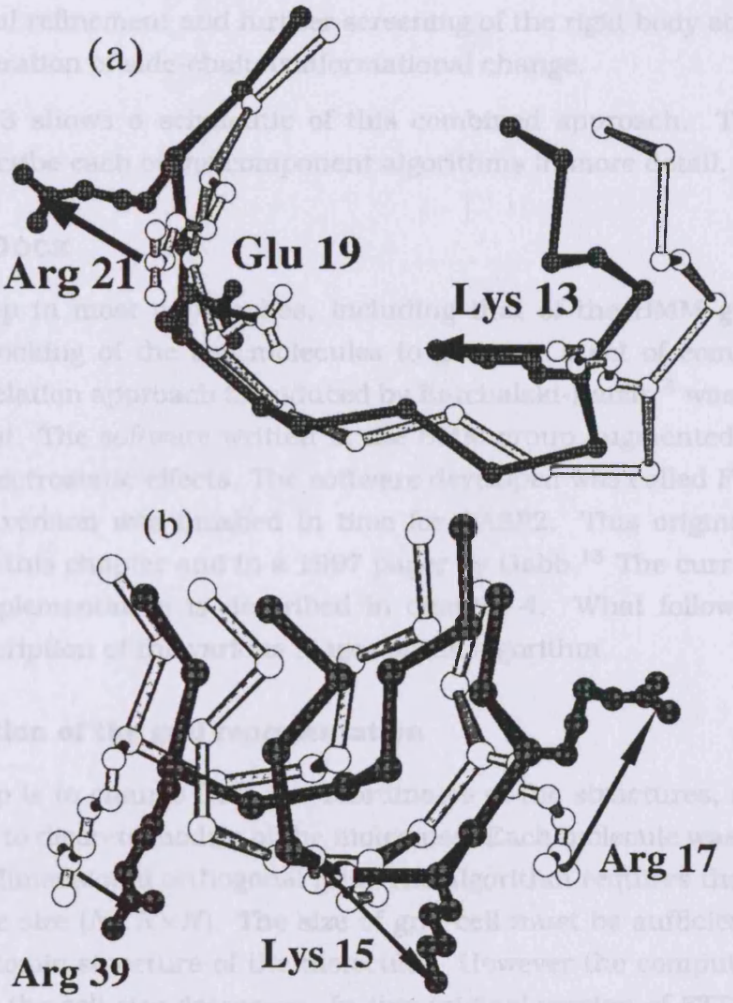


Figure 1.2: Examples of docking to illustrate induced binding in the interface. (a) Selected binding site residues of ovomucoid when free (2ovo - light grey) and when bound to α-chymotrypsin (1cho - dark grey). (b) Selected binding site residues of BPTI when free (4pti - light grey) and when bound to trypsin (2ptc - dark grey). Scoring functions used for predictive docking must be sufficiently "soft" to allow for conformational changes of this magnitude.

1. A rigid-body search for putative docked complexes that are favourable in terms of shape complementarity and electrostatics,
2. The use of biological distance constraints, particularly details of the binding sites in one or both proteins, to screen the putative complexes.
3. The final refinement and further screening of the rigid-body structure by a consideration of side-chain conformational change.

Figure 1.3 shows a schematic of this combined approach. The following sections describe each of the component algorithms in more detail.

1.3.1 FTDock

The first step in most approaches, including that of the BMM group, is the rigid-body docking of the two molecules to generate a set of complexes. The Fourier correlation approach introduced by Katchalski-Katzir[14] was used as the starting point. The software written in the BMM group augmented the method to include electrostatic effects. The software developed was called FTDock, and the original version was finished in time for CASP2. This original version is described in this chapter and in a 1997 paper by Gabb.[15] The current (finished in 2001) implementation is described in chapter 4. What follows is a more detailed description of the various stages of the algorithm.

The generation of the grid representation

The first step is to change from the coordinates of the structures, as described in PDB files, to discrete models of the molecules. Each molecule was placed onto a regular 3-dimensional orthogonal grid. The algorithm requires that both grids are the same size ($N \times N \times N$). The size of grid cell must be sufficiently small to model the atomic structure of the molecules. However the computational time increases as the cell size decreases. In this original version of FTDock, N was set at 64 at compile time. This resulted in grid cell sizes from approximately 1.3Å to 2.2Å on a side, over all the systems that were studied.

The discretisation is done with each molecule at the centre of its own grid. The empty space not filled by the molecule is necessary for the algorithm, as the convolution in Fourier space is cyclic. To perform the discretisation, each grid cell within which an atomic position is found is turned 'on'. Grid cells whose centre is within 1.8Å of any atomic position are also turned on. This value of 1.8Å was chosen to approximate an effective van der Waals radius for an atom combined with any hydrogen atoms that are bound to it. Thus the surface of the resulting grids represents the atomic surface of the molecules.

Next, the larger static molecule (S) is assigned a surface thickness below its atomic surface. This means that any grid cell that was turned on by the

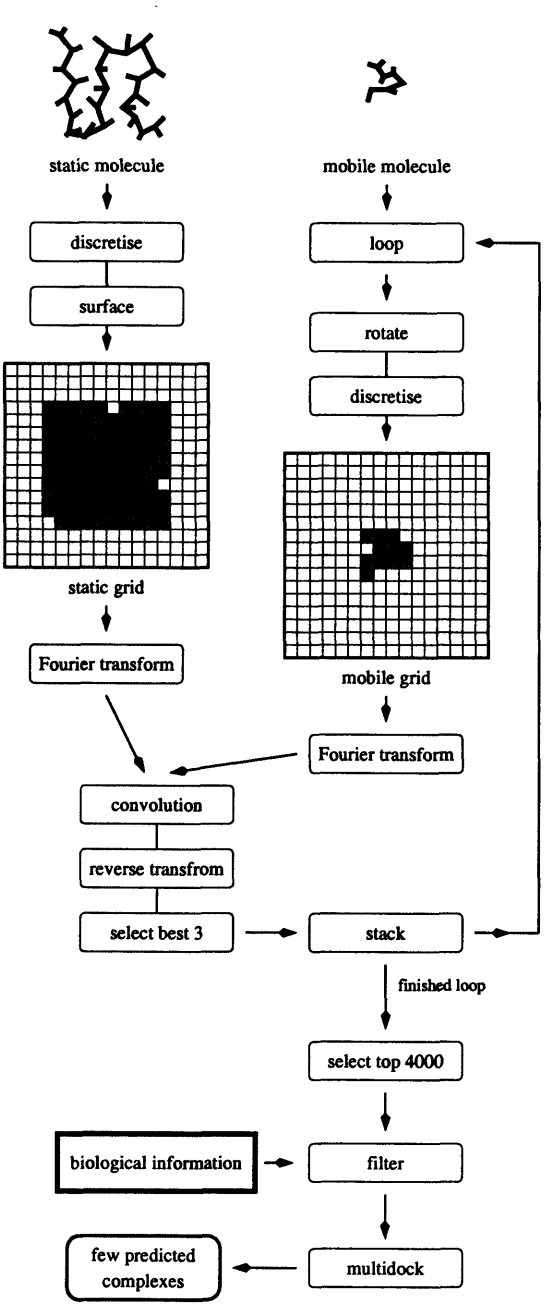
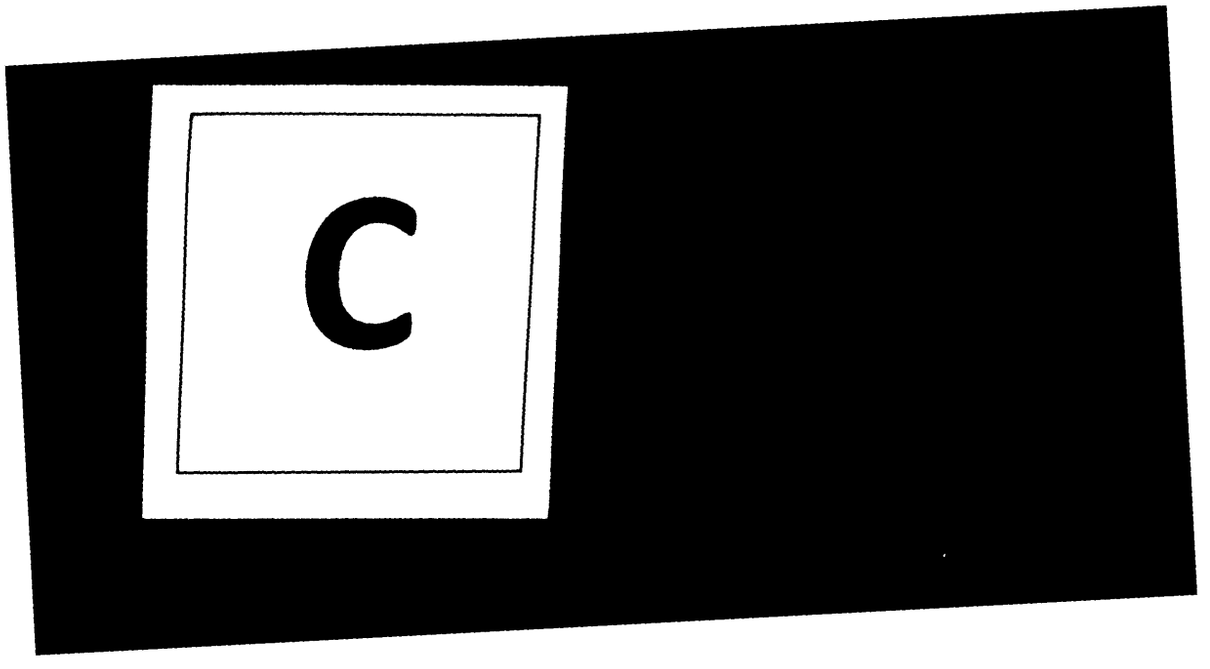


Figure 1.3: Flow chart of the docking algorithm in BMM lab in 1997 for CASP2.


C

discretisation algorithm, whose centre is within the surface thickness of a cell not turned on by that algorithm, now becomes assigned as a surface cell. This is essential in order to be able to calculate the quality of the fit. For algorithmic speed, this surfacing procedure is done to the static molecule, and accordingly it needs to be done only once by the program. The depth of this surface was 1.4Å. A slice through such a discretised and surfaced molecule can be seen in Figure 1.4. The abundance of cells assigned as surface is necessary for complex formation since we are using a rigid body approximation where steric clashes are inevitable.

Figure 1.4: Discretisation and surfacing of a slice of 1BRA by FTDOCK. *The protein is shown as spacefilled blue. Those grid cells switched to the internal deterrent value are shown as red, and those that are surface value are shown as green. In this diagram, the grid cell size is 0.7Å and the surface is 1.3Å. These are values that are used in the current version of* FTDOCK, *for which a subroutine was written than allowed for the viewing of the discrete grid as shown.*

Evaluation of shape complementarity

The grid values $s_{l,m,n}$ for the static molecule (S), that has been discretised and then surfaced, at grid cell l, m, n are given by

$$s_{l,m,n} = \begin{cases} 1 & \text{for grid points on the surface of the molecule} \\ \rho & \text{for the core of the molecule} \\ 0 & \text{for outside the molecule} \end{cases}$$

where ρ is negative (we use -15, see below).

For the second molecule (M), that has just been discretised, the grid values are given by

$$m_{l,m,n} = \begin{cases} 1 & \text{for the molecule} \\ 0 & \text{for outside the molecule} \end{cases}$$

The two grids can then be superimposed, and the mobile grid (M) discretely translated by α, β, γ. The value

$$s_{l,m,n}.m_{l-\alpha,m-\beta,n-\gamma}$$

gives the shape complementarity for grid cell l, m, n of the static molecule. If either grid cells are empty of molecule, then the product is zero, as would be desired. If the static grid cell has a surface value of 1, and the mobile grid has a molecule value of 1, then the product is 1, which is positive and favourable. However, if the static grid cell has a core value of ρ, and the mobile grid has a molecule value of 1, then the product is ρ, which is negative and unfavourable. The summation of all such values over the grid representing the static molecule provides the surface complementarity score for a given translation of α, β, γ. *i.e.*

$$c_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} s_{l,m,n}.m_{l-\alpha,m-\beta,n-\gamma}$$

which is a convolution, *i.e.*

$$c = s * m$$

Figure 1.5 shows this in a 2D analogy.

Use of discrete Fourier transforms

The value of c is a convolution and its calculation requires approximately N^3 multiplications (and a summation) for every N^3 translations of α, β, γ, resulting in calculation times proportional to N^6. Katchalski-Katzir[14] introduced the use

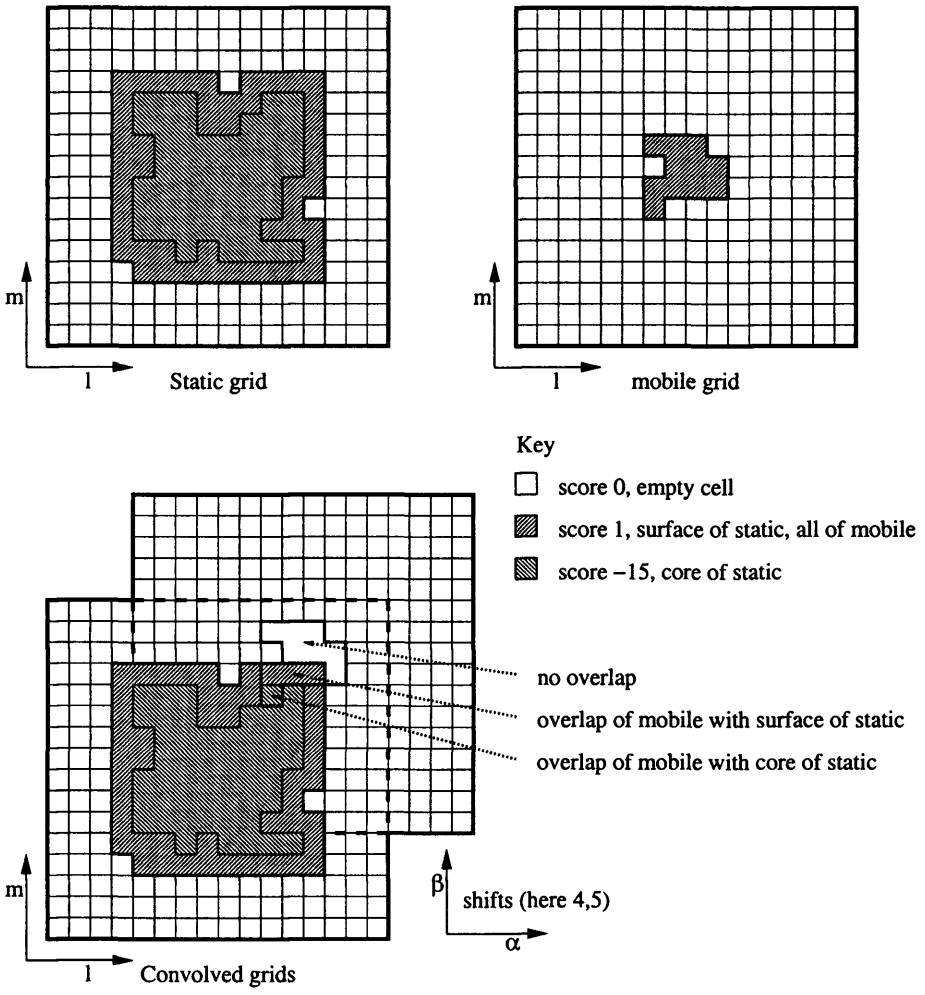


Figure 1.5: 2D analogy of a surface complementary calculation.

of discrete Fourier transforms to speed up the process of calculating c. First calculate the discrete Fourier transforms of the discrete functions $s_{l,m,n}$ and $m_{l,m,n}$, *i.e.* the two grids. Let the resulting transforms be denoted as $\mathcal{F}(s)$ and $\mathcal{F}(m)$. The multiplication of these two functions is the Fourier space equivalent to the convolution of the functions s and m, *i.e.*

$$\mathcal{F}(s * m) \equiv \mathcal{F}(s) \times \mathcal{F}(m)$$

so

$$s * m \equiv \mathcal{F}^{-1}(\mathcal{F}(s) \times \mathcal{F}(m))$$

where $\mathcal{F}^{-1}$ is the reverse Fourier transform. Hence c can be calculated using Fourier space. The multiplication in Fourier space is of the order N^3, and the Fourier transforms and reverse Fourier transform are dependent on the discrete Fourier transform algorithm. If performed efficiently then they are of the order $N^3 logN$. So, by using fast discrete Fourier transforms, the time for the calculation of c for all α, β, γ is reduced from N^6 to the order of $N^3 + N^3 logN$, a reduction by at least N^2. Where N is 64, this results in an approximately 3000 fold reduction in computing time.

The global search

The mobile molecule was rotated to sample all possible rotations in as fair a way as possible. Three rotational angles are required in order to describe the orientation of a three-dimensional object in a three-dimensional space. These three Euler angles were sampled at 15°. (Since only integer values are used to describe the rotations, this limits the angle step used to being an integer factor of 180.) This results in 360×360×180/15 = 6912 orientations. However, many of these orientations are degenerate, and so must be removed using the following relationship.[16]

$$\alpha = cos^{-1} \frac{tr(R_1 \times R_2^T) - 1}{2}$$

where R_1 is the rotation matrix of the first orientation, R_2^T is the transpose of the rotation matrix of the second orientation, and tr is the matrix trace. If $\alpha \leq 1°$ then the two orientations are degenerate. Removing degeneracies in this fashion yields 6385 unique orientations. A finer angular rotation rapidly results in more orientations, *e.g.* 22,105 for a sample of 10°.

Residue	Atom type	Charge (e)
All	C_α	0.0
All	C	0.0
All	N	0.5
All	O	-0.5
All	Terminal N	1.0
All	Terminal O	-1.0
Arg	NH	0.5
Glu	OE	-0.5
Asp	OD	-0.5
Lys	NZ	1.0
Pro	N	-1.0

Table 1.1: Charges assigned by FTDOCK *to the proteins, used for electrostatic complementarity calculations.*

Electrostatic effects

Both shape complementarity and electrostatic effects are important in the recognition process during protein complex formation. Accordingly, a treatment of electrostatics was introduced (which was not present in the original work by Katchalski-Katzir[14]). The charge-charge interaction is evaluated from point charges of the mobile molecule M interacting with the electric field potential from static molecule S. This choice results in having to perform the more computationally intensive potential calculation only once for the static molecule, whilst the quick charge calculation is performed for every rotation of the mobile molecule. In the above treatment of rigid-body docking based on shape complementarity, it is possible for a model complex to place two charges closer together than would be allowed by van der Waals packing. Since the potential energy of two interacting charges depends inversely on their separation, such close placement would result in an artificially very favourable or very unfavourable interaction. These artificial terms are prevented in the method used, by setting a lower limit (2Å) on the distance at which a charge effects the field potential.

Charges, as shown in Table 1.1, were assigned to the atoms of molecule S and the electrostatic potential evaluated from

$$\phi_{l,m,n} = \sum_j \frac{q_j}{\varepsilon(r_{ij}) r_{ij}}$$

where $\phi_{l,m,n}$ is the potential for grid cell l, m, n (position i), q_j is the charge on

atom j, r_{ij} is the distance between i and j (with a minimum value of 2Å to avoid artificially large values of the potential as mentioned above) and $\varepsilon(r_{ij})$ is a distance dependent dielectric function, namely the sigmoidal function of Hingerty,[17] given by

$$\varepsilon(r_{ij}) = \begin{cases} 4 & \text{for } r_{ij} \leq 6\text{Å} \\ 38r_{ij} - 224 & \text{for } 6\text{Å} < r_{ij} < 8\text{Å} \\ 80 & \text{for } r_{ij} \geq 8\text{Å} \end{cases}$$

This function was originally introduced for modelling the effective dielectric between atoms in proteins. The rationale for this function is that at close separation ($r_{ij} \leq 6$Å), when there is no intervening water molecules, the effective dielectric is that of protein atoms, and a value of 4 is appropriate. For separations of 8Å or more, the dielectric is dominated by the screening effect of the intervening water, and so the value for bulk water (80) is used. Between these two separations a linear interpolation is used. In FTDOCK, precise atomic positions are not used, but there is still a need to model the complex dielectric behaviour of proteins in solvent, and so this function was used and found to perform well.

The potential $\phi_{l,m,n}$ is only assigned to grid cells outside and on the surface region of molecule S. For the core of molecule S, where $s_{l,m,n} = \rho$, $\phi_{l,m,n}$ is zero. For the mobile molecule M, the charges on the charged atoms are distributed amongst the closest 8 grid cells.[18] Figure 1.6 shows the normalised coordinate system used.

Given the atomic position of the charged atom, normalised onto the shared vertex of the 8 neighbouring grid cells, as (x, y, z), the charge on the atom being q, and the normalised centre of any one of those closest 8 cells as (X, Y, Z), then the charge given to that cell is

$$\frac{q}{8} \times \frac{x+X}{X} \times \frac{y+Y}{Y} \times \frac{z+Z}{Z}$$

Each grid cell has a total pseudo-charge that is the summation of all charged atoms that are in its immediate neighbourhood.

Having now assigned an electric field to each grid cell of S, and electric charges to each grid cell of M, the electrostatic interaction $e_{\alpha,\beta,\gamma}$ for a translation of α, β, γ is given by

$$e_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \phi_{l,m,n} \cdot q_{l-\alpha,m-\beta,n-\gamma}$$

This equation can be seen to be analogous to the equation for the surface complementarity score $c_{\alpha,\beta,\gamma}$, derived above. Hence it is possible to treat the electrostatic charges in Fourier space in the same way as shape complementarity is used.

Generation of model complexes

In trials with FTDOCK, it was found that the electrostatic term worked best as a binary filter. This resulted in a method where complexes with unfavourable electrostatics were immediately discarded, and the remaining docked structures ranked by shape complementarity. For a given orientation of the movable molecule, the shape complementarity correlation function c was examined, and the three highest scoring models stored. After all orientations were sampled, the top 4000 of the models were kept for subsequent examination (Figure 1.3).

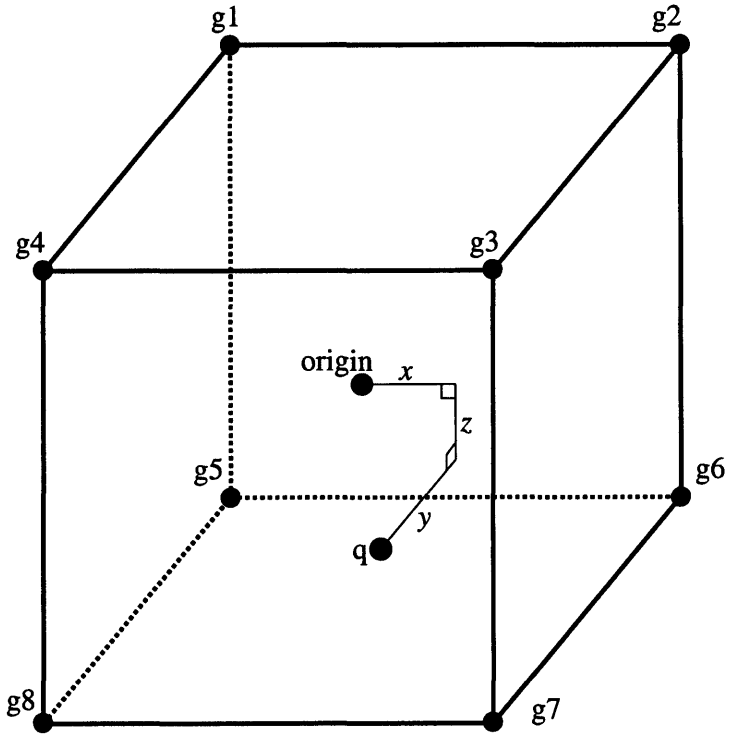


Figure 1.6: Schematic diagram of the method used to distribute an electronic point charge among its nearest eight grid cells.
The charge is q, with normalised position (x, y, z). The eight grid cells centres' are labelled $g1$ through to $g8$, and their common centre is the origin of the coordinate system.

1.3.2 Use of distance constraints

The location of the binding site can yield distance constraints to be used to filter the model complexes. An intermolecular residue-residue interaction was defined if any pair of atoms is closer than a 4.5Å distance cut-off. This cut-off value incorporates the van der Waals radii plus the probable error in conformation of any model complex. One feature of the Fourier correlation method, as implemented in FTDOCK, is that biological constraints cannot be used to reduce the initial search, so forcing any pruning of allowed model complexes to be after a global search.

1.3.3 Finescan

The original FTDOCK implemented a possibility for scanning the space immediately around an already identified model. This scan used a finer angle step, and the linear translation and orthogonal discretisation was based on a compile time grid size of 128. This step was used on all the models which passed the distance constraints filter. (A minor error existed in that this resulted in some duplicated refined models, but these were not identified and removed.) This step did improve the results, though not significantly.

1.3.4 Additional screening of models

The above strategy explores and ranks rigid-body model complexes. The next step is to allow for conformational changes in side-chains, and to consider the interactions between the proteins at the atomic level. Jackson developed a procedure,[19] MULTIDOCK, that models both side-chain conformational changes in a mean field approach, together with limited rigid-body shifts between the components of the model complex. The energy of interaction is evaluated from a molecular mechanics function.

Potential energy function

The proteins are represented at the atomic level by multiple copies of side-chains, on a fixed peptide backbone, modelled according to a rotamer library that gives the commonly occurring side chain conformations.[20] The use of a rotamer library means that side chains adopt a limited number of conformations, rather than being able to sample every value for bond rotation.

The model description needs to consider the various types of protein/protein interactions. The terms for van der Waals interactions are taken from the AMBER force field,[21] and atomic charges from the PARSE parameters.[22] A cut-off of 10Å is used in the calculation of the non-bonded interactions (van

der Waals and electrostatic) between atoms. The dielectric screening between charges is represented by the widely-used distance dependent dielectric

$$\text{dielectric } \varepsilon = \text{distance of separation in Ångstroms}$$

The effect of this dielectric model is that for close separations the value is low, representing the dielectric due to protein atoms. Thus at a separation of 4Å the dielectric is 4. For larger separation the dielectric increases, so representing the greater electrostatic screening due to water molecules (that are not included explicitly).

The starting model generated by FTDOCK (or another rigid-body docking approach) is unlikely to be a very good one. A consequence of this is that there can be unrealistically close approaches of atoms that would distort the modelling by MULTIDOCK, due to high repulsive van der Waals interactions or electrostatic effects of large magnitude. To reduce this effect, van der Waals interactions were truncated to a maximum value of 2.5 kcal/mol. Similarly, an electrostatic interaction scheme was used in which a minimum allowed distance separation between two interacting charges is set. This means that atom pairs that come closer than allowed are re-scaled to realistic values which are no greater than the approximate sum of their van der Waals radii. The minimum allowed distances for two charges are 3Å for two heavy atoms, 2Å for one heavy atom with a hydrogen, and 1Å for two hydrogen atoms. It should be noted that the treatment of the screening effect of water for larger separations and the prevention of anomalous electrostatic effects were also included in FTDOCK.

Refinement procedure

The object of the refinement procedure is to move from a completely rigid-body docking scheme to one that includes both flexibility of the side-chains together with a limited re-orientation of the two interacting molecules. Thus the refinement procedure is an iterative two step approach repeated until convergence involving

1. optimisation of the protein side chain conformations by a self-consistent mean field approach[23,24]
2. rigid-body energy minimisation to relax the protein interface

The mean-field approach

The side chain degrees of freedom are defined by a conformational matrix, **CM**, where each rotamer, k, has a probability of $\mathbf{CM}_{i,k}$, where the sum of the probabilities for a given residue, i, must be equal to 1. The potential of mean force, $E(i,k)$, on the k^{th} rotamer of residue, i, is given by

$$E(i,k) = V(\chi_{ik}) + V(\chi_{ik}, \chi_{mc}) + \sum_{j=1, j\neq i}^{N} \sum_{l=1}^{K_j} \mathbf{CM}_{j,l} V(\chi_{ik}, \chi_{jl})$$

where V is the potential energy, χ_{ik} are the coordinates of atoms in rotamer k of residue i, and χ_{mc} are the coordinates of atoms in the protein main chain. N is the number of residues in the protein, and K_j is the number of rotamers for residue j. The first term models the internal energy of the rotamer whilst the second represents the interaction energy between the rotamer and all the main chain atoms. These two values are constant for a given rotamer on a given main chain. The third term models the interaction energy between the rotamer and all the rotamers of other residues, weighted by their respective probabilities.

Given the effective potentials acting on all K_i possible rotamers of residue i, the Boltzmann principle can be used to calculate the probability of a particular rotamer

$$\mathbf{CM}_{i,k} = \frac{e^{-E(i,k)/RT}}{\sum_{k=1}^{K_i} e^{-E(i,k)/RT}}$$

where R is the Boltzmann constant and T the temperature. The values of $\mathbf{CM}_{i,k}$ are substituted back into the equation describing $E(i,k)$, and its new value recalculated. This process is repeated until values of $\mathbf{CM}_{i,k}$ converge. The predicted structure corresponds to the rotamer of each residue with the highest probability. Trials showed that the procedure converged to the same side-chain rotamers using a number of different schemes for initiating the starting probabilities in the **CM** matrix.

Rigid-body energy minimisation

After a complete cycle of mean field optimisation of side-chain conformation, a rigid-body minimisation is performed on the resultant coordinates of the new model complex. Only interface residues whose C_β atoms (C_α for Gly) are within 15Å of a C_β atom of the other molecule are included in the minimisation. The larger molecule is kept stationary while the three rotational and three translational degrees of freedom of the smaller mobile molecule are varied according to the path determined by the derivatives to minimise the intermolecular interaction energy. The steepest descents approach for minimisation is used.

1.3.5 Results

The benchmark used by both Gabb[15] and Jackson[19] consisted of six enzyme/inhibitor and four antibody/antigen complexes. With the exception of

System	PDB Code	N	N_{good}	Rank	RMSD (Å)
α-chymotrypsinogen / human pancreatic trypsin inhibitor	1CGI[25]	93	1	3	1.8
α-chymotrypsin / ovomucoid	1CHO[26]	85	5	11	1.2
kallikrein / bovine pancreatic trypsin inhibitor	2KAI[27]	349	16	128	1.5
trypsin / bovine pancreatic trypsin inhibitor	2PTC[28]	205	7	12	1.5
subtilisin / Streptomyces subtilisin inhibitor	2SIC[29]	26	2	8	1.8
D1.3 F_{ab} / lysozyme	1FDL[30]	636	2	149	2.1
D44.1 F_v / lysozyme	1MLC[31]	539	4	34	1.2
HyHel-5 F_{ab} / lysozyme	3HFL[32]	498	2	218	1.8
HyHel-10 F_{ab} / lysozyme	3HFM[33]	700	4	48	1.1

Table 1.2: Original FTDOCK *results as reported in Gabb* et al. *1997.*

For fuller descriptions of the test systems, see Section 2.2.1. N is the number of complexes left after filtering and finescan. N_{good} is the number of models that are considered good, by the criteria of having a RMSD $\leq$ 2.5Å from the crystallographic structure. RMSD is calculated over Cαatoms for all residues. The numbers in this table are in some cases slightly lower than those in reported in 1997 by Gabb[15] due to the removal of duplicate structures generated by local refinement after the global scan.

the coordinates of two antibodies (HyHEL5 and HyHEL10), all the coordinates were from separately determined unbound components. Table 1.2 presents the results from FTDOCK. For each complex system, 4000 model complexes were generated. The value of N shows the number that were left subsequent to a biological filter and local refinement. As all the enzymes were serine proteinases, the distance filter for them was that at least one residue in the inhibitor must be in contact with the one of the catalytic triad residues (*i.e.* His, Ser or Asp). For the antibody/antigen systems, the constraint was that the antigen must contact the third complementarity determining region of either the light or the heavy chain (CDR-L3 or CDR-H3).

In this study a good prediction was defined as within 2.5Å of the correct structure over the C_α atoms of all residues in the complex. The number of good predictions is shown as N_{good}. FTDOCK generated at least one good model in all bar one of the systems. The results for subtilisin docking to its inhibitor are not shown, as no good models were generated for this system.

The evaluation of MULTIDOCK was based on these lists. The program was run on the N models, and the energy function used to re-rank them. Table 1.3 shows the results, both for when MULTIDOCK was run *in vacuo*, and for when solvent was included. The solvent calculations took a large amount of computational time, so only the two antibodies with more favourable results from the *in vacuo* simulations were evaluated, and then only the top 50 models as ranked by those *in vacuo* simulations. It can be seen that the ranks are an improvement on those from FTDOCK for all the enzyme/inhibitors, and for two of the four antibody/antigens. For the enzyme/inhibitors the results are very good.

Figure 1.7 illustrates how MULTIDOCK is capable of moving a side chain out from a steric clash and towards its correct position (not shown in the figure).

1.3.6 Implementation of the docking suite

The original distributed version of FTDOCK (version 1.0) had two implementations of Discrete Fast Fourier Transforms (DFFTs). One was the DFFT routines from Numerical Recipes Software,[34] which can be implemented on most platforms. The other implementation used the more efficient Silicon Graphics library functions, suitable for those with SG machines, including parallel architectures.

The FILTER program implemented the distance constraints. The program takes a list of inter-molecular constraints which can either be residue to residue, chain to residue, or chain to chain. Only one from a given list of constraints needs to be satisfied for the program to accept a model complex as passing the filter (*i.e.* logical OR).

MULTIDOCK was made available as an executable for a Silicon Graphics

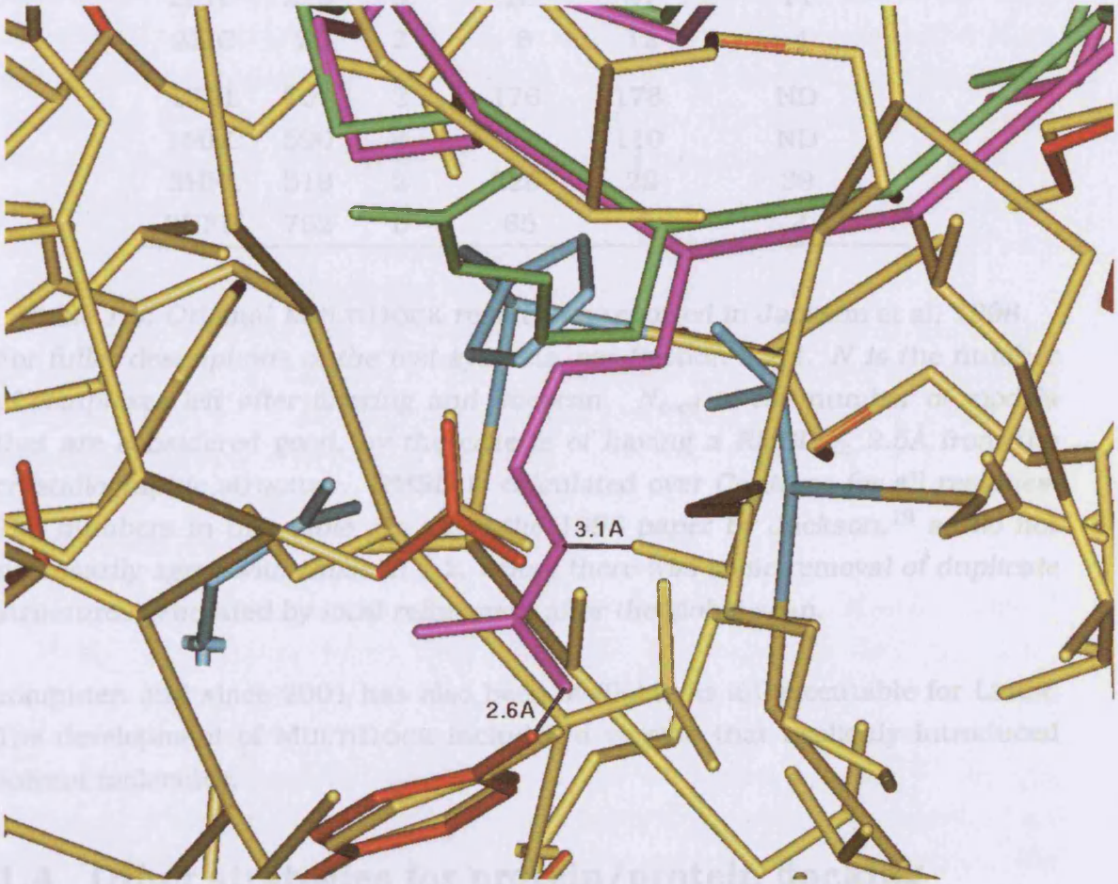


Figure 1.7: Movement of Arginine in interface of Trypsin complex (1BRC) by MULTIDOCK

The chains coloured orange (trypsin) and magenta (inhibitor) are from the rigid body docking. The chains coloured yellow (trypsin) and green (inhibitor) are from the MULTIDOCK *refinement. Most of the orange chain is in an identical position to that of the yellow, in which case only the yellow is seen. The cyan residues are the catalytic triad of the trypsin. Two distances are shown between an inhibitor Arginine and the trypsin in the rigid body docking positions, showing that in this position there is a steric clash. When moved by* MULTIDOCK*, the Arginine is no longer clashing (it is fully 'in front' of the catalytic triad Histidine).*

System	N	N_{good}	Rank FTDOCK	Rank E^{int}	Rank $E^{int} + \Delta\Delta G^{solv}$
1CGI	94	1	3	2	3
1CHO	86	5	11	1	1
2KAI	364	18	130	2	2
2PTC	229	8	16	47	11
2SIC	26	2	8	12	4
1FDL	707	2	176	178	ND
1MLC	590	4	41	110	ND
3HFL	519	2	228	29	39
3HFM	762	6	65	9	4

Table 1.3: Original MULTIDOCK *results, as reported in Jackson* et al. *1998. For fuller descriptions of the test systems, see Section 2.2.1. N is the number of complexes left after filtering and finescan. N_{good} is the number of models that are considered good, by the criteria of having a RMSD $\leq$ 2.5Å from the crystallographic structure. RMSD is calculated over Cαatoms for all residues. The numbers in this table are as in the 1998 paper by Jackson,[19] so do not necessarily agree with those in 1.2, where there was some removal of duplicate structures generated by local refinement after the global scan.*

computer, and since 2001 has also been available as an executable for Linux. The development of MULTIDOCK included a version that explicitly introduced solvent molecules.

1.4 Other strategies for protein/protein docking

Different groups have explored a variety of methods for protein/protein docking. This section reviews the major strategies that have been used, with emphasis on recent developments.

1.4.1 Evaluation of the results of docking simulations

There is a fundamental problem when comparing different reported results. This is that each study will tend to use a different calculation for the quality of the models. Although the calculations themselves are accurate, some values are more correctly descriptive than others.

In the following summary of the different algorithms to date, little assessment will be made of how each study evaluated its results. Given that it is not possible to re-calculate reported values to what may be considered a more correct value without having the actual coordinates of each model, there is little use in trying

to quantitatively compare different studies. For a more detailed description of all the possible measures that can be calculated, see 4.4.

Each study also used a different set of complexes on which it was assessed. It will be noted whether the study used bound complexes that were re-docked after having separated the components, or if it used components of unbound coordinates that were determined separately from the complex. The former does not reflect a real-world problem. The later is the real-world problem, and is more difficult since it requires the algorithm to be able to cope with conformational changes on association. Also, although it will be reported which complexes that were studied, it will not be noted how hard these cases are. It is however generally true that smaller enzyme/inhibitor complexes are easier to get good results for than larger antibody/antigen complexes, and so a study that only tackled easier cases is possibly of less widely applicable value.

In all the studies, the models that result from the algorithm are sorted by whichever energy function is used. This results in a ranked list of model complexes, and in all studies, rank 1 is the best rank, and rank n, where the algorithm generates n models, is the worst.

1.4.2 Fourier correlation methods

The Fourier correlation method was originally introduced by Katchalski-Katzir[14] in 1992. The original paper emphasised the application to docking bound complexes. This is the method used, with additions, by FTDOCK.

One of the authors of this work was Vakser, who has continued with this approach and explored its application to low resolution docking.[35] In this approach, a low resolution grid (typically several Ångstroms) is used. This leads to representing only the major spatial features of the two molecules. The surface is therefore smoother, *i.e.* not at the atomic level, and the search much faster as the number of cells used to represent the two molecules is far fewer. This low resolution docking was tested on a database of 475 co-crystallised protein/protein complexes.[36] Most of the database were multimeric proteins, though it also included complexes of the type conventionally studied in docking. The study considered the binding region to be defined as being within a 10Å region traced out from the centre of mass of the ligand in the true complex. Hence, a good docking is one which places the centre of mass of the ligand within 10Å of its correct position. Note that the orientation of the ligand is not considered, only the location of its centre of mass, so a good model will by this criteria have identified the receptor binding site but not necessarily the correct section of the ligand surface (see 4.4 for a more detailed evaluation of this). The docking program GRAMM was run at grid spacing of 6.8Å and a rotational step of 20°. The receptor binding site was recognised by the above criteria in 52%

of the complexes used. For 113 complexes with a large interface area (> 4,000 Å^2), the success rate rose to 76%. Hence this low resolution docking provides a possible tool to delineate a putative binding site in a protein. This could then be used to provide a filter for results of higher resolution procedures.

Palma[37] has implemented the Katchalski-Katzir algorithm, with various modifications, but without making use of Fourier space mathematics. By using grid representations that use logical states, as opposed to numerical values, and some very fast logical matrix algorithms, they report speeds for simulations that are in fact faster than those of either FTDOCK or GRAMM. They report results for systems starting with unbound components which are generally useful. However, both of the antibody/antigen systems failed.

A new approach for rapid protein/protein docking, HEX, has been introduced by Ritchie and Kemp,[38] that employs spherical polar Fourier correlations. This removes the time consuming requirement of FTDOCK and GRAMM for explicitly generating different orientations for the mobile molecule. The algorithm considers both shape complementarity and electrostatic effects. Note that the approach uses a Fourier correlation but does not use a discrete Fourier transform as employed by FTDOCK and GRAMM. A full search on a single workstation is reported as taking around two hours. Importantly, constraints on the location of the receptor binding site can readily be incorporated, so reducing the calculation to the order of minutes. In the conventional Fourier approach (FTDOCK and GRAMM), knowledge of the binding site in either of the two molecules was not used to provide such a constraint. The program has been extensively benchmarked on unbound docking and yields results that suggest it is a valuable tool.

It has been standard with those methods based on algorithms using Fourier convolution, such as FTDOCK, that filtering is done after a global scan. Since the grid representation loses information about which residues are being represented, it was not obvious how a biological filter could be introduced during a global scan. However, this problem has been solved by Ben-Zeev and Eisenstein,[39] using complex numbers in the grid, and using the imaginary part to represent biologically derived constraints.

ZDOCK[40] is a more recent rigid body Fourier grid method. The scoring function uses pairwise shape complementarity, desolvation and electrostatics. The shape complementarity function is somewhat different from that used by FTDOCK. A favourable value is given by a count of all atom pairings across the interface. Unfavourable values are then assigned to all overlap cases, with increasing severity of 9, 27, and 81 to surface/surface, surface/core, and core/core overlap. The desolvation term is calculated from the ACE[41] potential. The electrostatic term is not simply a binary filter as used in FTDOCK. These three terms are variously weighted to what is considered to allow for the

correct balance of the respective terms towards an overall potential. RDOCK[42] is a second stage refinement that performs an energy minimisation using the molecular mechanics software CHARMM.[43]

1.4.3 Other rigid-body docking approaches

Janin and coworkers[44] in 1991 used spheres of different radii to model residues, and docked the resultant structures so as to maximise the buried surface area. The resultant initial models were refined using a Monte Carlo search. In a Monte Carlo search, trial perturbations are generated randomly from the existing state. A perturbation is always accepted if it reduces the energy, and may also be accepted if the energy increases but is below a statistical value determined by Boltzmann's principle. After the Monte Carlo search, a full atomic representation (apart from hydrogen) was restored. An energy minimisation was then performed that allowed the side-chains in the predicted interface to move. Note that the selection of initial models examined to include flexibility is constrained by the rigid-body approximation. The method was tested on two unbound docking systems,[44] resulting in one system for which a native-like solution was generated at rank three.

The DOCK algorithm, developed by Kuntz and his group, is widely used in the community for the docking of low molecular weight ligands to protein receptor. It can however also be applied to protein/protein docking.[45] The DOCK approach fills the binding site of one protein (the receptor) with a cluster of overlapping spheres. Then the algorithm matches the sphere centres of this cluster with similar clusters from the ligand protein. Predicted models are then ranked in terms of a score for residue-residue contact. In the early study on docking starting with unbound complexes,[45] individual atoms from the molecules had to be manually truncated to obtain good results. Although there were suggestions of which atoms were mobile, from the crystallographic thermal parameters present in the coordinate file, in other simulations it may not be known in advance which atoms need to be truncated. This approach has been developed further for macromolecular docking by Hendrix.[46] The method consists of three steps: defining the shape-based sites that define putative docking locations; docking using these site points; and scoring the docked complex. Complexes were scored using van der Waals and electrostatic interactions calculated from the AMBER program with united-atom parameters.[21] The method was benchmarked on re-docking several complexes starting from the bound coordinates. The approach was then used to suggest a model for the docking of human growth hormone to its receptor.

The groups of Wolfson and Nussinov have developed an approach for docking approach based on matching critical points.[47–49] These points define the knobs

and holes on the two interacting surfaces. Both surface points and surface normals are matched. The approach can be implemented as a fast program, and a global search takes of the order of minutes. After the search, putative solutions are checked to penalise overlap. In addition, the extent of hydrophobic packing across the interface is assessed. The method was recently tested on four different protease/inhibitor systems starting with unbound components and two bound antibody / unbound antigen complexes. For each system studied, several sets of coordinates were used in different runs of the program. The RMSD was taken between the true and predicted ligand after optimally superposing the receptor. Rankings of the first structure with an RMSD of 5Å or lower were between 1 and 600, with several systems having rank 10 or better. The variance of results with the different sets of coordinates showed that the method proved sensitive to the precise starting set of coordinates for a given biological complex. This highlights that for other algorithms (including FTDOCK) it is important to assess the dependency of the results on the precise starting coordinates.

Lenhof[50] has developed an approach for docking based on the identification of points on the surface of each molecule that could be equivalenced in a close-packed docked complex. The search for possible rigid-body models is then sped up by considering which sets of three points on one molecule could be equivalenced to three points on the other molecule. Suitable model complexes are then scored in terms of the geometric match between atoms followed by consideration of the chemical complementarity of the match. Trials starting from unbound components showed that for several, but not all systems, the method yielded lists of a few complexes (< 10), one of which was close (< 4Å RMSD) to the native. This method has been extended to include a treatment of side-chain flexibility in a subsequent screening[51] [1.4.6].

Ausiello[52] has developed a docking procedure named ESCHER. The method starts with shape complementarity based on slices along the z axis of the protein surface mapped to sets of polygons. Complementarity is assessed by the close approach of polygon vertices between the two docked molecules. Steric clashes and charge complementarity are then evaluated. The quality of the results is assessed from the RMSD of the predicted model complex from the best attainable model. Only one true unbound docking system was studied (chymotrypsin / ovomucoid inhibitor), and a good structure (RMSD for the complex < 2.0Å) was obtained at rank three. In addition, similar results were obtained for two complexes in modelling the docking of bound antibody to unbound antigen.

1.4.4 Flexible protein/protein docking

Totrov and Abagyan[53] developed a method that introduces side-chain flexibility early in the search procedure. The approach was applied to the unbound

docking of hen lysozyme to the bound conformation of the combining site of antibody HyHEL5. An initial set of starting conformations are generated that sample space. Then a Monte Carlo search is performed starting with random rigid-body translations (see 1.4.3 for a brief description of the Monte Carlo procedure). Following each translation, side chain torsion angles were allowed to vary, and the energy of the resultant model complex minimised. This procedure was run to identify the set of conformations close to (within 20 kcal/mol) to the best possible model. In this system, this leads to 30 models that were then subjected to extensive energy minimisation, including both rigid-body translations and side-chain rotations. The lowest energy model complex was close the native complex. When the antibody coordinates were superimposed, the RMSD for the C_α, C and N backbone lysozyme atoms was 1.6Å. The primary limitation to this method was its use of bound conformation antibody, therefore requiring the algorithm to introduce conformational movement in only one component of the complex. This makes the method of limited value to a real-world problem. However, a major factor in deciding to show the ability of the algorithm in this limited fashion, must have been resources. The procedure was reported in 1994, and was very time consuming, taking 500 hours (three weeks) of computing time on a state-of-the-art workstation of that time (AXP3000/400).

More recently, the above method has been updated. The method is essentially the same, but a rigid body Monte Carlo search is done prior to the flexible Monte Carlo search.[54,55]

1.4.5 Rigid-body treatment to re-rank model complexes

Evaluation of a scoring function to assess the stereochemistry of model complexes is a central feature of all docking algorithms. However, additional methods can be applied to re-rank model complexes. We will distinguish between whether the re-evaluation treats the model complexes as rigid bodies, or if the procedure includes conformational flexibility to refine the model before re-ranking. This section deals with the former, the next section with the later.

One rigid body method was developed by Jackson and Sternberg,[56] and is referred to as the continuum model. This approach evaluated the electrostatic and hydrophobic energy contributions to a set of putative docked complexes, treating the solvent macroscopically (*i.e.* as a continuum), rather than explicitly including the solvent atoms. The total electrostatic energy of binding involves the loss of interaction between the solvent and each of the protein components independently, followed by the interaction between the two protein components of the complex. The algorithm treated each protein as a low dielectric surrounded by a high dielectric solvent. The electrostatic contributions were evaluated using the program DelPhi.[57] This program maps the protein/solvent

system onto a grid, and then calculates the resultant electrostatic effects considering both the local dielectric (protein or solvent) and the local charge distribution. In the continuum model, the position of polar hydrogens were optimised prior to the calculation of the electrostatic effects.

The hydrophobic effect was quantified in terms of the change to the molecular surface (MS) when the two proteins dock.[58] Generally the hydrophobic effect is quantified in continuum modelling as being proportional to the change in solvent accessible surface (SAS) area,[59] where SAS is the surface traced by the centroid of a hypothetical water molecule (solvent probe) as it rolls along the surface of the protein.[58] Molecular surface can be considered as the surface representing the protein/solvent-probe interface. Previously Jackson and Sternberg[60] suggested that MS provides a better model for the hydrophobic effect than SAS.

The continuum method was applied to re-rank the results on three enzyme/inhibitor systems generated from the docking of unbound components using DOCK.[45] The continuum model was able to identify a near native solutions as having particularly low energies.

Another approach to re-rank model complexes treated as rigid bodies was developed by Robert and Janin.[61] They developed a new soft mean-field potential derived from analysis of protein/protein contacts in crystal structures. A hydrophobic-hydrophobic atom potential was applied to screen putative docked complexes generated by the approach of Cherfils,[44] see 1.4.3 above. Four systems were studied: a reconstitution of the bound components of barnase and barstar, two systems starting with unbound coordinates (β-lactamase / β-lactamase inhibitor, and chymotrypsin / ovomucoid), and one system of bound antibody to unbound lysozyme. For the first three of these systems studied, the lowest free energy model complex was considered a good prediction, being within 2.5Å of the true complex. In the fourth system, such a good solution was found at rank two. Thus the potentials were highly effective in screening for a good model complex in this limited set of systems.

1.4.6 Introduction of flexibility to re-rank putative docked complexes

A method to include side-chain flexibility into the refinement and re-ranking of docked complexes has been developed by Weng.[62] The algorithm was tested on three enzyme/inhibitor systems generated from the docking of unbound components using DOCK,[45] the same as those studied for the evaluation of the continuum model by Jackson and Sternberg[56] (see 1.4.5 above). The conformation of inhibitor side chains buried in the docked complex with the enzyme were examined with an exhaustive conformational search for energetically more stable positions using CONGEN.[63] The resultant models were

then scored by a measure of their relative energetic stability, the function considering the electrostatic interaction between the molecules, desolvation and side-chain conformational entropy. Desolvation was evaluated using the rapid approach of being considered proportional to the change in accessible surface area, where the constant of proportionality depends on the nature of the atom.[64] Side-chain conformational entropy was assumed to be proportional to the change in solvent accessible surface.[65,66] For each model complex, the procedure lead to the identification of a native-like model as having the lowest energy of association. In addition, there was a general improvement in the agreement between the native and predicted geometry of the side-chains whose conformations were adjusted.

A method that provides an extensive sampling of side-chain conformations has been recently reported by Althaus.[51] The set of rigid-body model complexes for screening were generated by the method of Lenhof,[50] described in section 1.4.3. All side chains with rotatable bonds to non-hydrogen atoms that are part of the interface of the model complex are considered. A combinatorial search for favourable orientations is undertaken using computational methods (dead-end elimination and branch-and-bound) to prune the search space and thereby speed up the calculation. The model complexes are evaluated in a method similar to the continuum model of Jackson and Sternberg.[56] The study considered three enzyme/inhibitor systems, for each of which the lowest energy conformation was reported as close to the native complex.

More recent work by Gray and Baker has brought together a wide range of algorithms and scoring functions, many of which have been used separately in other studies. The method[67,68] uses a Monte Carlo search in a rigid-body approximation, followed by a Monte Carlo refinement of the models allowing backbone and side-chain movement. The resulting models are finally ranked by an energy function. This energy function is made up of several calculated values, including van der Waals interactions and a solvation model.

1.5 Blind trials of protein/protein docking

As mentioned above (1.4.1), there are several problems in comparing the success of docking approaches from different groups. In 4.4 there is a full discussion of the different measures that can be used to report the agreement between a predicted and the true complex. The largest problem in comparing between different groups' reported results is the difficulty of translating between the different measures. (It is in fact impossible unless the coordinates of every model are made available.) Another problem is that of optimising an approach until it is successful when working on reproducing the docked structures of known complexes. In addition, a developer may be aware of specific features of

the stereochemistry of the known complexes, and include these features in an algorithm, leading to a bias towards known rather than unknown targets.

In recognition of these problems the docking community has had, and continues to have, blind tests of protein/protein docking. In these evaluations, docking groups were supplied coordinates of the components of a protein/protein complex. The challenge was to predict the structure of the complex prior to its structure being reported.

1.5.1 The Alberta Challenge

The first test, the Alberta challenge, was organised by James and Strynadka in 1996.[69] It involved docking the coordinates of unbound β-lactamase to those of its unbound inhibitor. Six groups submitted entries (see Table 1.4). The results were impressive because all entrants identified as their favoured suggestion a model that had an RMSD for superimposed C_α atoms of no more than 2.5Å. However some groups submitted other entries that were far from the correct structure. Many of the entries used a biological filter requiring that the inhibitor docked to the known active site of β-lactase.

The closest prediction to the true structure, submitted by Eisenstein and Katchalski-Katzir, had an RMSD of all superimposed C_α atoms of 1.1Å, which corresponds to an RMSD for the inhibitor C_α atoms of 4.6Å when the enzyme was optimally superposed. Their successful approach employed the Fourier correlation method developed by Katchalski-Katzir.[14] Their version only considered rigid-body shape complementarity which clearly was sufficient in this system for a successful prediction.

Four other groups each performed a global search and submitted a model between 1.9Å and 2.5Å RMSD over all C_α atoms from the true complex, corresponding to an RMSD for the inhibitor C_α atoms of between 4.0Å and 6.6Å. Two methods were totally rigid-body dockings: the DOCK approach from Shoichet and Kuntz[45] (section 1.4.3), and the comparison of protein surfaces using a smoothed representation from Duncan, Rao and Olson.[70] The model submitted by Janin, Cherfils and Zimmerman[44] (section 1.4.3) started with a rigid-body docking, and then included side-chain optimisation. Only one approach, from Abagyan and Totrov, used a procedure[53] (section 1.4.4) that incorporated flexibility at an early stage of the search procedure.

These results suggest that for this system, the rigid-body approximation is appropriate for docking simulation. In addition, the different approaches to match surfaces yield broadly similar results.

In contrast to the other five submissions that performed a global search, Jackson & Sternberg used an implementation of the continuum model[56] (section 1.4.5) for screening results generated by the DOCK algorithm. They

Group	Number of models	RMSD (Å) of best ranked (Over whole complex)	RMSD (Å) range of other models (Over whole complex)	RMSD (Å) of best ranked (Over inhibitor)
Abagyan & Totrov	3	1.9	11.3 → 16.2	6.6
Duncan, Rao, & Olson	14	1.9	2.0 → 17.7	4.5
Eisenstein & Katchalski-Katzir	3	1.1	13.4 → 14.1	3.4
Jackson & Sternberg	1	1.9	N/A	4.0
Janin, Cherfils, & Zimmerman	4	2.5	2.5 → 16.0	6.1
Shoichet & Kuntz	15	1.8	2.3 → 18.7	3.8

Table 1.4: Results of the Alberta Docking Challenge.[69]

There are two RMSD values used. Over whole complex was calculated over the main-chain atoms for the whole complex. Over inhibitor was calculated over the main-chain atoms for the inhibitor, after optimally superposing the enzyme.

were able to identify a single preferred complex that was close to the native.

1.5.2 CASP2

The second test was as part of the Second Critical Assessment of Techniques for Protein Structure Prediction (CASP2).[71] The target was an antibody/haemagglutinin complex. Coordinates of unbound haemagglutinin and bound antibody were supplied. This was a difficult target given the size of the complex, and only four groups entered. Multiple entries were allowed, to which confidences then had to be assigned such that the total was 100%.

No group submitted any entry that was close to the true complex (Table 1.5). The best averaged prediction was from Vakser.[72] Only one model was submitted and this yielded an RMSD for the interface C_α atoms of the antibody of 9.5Å, calculated after optimally superposing the haemagglutinin. However this prediction did not have any correct contacts. A correct contact was defined as trans-interface residues with atoms that are separated by less than their van der Waals radii plus 1Å. The submission was based on the Katchalski-Katzir Fourier method, implemented for low-resolution search in Vakser's program GRAMM.[35,72] The single prediction that was closest to the native (an RMSD of 8.5Å calculated as before) was also based on the Fourier correlation method, as implemented by the ICRF group of Sternberg, Jackson and Gabb using FTDOCK and MULTIDOCK. However, since the approach did not provide a clear

Group	Number of models	RMSD (Å) mean	RMSD (Å) minimum	Number of Correct Contacts mean	Number of Correct Contacts maximum
DeLisi	2	18.3	15.1	4.5	5
Rees	2	32.3	30.6	0	0
Sternberg	8	20.2	8.5	1.8	8
Vakser	1	9.5	9.5	0	0

Table 1.5: Results of CASP2 Docking Challenge.
The RMSD calculations are of interface Fab C_α atoms after optimal superposition of the haemagglutinins. The interface atoms of the antibody are those within 8Å of the haemagglutinin. Correct residue–residue contacts are defined as where the trans-interface residues have at least one atom–atom distance less than the sum of their van der Waals radii plus 1Å. There were a total of 59 contacts in the true complex. The mean values refer to the weighted scores from all the predictions.

single confident prediction, 8 entries were submitted with associated confidence values, as allowed by the competition. This best single prediction did contain correct contacts. There were two other entries using other approaches. Rees and coworkers used a matching of surfaces using graph theory,[73] and DeLisi and coworkers used an implementation of Weng[62] (see 1.4.6).

From only two blind trials, one cannot draw definitive conclusions. However, in both challenges the Fourier correlation approach of Katchalski-Katzir yielded the best submission, which suggests that it can be considered as a valuable strategy for macromolecular docking.

1.5.3 CAPRI

Occurring after the work in this thesis was completed, the Critical Assessment of PRediction of Interactions (CAPRI) evaluations are a major step forward from the above two tests. 13 targets over 4 rounds of the competition have been assessed to date (April 2004). The number of groups/people submitting entries has varied from 16 to 25 (the latest round). These include both groups who have been in the field for over a decade, and others who are new to it.

The assessment of the quality of structures[74] set three bands of quality of prediction (plus failure). These are shown in Table 1.6. The High quality is a great deal more stringent than most people have used to date in papers to show the success of their algorithms, and could be said to be excellent. The Medium quality is closer to what most would consider a good model. The Acceptable quality is only really acceptable in the sense that it would be a good starting point for further algorithms to refine a model from, and in itself is very poor.

Looking at which groups have done well after 13 targets, it is possible to start ranking the abilities of the algorithms. However, it should be noted that the algorithms are being changed constantly, and specifics of these changes after the second round are not necessarily known. The Abagyan group, Baker group, Camacho and Weng groups have so far modelled four different targets at Medium or High qualities. Ritchie, the Ten Eyck and Wang groups have so far modelled three targets at Medium or High qualities, and ClusPro, the Eisenstein and Sternberg groups two. Other groups have modelled one target to date to such a quality.

The software in this thesis was used by the Sternberg group. Collectively referred to as 3D-Dock, the only additional algorithm to this work was the use of a clustering algorithm before the use of MULTIDOCK. Manual intervention was also used immediately prior to submission. Particular success was achieved with Target 06; alpha-amylase complex camelid antibody VH domain 3. The RMSD was below 1Å, and 56 out of 65 correct interface pairs were modelled.[11]

SmoothDock, used by the Camacho group,[75] is a combination of using

Rank	f_{nat}	L_rms (Å)	or I_rms (Å)
High	≥ 0.5	≤ 1.0	or ≤ 1.0
Medium	≥ 0.3	$1.0 < x \leq 5.0$	or $1.0 < x \leq 2.0$
Acceptable	≥ 0.1	$5.0 < x \leq 10.0$	or $2.0 < x \leq 4.0$
Incorrect	< 0.1		

Table 1.6: Criteria for Ranking the CAPRI Predictions.[74]
Column 1 defines the quality of a prediction. f_{nat} is the fraction of native contacts defined as the number of native residue-residue contacts in the predicted complex divided by the number of native contacts in the target. A contact is defined as when a pair of residues on different sides of the interface have any of their atoms within a 5Å distance. L_rms is the backbone rms displacement of the ligands in the predicted versus the target structures computed after the receptors of these structures have been superimposed. The I_rms is the rms displacement of the interface residues only, in the predicted versus the target complexes. An interface residue is defined as a residue that looses accessible surface area when the two proteins in the target complex associate.

DOT,[14] a desolvation and electrostatics filter, clustering and refinement. It is of particular interest because it does not really represent anything new in terms of algorithms, but shows good success by its combination of already existing tools. Of particular note is the way in which the filters are applied in a logical OR manner. 500 models are allowed through by the desolvation filter, and 1500 by the electrostatics filter. This is aimed at trying to include different modes of association, and the refinement step can still remove models that are unfavourable by either criteria.

1.6 Energy landscape for protein docking

This chapter has considered predictive protein/protein docking. There remains, however, a related question of how two molecules can associate within the time observed biologically. The problem is that association rates for protein/protein docking would be of the order $10^3 M^{-1}s^{-1}$ if they were just governed by diffusion and a correction for orientational constraints. However, observed rates are typically far faster, being of the order of $10^5 M^{-1}s^{-1}$. This large difference is often attributed to long range effects, particularly long range electrostatic steering. Janin[76] has shown that long-range electrostatic steering can enhance association rates by up to 10^5 fold.

Zhang[77] has since modelled the energy surface near the native docked complex in terms of empirical atomic contact energies and Coulomb electrostatic

interactions. The Coulomb interaction has a distance dependent dielectric and a cut-off of 20Å. Thus the energy surface does not consider long-range effects. The study showed that the energy gradient provided by the surface provides a funnel towards the docked structure that increases the probability that an encounter will evolve into the stable complex by about 400 fold. Given the simplified treatment of the interaction energy, in real systems energy funnels could provide even greater enhancements for the rates of association. Thus, even without long-range electrostatic effects, energy funnels provide a possible explanation for the observed relatively rapid association rates. More generally, the role of funnels in directing protein folding and function has been reviewed by Tsai.[78]

1.7 Conclusion

This chapter has shown that there are a variety of algorithms and methods available to tackle protein/protein docking. Although they have had varying success, together they help further development in the field.

Chapter 2 will show the work which constituted the first use of empirically derived residue pair potentials to evaluate the quality of possible models resulting from a docking algorithm.

Chapter 3 contains both some of the first work in docking protein repressors to DNA, and the first report of protein/DNA interface pair potentials.

Chapter 4 displays a thorough testing of the work started in Chapter 2, and also develops better pair potentials.

These chapters do not provide a best solution to the docking problem. However, they do provide a different approach, and this is useful in itself. This was original work, and constitutes a useful addition to the toolkit available for protein/protein docking. Residue level pair potentials have an advantage of being fast to calculate, making it easily possible to incorporate them into other methods that currently do not use them.

Chapter 2

Protein/Protein Docking

2.1 Introduction

The diversity of interactions between residues provides the specificity of recognition in protein folding and ligand binding. A simple model for these interactions is provided by residue/residue pair potentials. These have been widely used to evaluate the stability of protein fold predictions. In this chapter, pair potentials are used to identify a near-native predicted model for a protein/protein complex from decoys of false positives obtained from the FTDOCK rigid-body docking program.

The protein docking problem is to start with coordinates of two molecules in their uncomplexed state and hence predict the structure of the complex. A solution to this problem is becoming increasingly important as the number of experimentally determined protein structures (or protein domains) is increasing rapidly, without the corresponding characterisation of their docked complexes. Advances in computing have lead to the development of several algorithms that tackle the step of exhaustively searching all rigid-body dockings. The approaches primarily match shape complementarity without too many steric clashes. Some then filter on burial of hydrophobic surfaces and/or electrostatic complementarity (see Sections 1.4.2 & 1.4.3).

For many test systems, these approaches generate one or more complexes that are close to the native (typically root-mean-square distance (RMSD) for C atoms of less than 2.5Å at the interface) but also generate several false positives of comparable score to the true positive. The scoring function used during exhaustive searching must be fast to evaluate. However, more sophisticated and time consuming treatments can be applied as a subsequent step of screening a limited set of alternative dockings (see Sections 1.4.5 & 1.4.6).

Strategies that have been explored for a subsequent screening include atomic solvation potentials, empirical functions for atom/atom surface contacts, and continuum models with Poisson-Boltzmann electrostatic calcula-

tions.[45,56,62,79,80] These approaches require a decomposition of the effects stabilising the complex with consequential simplifications and omissions. In addition, the treatment of electrostatic effects is particularly sensitive to atomic positions, so these screening approaches tend to have a limited radius of convergence. Also, some of these approaches are time-consuming and so are less appropriate for screening hundreds of complexes. Therefore there is a requirement for an alternative strategy for screening that is both robust, with respect to the detailed atomic interaction, and fast enough to be applied to a large set of complexes.

These considerations have led us to evaluate the use of residue/residue pair potentials for screening docked complexes. Following earlier work,[81,82] several groups have derived these potentials from frequencies of residue/residue pairs in an appropriate database of experimentally determined protein structures.[83–89] The theory is that by applying Boltzmann's principle to the ratio of observed to expected frequencies of pairings between two residue types one obtains an estimate of the mean force potential between those two residue types. This potential should then incorporate all the pertinent thermodynamic effects, including protein/solvent effects, interresidue van der Waals forces, and electrostatic interactions. The use of residue level (rather than atomic level) potentials provides a smoothness in the energy landscape that is likely to reduce the sensitivity of the function to precise atomic position. In addition, residue/residue potentials are fast to evaluate.

Residue pair potentials are often used in protein fold recognition (*i.e.* threading) to evaluate the fit of a sequence of unknown structure onto a known fold.[83–86,90] In addition, the potentials can be used to evaluate simplified folding simulations,[91] including those on lattices. However, several problems have been identified in simply applying Boltzmann's equation to observed frequencies to obtain a potential of mean force.[92–94] In particular, the difficulty in correctly identifying the random state and the validity of the quasi-chemical approximation that neglects the chain connectivity have been emphasised by some investigators.[92,94] With the quasi-chemical approximation there remains several possible reference states, including one that is purely compositional (mole-fraction) or one that incorporates the differing tendencies of residues to make pairs (contact-fraction).

In this chapter, the problems of deriving potential of mean force from pairing frequencies are acknowledged, and the formalism is simply used to derive a statistical log odds ratio. These log odds were then used to screen docked complexes generated by FTDOCK in the study by Gabb.[15]

This chapter describes the same work as in the 1999 paper by Moont *et al.*.[7] Although chapter 4 greatly extended the work on the datasets used to generate the pair potentials, this work established some basic facts which encouraged

further work. Possibly the most important of these is the ability of pair potentials to avoid low rankings, extreme false negatives, as compared to other scoring methods. It also showed that pair potentials have a usefully large radius of convergence. Finally, it showed that the mole-fraction random model for the expected pairings was the better model to use for pair potentials in docking.

2.2 Methods

2.2.1 protein/protein complexes

All of the methods of residue potentials were evaluated on the set of possible complexes generated by FTDOCK, on the same systems as used in the study of FTDOCK by Gabb[15] (Table 1.2). The ten systems used in that study were a large dataset for that time. It was known that there was at least one correct docking in these lists of possible complexes for 9 out of the 10 systems. A correct docking was described as when the RMSD between the prediction and the experimentally determined complex was 2.5Å or less for the C_α of the interface. The interface is considered to comprise of residues within 10Å of the opposing protein, and the superposition was done using all C_α atoms in the complex. The one system where there was no correct docked structure to be found was not used.

The enzyme-inhibitor systems consisted of the following experimentally determined complexes and components of those complexes (with their PDB codes):

1. CHI, human pancreatic trypsin inhibitor (1apt)[95] / α-chymotrypsinogen (1chg).[96] The PDB code of the complex is 1cgi.[25]

2. CHO, ovomucoid (2ovo)[97] / α-chymotrypsin (5cha).[98] The PDB code of the complex is 1cho.[26]

3. KAI, bovine pancreatic trypsin inhibitor (1bpi)[99] / kallikrein A (2pka).[27] The PDB code of the complex is 2kai.[27]

4. PTC, bovine pancreatic trypsin inhibitor (4pti)[100] / trypsin (2ptn).[101] The PDB code of the complex is 2ptc.[28]

5. SNI, chymotrypsin inhibitor 2 (2ci2)[102] / subtilisin (1sup).[103] The PDB code of the complex is 2sni.[29]

The antibody-antigen systems used consisted of the following F_{ab}'s and F_v's bound to lysozyme (1lza):[104]

1. FDL, D1.3 F_{ab} (1vfa),[105] complex (1fdl).[30]

2. MLC, D44.1 F_v (1mlb),[31] complex (1mlc).[31]

3. HFL, HyHel-5 F_{ab} (2hfl).[32]

4. HFM, HyHel-10 F_{ab} (3hfm).[33]

Native crystal structures for the antibody in 2hfl and 3hfm were not yet solved. Therefore the bound forms of the F_{ab}'s were used in HFL and HFM docking. Only the F_v regions of 1mlb, 2hfl, and 3hfm were used during docking.

2.2.2 Pair potentials

In order to generate an empirical pair potential there were three main considerations; which dataset to use, what 'level' the potentials should be at, and which random model to use in order to calculate the expected values.

Datasets

The ideal dataset for generating a pair potential to be used across a protein/protein interface would be one generated from other such interfaces. However, at the time of the work being done, the number of such interfaces was small. Using a recent study of the time,[106] 11 non-homologous interfaces with resolutions of 2.5Å or better were found. From the same study, 23 homodimer interfaces with resolutions of 2.5Å or better were also found. These two datasets were both tested, though their small size was of evident concern.

The other dataset used was that of a set of non-homologous domains. Although this would result in a pair potential generated from intramolecular pairings, as opposed to the intermolecular pairings across a docked interface, there is evidence that at least some docked interfaces have a composition closer to that of a protein domain core than the average protein surface.[12] The dataset was created by using the Structural Classification of Proteins (SCOP)[107] database (version 1.37). The best resolution structure of each superfamily was taken for each of the superfamilies in the first four fold classes (α, β, α/β, $\alpha+\beta$). Superfamilies where there was no structure with a resolutions of 2.5Å or better were ignored. The dataset totalled to 385 domains, listed in Table 2.1.

Table 2.1: *Dataset of 385 domains used to generate matrices. The range can be either "all" of the PDB file, or a chain identifier, followed by (a colon and) a range, which can either be "all" or the range of residue IDs.*

PDB	range	PDB	range	PDB	range
135l	all	1aac	all	1aba	all
1abr	B:1 – 140	1ads	all	1afw	A:25 – 293

continued on next page

Dataset of 385 domains used to generate matrices. The range can be either "all" of the PDB file, or a chain identifier, followed by (a colon and) a range, which can either be "all" or the range of residue IDs.

continued from previous page

PDB	range	PDB	range	PDB	range
1ah6	all	1ahs	A:all	1aie	all
1aih	A:all	1aij	H:36 - 256	1ail	all
1ajs	A:all	1ak4	C:all	1ako	all
1akz	all	1alk	A:all	1alo	-:81 - 193
1alo	-:194 - 310	1alo	-:311 - 442	1aly	all
1amm	-:1 - 85	1aof	A:134 - 567	1aol	all
1aor	A:211 - 605	1aor	A:1 - 210	1arb	all
1aru	all	1axn	all	1ayl	all
1bam	all	1bco	-:481 - 560	1bdo	all
1beo	all	1ber	A:9 - 137	1bgl	A:220 - 333
1bgl	A:731 - 1023	1bia	-:64 - 270	1bkf	all
1bme	all	1bp1	-:1 - 217	1bpy	A:10 - 91
1brs	D:all	1btk	A:all	1bv1	all
1cei	all	1cem	all	1cex	all
1cfy	A:all	1chd	all	1chm	A:2 - 156
1ciy	-:256 - 461	1cka	A:all	1ckm	A:11 - 238
1cmb	A:all	1cpo	-:1 - 119	1cse	E:all
1cse	I:all	1csh	all	1ctf	all
1ctj	all	1ctt	-:1 - 150	1cuk	-:156 - 203
1cuk	-:65 - 142	1cxs	A:626 - 780	1cxs	A:1 - 625
1cyo	all	1dar	-:283 - 400	1dar	-:477 - 599
1dar	-:600 - 689	1dco	A:all	1ddt	-:381 - 535
1dea	A:all	1der	A:2 - 526	1der	A:137 - 409
1dik	-:510 - 874	1dik	-:377 - 505	1dnp	A:201 - 469
1dnp	A:1 - 200	1dor	A:all	1dos	A:all
1dup	A:all	1ecm	A:all	1efn	B:all
1efu	A:297 - 393	1ema	all	1epn	E:all
1esc	all	1esf	A:1 - 120	1esf	A:121 - 233
1ezm	-:154 - 298	1ezm	-:1 - 153	1fbt	A:all
1fid	all	1fie	A:516 - 627	1fip	A:all
1fjm	A:all	1fnb	-:19 - 154	1fnb	-:155 - 314
1frd	all	1fua	all	1fui	A:356 - 591
1fui	A:1 - 355	1fur	A:all	1fvk	A:65 - 128

continued on next page

Dataset of 385 domains used to generate matrices. The range can be either "all" of the PDB file, or a chain identifier, followed by (a colon and) a range, which can either be "all" or the range of residue IDs.

continued from previous page

PDB	range	PDB	range	PDB	range
1fxd	all	1gad	O:149 - 312	1gar	A:all
1gdo	A:all	1gln	-:306 - 468	1gof	-:1 - 150
1gof	-:151 - 537	1got	B:all	1gpb	all
1gpl	-:337 - 449	1gpm	A:3 - 207	1gpm	A:405 - 525
1gpr	all	1grj	-:2 - 79	1grj	-:80 - 158
1gtq	A:all	1gtr	A:339 - 547	1gua	B:all
1gvp	all	1gym	all	1gzi	all
1ha1	-:8 - 92	1han	-:2 - 132	1hcr	A:all
1hcz	-:1 - 250	1hcz	-:168 - 230	1hiw	A:all
1hoe	all	1hpm	-:4 - 188	1hrd	A:1 - 194
1hsb	A:1 - 181	1hxn	all	1hxp	A:2 - 177
1idk	all	1ido	all	1ifc	all
1igd	all	1ihf	A:all	1iow	-:1 - 96
1iow	-:97 - 306	1isa	A:1 - 82	1iso	all
1jac	A:1 - B:18	1jbc	all	1jet	A:all
1jhg	A:all	1jpc	all	1kap	P:247 - 470
1kid	all	1knb	all	1kpt	A:all
1lam	-:1 - 159	1lba	all	1lbu	-:1 - 83
1ldg	-:164 - 329	1lis	all	1lit	all
1lkk	A:all	1lla	-:2 - 379	1lts	A:4 - C:236
1luc	A:all	1mka	A:all	1mla	-:3 - 307
1mla	-:128 - 197	1mng	A:93 - 203	1mol	A:all
1mrj	all	1msc	all	1msk	all
1mty	B:all	1mty	G:all	1mzm	all
1nba	A:all	1nbc	A:all	1nci	A:all
1nfn	all	1nox	all	1npk	all
1nsy	A:all	1nul	A:all	1nzy	A:all
1oac	A:301 - 724	1oac	A:91 - 185	1oac	A:5 - 90
1obw	A:all	1one	A:142 - 436	1one	A:1 - 141
1onr	A:all	1opd	all	1orc	all
1osp	O:all	1otf	A:all	1oun	A:all
1pam	A:583 - 686	1pau	A:150 - B:401	1pax	-:662 - 796
1pbn	all	1pca	-:4A - 99A	1pda	-:220 - 307

continued on next page

Dataset of 385 domains used to generate matrices. The range can be either "all" of the PDB file, or a chain identifier, followed by (a colon and) a range, which can either be "all" or the range of residue IDs.

continued from previous page

PDB	range	PDB	range	PDB	range
1pdo	all	1pfk	A:all	1pgs	-:4 - 140
1phc	all	1php	all	1phr	all
1pky	A:70 - 167	1pky	A:351 - 470	1plq	-:1 - 126
1pmi	all	1pne	all	1poa	all
1pox	A:183 - 365	1ppn	all	1ppr	M:1 - 156
1prc	C:all	1puc	all	1pud	all
1pya	A:1 - B:310	1qba	-:201 - 337	1ra9	all
1rai	A:1 - 150	1rai	B:1 - 100	1rcf	all
1reg	X:all	1req	A:2 - 560	1req	A:561 - 728
1rge	A:all	1rgp	all	1ris	all
1rla	A:all	1rlr	-:10 - 221	1rlr	-:221 - 748
1rpo	all	1rro	all	1rsy	all
1rtm	1:73 - 104	1rvv	A:all	1scu	A:122 - 288
1sei	A:all	1sfe	-:93 - 176	1sfe	-:12 - 92
1sft	A:2 - 383	1sft	A:12 - 244	1slu	A:all
1smd	-:404 - 496	1smn	A:all	1smp	I:all
1sri	A:all	1sry	A:1 - 110	1stm	A:all
1tad	A:57 - 177	1taf	A:all	1tca	all
1tdt	A:all	1tfe	all	1tfr	all
1thw	all	1tif	all	1tig	all
1tml	all	1tph	1:all	1trk	A:3 - 337
1trk	A:535 - 680	1tta	A:all	1tul	all
1tup	A:all	1tvx	A:all	1tys	all
1tyu	all	1u9a	all	1uae	all
1ubi	all	1utg	all	1uxy	-:3 - 200
1uxy	-:201 - 342	1v39	all	1vao	A:274 - 560
1vcc	all	1vdf	A:all	1vhh	all
1vhr	A:all	1vie	all	1vin	-:181 - 308
1vls	all	1vmo	A:all	1vnc	all
1vom	-:34 - 79	1wap	A:all	1wba	all
1whi	all	1who	all	1wpo	A:all
1xgs	A:195 - 271	1xgs	A:1 - 295	1xso	A:all
1xxa	A:all	1xyz	A:all	1yge	-:150 - 839

continued on next page

Dataset of 385 domains used to generate matrices. The range can be either "all" of the PDB file, or a chain identifier, followed by (a colon and) a range, which can either be "all" or the range of residue IDs.

continued from previous page

PDB	range	PDB	range	PDB	range
1yge	-:1 - 149	1ytb	A:61 - 155	1ytf	B:all
1ytf	C:all	1yve	I:308 - 595	1yve	I:83 - 307
256b	A:all	2abk	all	2arc	A:all
2bbk	H:all	2bnh	all	2bop	A:all
2cba	all	2chs	A:all	2cpl	all
2ctc	all	2dtr	-:65 - 140	2dtr	-:148 - 226
2end	all	2eng	all	2erl	all
2gyi	A:all	2hft	-:1 - 106	2hmz	A:all
2ilk	all	2kau	A:all	2kau	B:all
2kau	C:2 - 475	2kau	C:130 - 567	2mas	A:all
2mcm	all	2nac	A:1 - 374	2ora	-:1 - 149
2pcd	A:all	2phl	A:11 - 210	2phy	all
2pii	all	2reb	-:269 - 328	2rhe	all
2rn2	all	2rsl	A:all	2sil	all
2sns	all	2spc	A:all	2tct	-:2 - 67
2tct	-:68 - 208	2tmd	A:341 - 729	2trc	P:all
2ts1	-:228 - 319	2ts1	-:1 - 217	2tys	A:all
2tys	B:all	2zta	A:all	3bcl	all
3bto	A:1 - 374	3chy	all	3cla	all
3cox	-:319 - 450	3dpa	-:1 - 119	3dpa	-:120 - 218
3grs	-:18 - 363	3grs	-:364 - 478	3mdd	A:242 - 395
3min	A:all	3pmg	A:1 - 190	3pmg	A:421 - 561
3sdh	A:all	3sic	I:all	4aah	A:all
4fgf	all	4pga	A:all	5eas	-:24 - 220
5eas	-:221 - 548	5p21	all	6gsv	A:85 - 217
7acn	-:529 - 754	7acn	-:2 - 528	7rsa	all
8abp	all	8ruc	A:148 - 475	8ruc	A:9 - 147

end

Level

For this work, several different 'levels' of pair potentials were tested. The level is a measure of how specific the pair potential is to individual components of protein structures. The four levels studied here were; (i) residue level using C_β

atoms, (ii) residue level using all atoms, (iii) residue level using all side-chain atoms, and (iv) atom level. This is best clarified by showing how each type of pair potential was generated.

To generate the C_β potentials, the number of pairings between each type of residue were counted. A pair, $p_{i,j}$, was defined as occurring between residues i and j if the C_β atoms in the two residues were within a given distance cut-off (C_α for Gly). In the case of the intramolecular pairings, the pairs were within a domain. In the case of the interfaces and homodimers, the pairs spanned an interface.

For the residue level potential based on all atoms, a pair, $p_{i,j}$, was defined as occurring between residues i and j if any of the atoms in the two residues were within a given distance cut-off. Similarly for the residue level potential based on all side-chain atoms, a pair, $p_{i,j}$, was defined as occurring between residues i and j if any of the side-chain atoms in the two residues were within a given distance cut-off (C_α was counted as side-chain for Gly).

For the atom level potential, each atom on every residue was assigned an atom type. We used 40 atom types, the same used as in a previous study of atom level pair potentials.[108] To generate the potential we then did essentially the same as was done for the residue level potentials. A pair was defined as occurring between atom types i and j if they were within a given distance cut-off.

Random model

There were two methods used for calculating the expected number of pairs between residues i and j. Both assume a quasi-chemical approximation that the amino acids are not parts of connected polymers.[94] One, the mole-fraction method, $e_{(mole-fraction)i,j}$, is proportional to the product of the fractional abundances of the two residues in the pair. The other, the contact-fraction method, $e_{(contact-fraction)i,j}$, is proportional to the propensities of the two residues to be paired with any residue at all. *i.e.*

$$e_{(mole-fraction)i,j} = P \times \frac{n_i}{N} \times \frac{n_j}{N}$$

$$e_{(contact-fraction)i,j} = P \times \frac{p_i}{P} \times \frac{p_j}{P}$$

$$p_i = \sum_{j=1}^{j=20} p_{i,j}$$

$$P = \sum_{i=1}^{i=20} p_i$$

$$N = \sum_{i=1}^{i=20} n_i$$

where n_i and n_j are the total occurrences of each residue.

For all the types of pair potential that were generated, it was considered necessary for the expected value of any given pair to be at least 5. If the expected value was lower than this it was considered to show that there was not enough data to generate a useful pair potential value. Where this occurred, the pair potential value for that pair was made equal to zero. This was not common for most types of pair potential, but did sometimes occur with small distance cut-off values. However, in the case of the atom level calculations using the heterodimer and homodimer datasets, there was not enough data to make even a small number of the expected values large enough to be acceptable. Therefore we did not use those datasets for atom level pair potentials.

The score, $s_{j,i}$, for each pair was then taken as the log fraction of the actual count and the expected count.

$$s_{i,j} = s_{j,i} = \log_{10}(\frac{p_{i,j}}{e_{i,j}})$$

The value of the score for each pair can be considered as a statistical measure of the likelihood of that pair occurring. Since the quantity is a log fraction, the total likelihood for a structure is the sum of all the individual scores. A widely used approach is to equivalence this method to Boltzmann's law,[109] and thereby relate the negative of the log fraction to an estimate of relative free energies for different residue pairings. This was not done for this work, though it would not alter the actual results.

Figure 2.1 shows the 205 different scores of a residue level potential based on all atoms, calculated using $e_{(mole-fraction)}$ with a distance cut-off of 8Å. There is a score value for each type of residue-residue pair, including pairings where the type is identical. The charge-charge interactions have score values of the expected sign (apart from the case of arg-arg), and pairings between hydrophobic residues have generally positive values. This is because hydrophobic residues tend to pair, yet the mole-fraction method does not take into account this information. The values calculated using $e_{(contact-fraction)}$ exhibited this feature to a lesser extent.

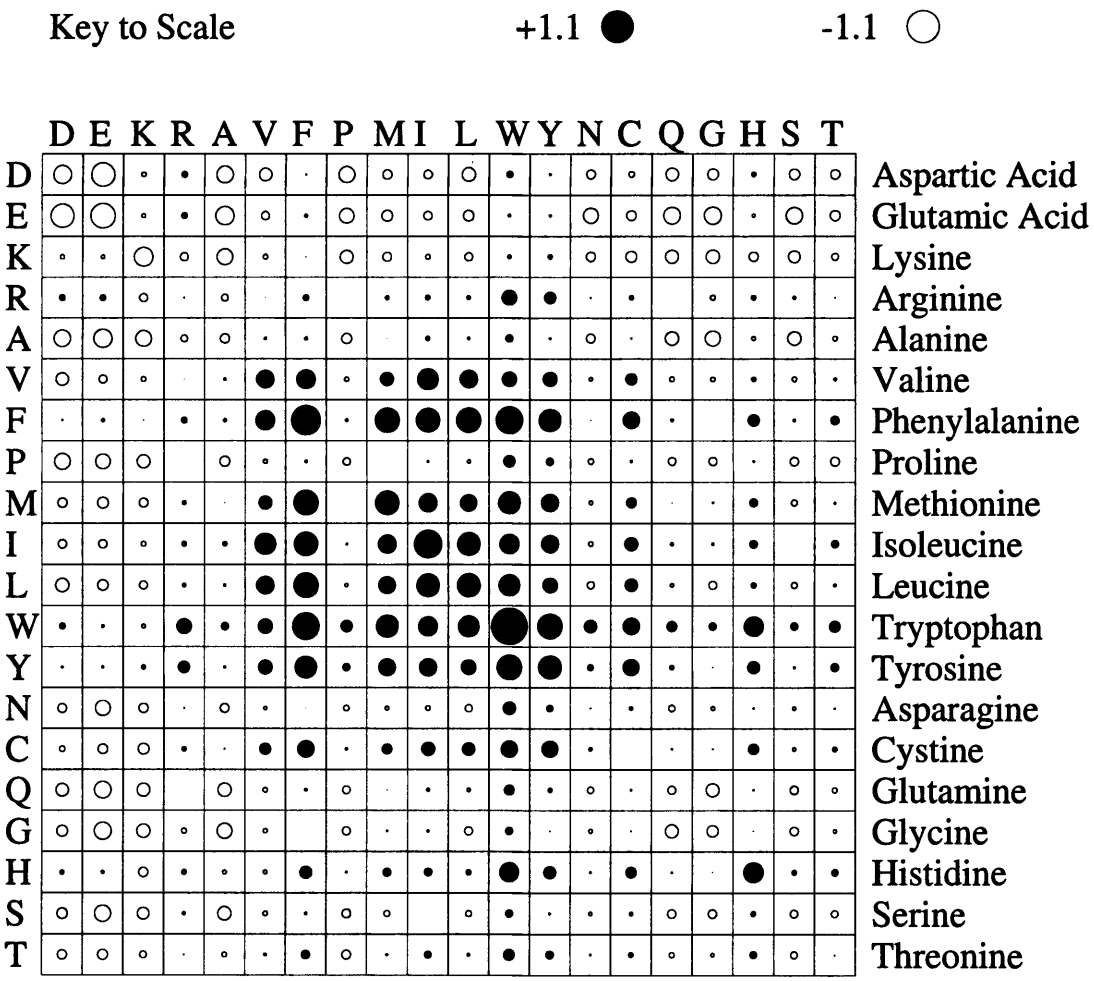


Figure 2.1: Example Pair Potential Matrix.
A graphical representation of an example matrix, generated from 385 SCOP domains, using all atoms and $e_{(mole-fraction)}$ with a 8Å cut-off.

Scoring docked structures

A score was calculated for a complex by summing the appropriate scores of pairs that spanned the interface of the complex. The pairs were considered to exist in exactly the same way as when generating the matrices. That meant that the exact method of scoring was different for each different type of pair potential.

A minimum relative surface accessibility (MRSA) was used as a further constraint on whether two residues are paired.[110] The program used to calculate this value was *naccess*, written by Simon Hubbard when at University College London (present URL is *http://wolf.bms.umist.ac.uk/naccess/*). This was used to assign a relative percentage accessibility value to a reside while in the unbound state, by dividing the accessible area of the residue by its accessibility in a standard conformation. By making the constraint that both residues in the

bound pair have to have had at least a given MRSA while unbound, residues that were not accessible when unbound (either buried or on the surface but largely unexposed) could be ignored.

There were two parameters which were varied to find the optimum values; the distance cut-off and the MRSA. For the residue level potential based on C_β atoms, the distance cut-off was varied from 5 to 15Å at 1Å steps. For the other three types of potential, the variation was from 4 to 10Å at 1Å steps. The difference was due to the first type being essentially a measurement of interresidue distance, whereas the other three are all measuring interatom distance. The MRSA for all methods was varied at values of 0, 5 and 20%.

The list of structures were then sorted according to their pair potential scores, and the positions in the sorted list of correct structures was determined.

Where it was found that the native complex, with a given parameter set, would have less than 20 pairs across the interface, that parameter set was not used. This was because the results were found to be too erratic when less than this number of pairs were involved, especially if a good result was due to, for example, only 3 pairs. This often occurred when the MRSA was high and the distance cut-off was low.

The primary value of interest was the rank at which at least one correct structure could be found in all of the test systems. To enable fair comparison between different test systems, the absolute rank was converted in to a percentage rank. This was because filtering of the output of FTDOCK had resulted in the length of lists differing by over an order of magnitude, from 26 to 762, and an absolute rank would not take this into consideration.

2.3 Results

2.3.1 Screening unbound complexes with pair potentials

Table 2.2 shows all the results using the optimal parameters for each method. The majority of the pair potential methods improve in ranking correct dockings high up the list of complexes compared to the ranking by FTDOCK (which puts a correct docking within 43.8%). The FTDOCK rank is from the shape complementarity value given by the FTDOCK program.[15] Although the FTDOCK algorithm is good at being able to generate a correct docking in a small list of complexes, it is not so succesful at selecting within that list.

The key observations are that the best dataset for any given level or random model of pair potential is intramolecular, and that the mole-fraction random model is better for all levels and datasets. The only exception to this is for the all side-chain atoms pair potential with the homodimer dataset, where the contact-fraction random model is better. It is not clear why this type of pair potential

Type	Percentage Ranks					
	Heterodimers		Homodimers		Intramolecular	
	molar	contact	molar	contact	molar	contact
C_β– C_β	32.3	72.9	33.6	48.4	29.9	35.8
Residue All atoms	22.0	43.4	27.3	46.2	16.3	30.1
Residue Side-chain atoms	21.3	60.7	29.8	18.9	11.8	41.9
Atom - Atom	———	———	———	———	38.7	53.8
	Parameters					
	Heterodimers		Homodimers		Intramolecular	
	molar	contact	molar	contact	molar	contact
C_β– C_β	14Å, 0%	11Å, 0%	14Å, 5%	14Å, 0%	6Å, 5%	14Å, 0%
Residue All atoms	10Å, 0%	7Å, 0%	8Å, 5%	4Å, 5%	6Å, 5%	8Å, 0%
Residue Side-chain atoms	7Å, 0%	5Å, 0%	5Å, 0%	4Å, 0%	6Å, 5%	10Å, 0%
Atom - Atom	———	———	———	———	7Å, 0%	3Å, 5%

Table 2.2: Results for all Datasets, Levels, and Random Models.

The top half of the table shows the percentage down the list at which at least one **good** *(within 2.5Å of the correct structure) complex is found for all the systems, for that dataset, level, and random model. The lower half gives the parameters used; the distance cut-off (in Å) and the MRSA (as a percentage).*

breaks the trend.

Two types of pair potentials performed particularly well; the residue level potential based on all atoms and the residue level potential based on all side-chain atoms. The best result is for the residue level all side-chain atoms type using the intramolecular dataset and a mole-fraction calculated expected. There is very small difference between these two types with the other two datasets using the mole-fraction calculated expected method.

Table 2.2 also shows that the MRSA value of the optimal parameters is either 0% or 5%, but never 20%. This may show that using 20% excludes some still useful pairings. However, since there were results with a 20% MRSA parameter (not shown) not significantly worse than the best parameters shown in the table, it is hard to be sure of making any firm conclusion from this.

Table 2.3 shows the absolute rankings for the best results, namely using the residue level all side-chain atoms potential with a cut-off of 6Å and an MRSA of 5%. Pair potentials substantially improve on the ranking for all the systems, apart from for PTC, where the rank is only just below. Table 2.3 also shows that the pair potentials are better at ranking all the correct structures towards the top. The worst rank they produce is still in the top 40% of the list. Compared to this, FTDOCK can put correct dockings right at the end of the list.

Table 2.4 shows where the actual crystal structure is ranked when it is included into the list of possible dockings produced by FTDOCK. The ranking is that given when using the same method as produced the best results (*i.e.* all side-chain atoms potential with a cut-off of 6Å and an MRSA of 5%). Clearly the rank is rarely the ideal top rank that was desired, and although the ranks for the enzyme-inhibitor systems still seem to be acceptable, the ranks given for the antibody-antigen systems are very bad. This shows that however good the method to produce a list of possible dockings for the pair potentials to evaluate is, the pair potentials used here are on their own unlikely to improve much on what they are currently capable of when given results from the present version of FTDOCK.

Figure 2.2 shows how the values for the best ranks vary with the distance cut-off parameter (MRSA = 5%). It shows that the method is stable around the optimal parameter for the best method. Therefore, even if the optimal parameters chosen using these test systems are not ideal for another system, they should still produce results which are useful.

2.3.2 Control - bound complexes

To investigate how sensitive the pair potential algorithm is to precise atomic positions, the experiment was repeated using the best performing pair potential (side-chain atoms, 6Å cut-off, 5% MRSA, $e_{(mole-fraction)}$), this time using the

System			Rank at which ...			
			first correct solution found		all correct solutions found	
	N	N good	FTDOCK	pair potentials	FTDOCK	pair potentials
CGI	93	1	3	2	3	2
CHO	85	5	11	6	39	23
KAI	349	16	128	13	336	131
PTC	205	7	12	14	145	49
SNI	26	2	8	1	23	4
FDL	636	2	149	75	401	89
MLC	539	4	34	24	493	182
HFL	498	2	218	36	416	153
HFM	700	4	48	6	342	220

Table 2.3: Best Results : Residue Side-chain atoms potential with a 6Å cut-off and MRSA of 5% using $e_{(mole-fraction)}$.

System	N	Rank
CGI	93	5
CHO	85	12
KAI	349	9
PTC	205	5
SNI	26	1
FDL	636	283
MLC	539	259
HFL	498	246
HFM	700	120

Table 2.4: Ranking of Correct Structure : Residue Side-chain atoms potential with a 6Å cut-off and MRSA of 5% using $e_{(mole-fraction)}$.

Key for Tables 2.3 and 2.4: **N** *is the number of possible dockings generated by* FTDOCK *for that system, after biological filtering.* **N good** *is the number of those possible dockings which are within 2.5Å of the correct structure. The parameter values are the distance cut-off (in Å), and the minimum relative surface accessibility (MRSA) as a percentage.*

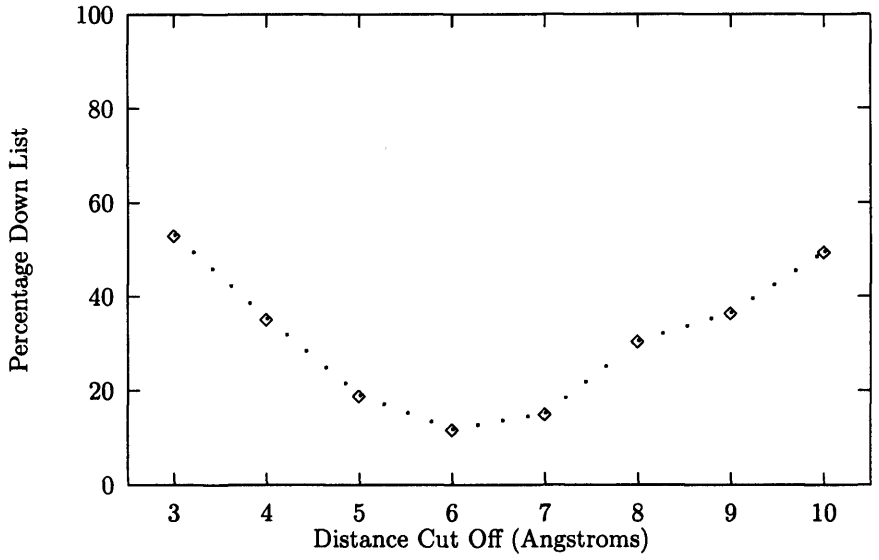


Figure 2.2: Stability of Results around the Minima.

bound forms of the two parts of each system. Table 2.5 shows that the results of the bound forms are no better than those of the unbound. Hence the algorithm is clearly able to cope well with the side-chain flexibility that occurs when two unbound proteins dock.

System			Rank at which ...			
			first correct solution found		all correct solutions found	
	N	N good	FTDOCK	pair potentials	FTDOCK	pair potentials
CGI	123	11	2	1	59	12
CHO	170	9	23	25	138	65
KAI	370	18	24	14	287	188
PTC	410	7	59	47	309	151
SNI	44	11	5	1	42	20
FDL	574	1	210	62	210	62
MLC	464	5	2	48	215	183
HFL	708	5	68	43	299	210
HFM	578	1	94	13	94	13

Table 2.5: Best Bound Results : Residue Side-chain atoms potential with a 6Å cut-off and MRSA of 5% using $e_{(mole-fraction)}$.
For key see Table 2.3.

2.3.3 Combining algorithms

Pair potentials, although producing false positives, never rank a correct structure as completely wrong. This is shown by the all-inclusive rank, which is the lowest rank of any correct docking (Table 2.3). In contrast, the scoring functions of FTDOCK and MULTIDOCK (see below), produce large all-inclusive ranks that are of little use. The MULTIDOCK algorithm, developed by Richard Jackson while in the group, is a simulation which allows for movement of side-chains into lower energy states, once a complex is formed. When run on the same dataset as used in this study,[19] the best ranks were comparable to those given by pair potentials, but the all-inclusive rank showed that correct dockings can be placed as completely wrong.

Accordingly, it was investigated whether the two algorithms could be run sequentially in order to produce a useful combined rank. By cutting the pair potential score ranked list of complexes at a given percentage down the list, it could still be guaranteed that at least one correct docking was still present, so allowing for fewer complexes to be evaluated by MULTIDOCK. MULTIDOCK is a computationally intensive and time consuming program, so this trimming of the number of structures for MULTIDOCK to evaluate has a clear added benefit of greatly reducing the computational requirements. It was decided to cut the list at 25% down the list, over double the length at which a correct docking was always found. This whole combined process is illustrated in Figure 2.3.

From the previous work,[19] the ranks that MULTIDOCK gave for the systems were already available. Once the lists were cut to the smaller sizes, the MULTIDOCK ranks were re-ordered, so giving a combined rank. Table 2.6 shows all the results together. The FTDOCK ranks were not used in the combination, and are for comparison only.

The combined rank clearly improves beyond either pair potentials or MULTIDOCK alone. Due to the order in which the algorithms were applied, it is impossible for the combined rank to be worse than that for MULTIDOCK alone, but it is possible for it to be worse than the rank given by pair potentials alone. This shows in SNI where there is a deterioration in its rank from the pair potential rank. However, overall, the distance down the list of complexes at which all systems have a correct docking is reduced. This is particularly found in the antibody-antigen systems, where there was more scope for improvement. All the systems now have a correct docking within the top 8%, as opposed to the top 12% for pair potentials alone, or top 34% for MULTIDOCK alone.

This therefore shows that a combined approach to filtering a list of complexes can yield better results than single algorithms alone, and that pair potentials are particularly useful in reducing a list of possible complexes.

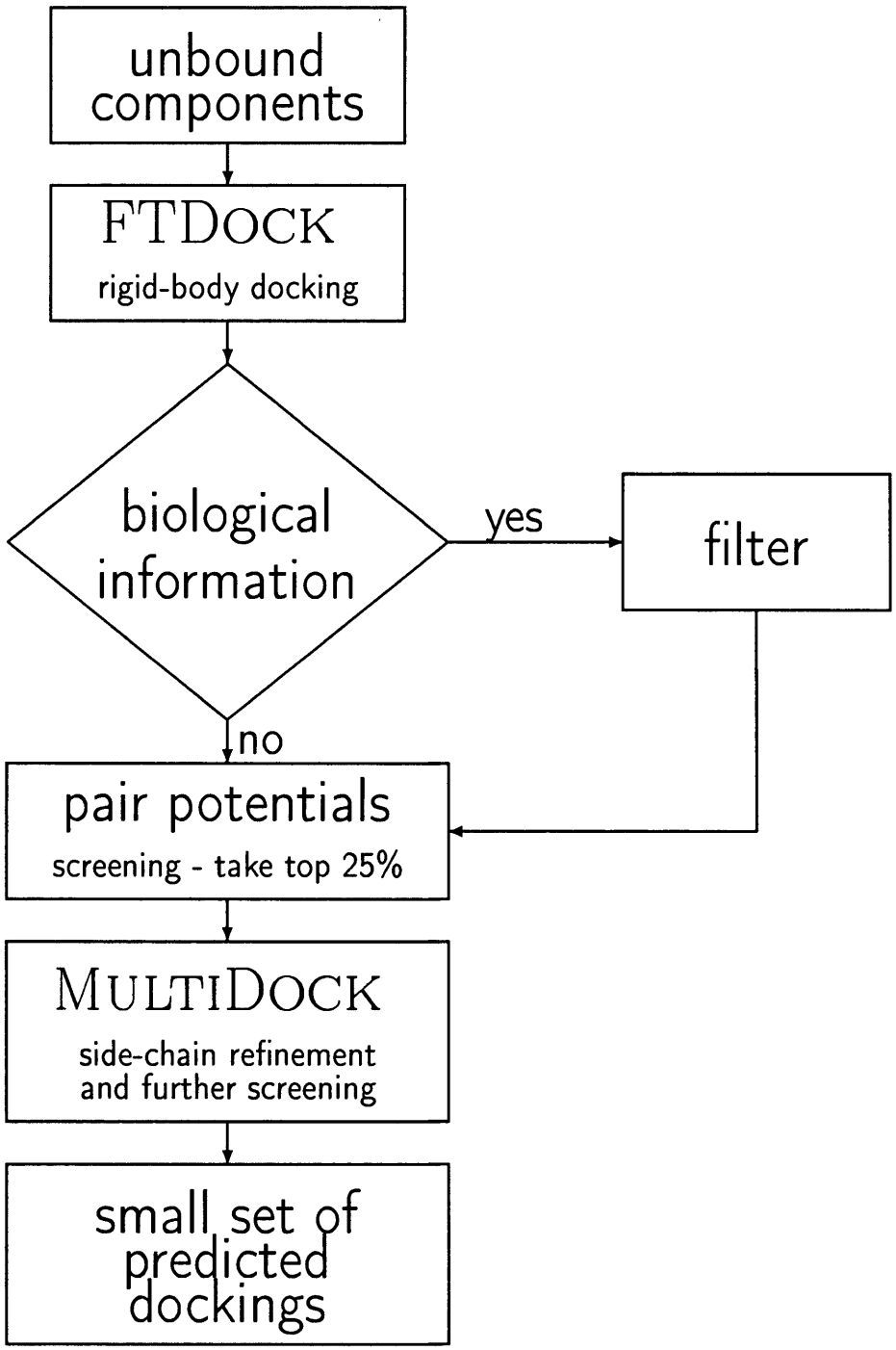


Figure 2.3: Flowchart summarising the combined methods.

System			Rank at which at least one correct solution found			combined	
	N	N good	FTDOCK	pair potentials	MULTIDOCK	N good	Rank
CGI	93	1	3	2	2	1	2
CHO	85	5	11	6	1	4	1
KAI	349	16	128	13	2	12	2
PTC	205	7	12	14	23	7	3
SNI	26	2	8	1	12	2	2
FDL	636	2	149	75	211	2	38
MLC	539	4	34	24	101	2	21
HFL	498	2	218	36	29	1	29
HFM	700	4	48	6	9	2	2

Table 2.6: Best Combined Results : Residue Side-chain atoms potential with a 6Å cut-off and MRSA of 5% using $e_{(mole-fraction)}$.
For key see Table 2.3.

2.3.4 Control - significance of results above random

In some of the systems studied, FTDOCK generates several dockings which are correct. It is possible that in a list of a limited size, with a large number of correct solutions to find, that our algorithms for screening a correct docking from the complexes generated by FTDOCK, though seemingly impressive, were not performing any better than chance. We therefore ran simulations of placing N_{good} solutions in a list of length N, to see what the probability was of obtaining our ranks or better. For each system, the computer simulation was run 10,000 times, and used a random number generator.

Table 2.7 shows that for most systems, the probability of obtaining by chance the observed rank or better is small. As the probabilities are independent of each other, the total probability of getting all these ranks or better together is the multiple of the probabilities for each individual system. Thus the success of each algorithm can be expressed as a single number. These values show that all the algorithms are well above random, and there is a clear progression to better values in the order FTDOCK $\Longrightarrow$ MULTIDOCK $\Longrightarrow$ pair potentials $\Longrightarrow$ combined method. If for a given algorithm, the probability of getting the rank given for any system was 50/50 (*i.e.* a coin toss), then the combined probability of success in 9 systems would be $(0.5)^9$, which is 0.002. All of our algorithms do better than this.

System			Probability of getting rank			
	N	N good	FTDOCK	pair potentials	MULTIDOCK	combined
CGI	93	1	0.031	0.023	0.023	0.023
CHO	85	5	0.504	0.311	0.058	0.058
KAI	349	16	0.999	0.466	0.092	0.092
PTC	205	7	0.352	0.401	0.579	0.099
SNI	26	2	0.530	0.082	0.724	0.159
FDL	636	2	0.417	0.218	0.554	0.114
MLC	539	4	0.230	0.166	0.567	0.144
HFL	498	2	0.690	0.144	0.116	0.116
HFM	700	4	0.256	0.034	0.051	0.010
Total Probabilities			$5 * 10^{-5}$	$2 * 10^{-8}$	$1 * 10^{-7}$	$4 * 10^{-11}$

Table 2.7: Probabilities showing significance of results.
For key see Table 2.3.

System	FTDOCK		MULTIDOCK		pair potentials		combined	
	RMSD	PCP	RMSD	PCP	RMSD	PCP	RMSD	PCP
CGI	11.61	0.0	6.05	20.0	7.96	0.0	6.05	20.0
CHO	8.29	17.4	1.52	73.9	6.20	21.7	1.52	73.9
KAI	7.13	11.1	4.85	5.6	6.28	5.6	4.85	5.6
PTC	5.98	10.5	6.57	5.3	7.79	0.0	5.00	21.1
SNI	5.98	4.8	7.52	9.5	1.56	28.6	8.49	4.8
FDL	8.59	0.0	12.96	0.0	9.46	11.1	4.68	22.2
MLC	12.20	0.0	9.28	21.1	9.86	10.5	10.04	0.0
HFL	10.68	15.4	10.27	0.0	12.92	0.0	10.27	0.0
HFM	17.94	0.0	17.70	3.2	10.09	0.0	13.53	3.2

Table 2.8: False Positives : RMSD (Å) and Percentage Correct Pairs for top ranks. Key: **RMSD** *RMSD (in Ångstroms) from crystal structure.* **PCP** *Percentage Correct Pairs compared to crystal structure. Pairs considered up to distance of 6Å. Pair Potential is Residue Side-chain atoms potential with a 6Å cut-off and MRSA of 5% using* $e_{(mole-fraction)}$.

2.3.5 False positives

Table 2.8 shows the RMSD and percentage correct pairs for the top ranked structure for each method. It can be seen that FTDOCK does not have any top rank with both less than 6Å RMSD and more than 20% correct pairs. Both MULTIDOCK and pair potentials have one top rank with an RMSD of less than 6Å and more than 20% correct pairs. The best method by these criteria is the combined method with three such top ranks.

The reasons why the false positive structures are ranked at the top are not consistent. Since all the structures in the ranked list have been filtered, there is for the four trypsin complexes, at least one residue of the catalytic triad is in the interface, and for the antibody complexes, at least a single antigen residue in contact with the H3 or L3 CDRs. However, compared to any lower ranked incorrect structure, there is no increase in hydrophobic pairs across the interface, or salt bridges for the top rank. Overall, there is no discernible single reason why the false positives are ranked as they are.

2.3.6 Relationship of score to correct pairs

Looking at all the systems together, an initial examination for a relationship between rankings and RMSD for the structures being screened was performed for each of the algorithms; FTDOCK. MULTIDOCK and pair potentials (RPDOCK). None was observed. However, when the percentage of correct pairs formed by the known complex, which are then found in the complex, were considered instead of RMS, relationships were observed (Figures 2.4, 2.6 and 2.5). A pair was considered to exist between two residues if any of the side chain atoms in the two residues were within 6Å of each other. There is still no discernible relationship with shape complementarity ranks from FTDOCK. However, pair potentials now show a clear relationship, even when the percentage of correct pairs is as low as 50%. MULTIDOCK also shows a relationship, which although not as good overall, is in fact better when the percentage of correct pairs is above 70%. This shows that though pair potentials have a larger radius of convergence than MULTIDOCK, once a complex is very near to the correct solution, MULTIDOCK will select it to a higher rank.

These results are consistent with the level of representation used in the modelling. Only surfaces are considered by FTDOCK, pair potentials are at the residue level, while MULTIDOCK is at the atomic level. The order in which the two rankings are integrated to produce the combined rank is a consequence of their different relationships and radii of convergence. For protein complexes with limited conformational change on association, this stepwise refinement, from a discretised molecular representation via residue pair potentials to an atomic representation, provides a useful strategy to predict docked protein complexes.

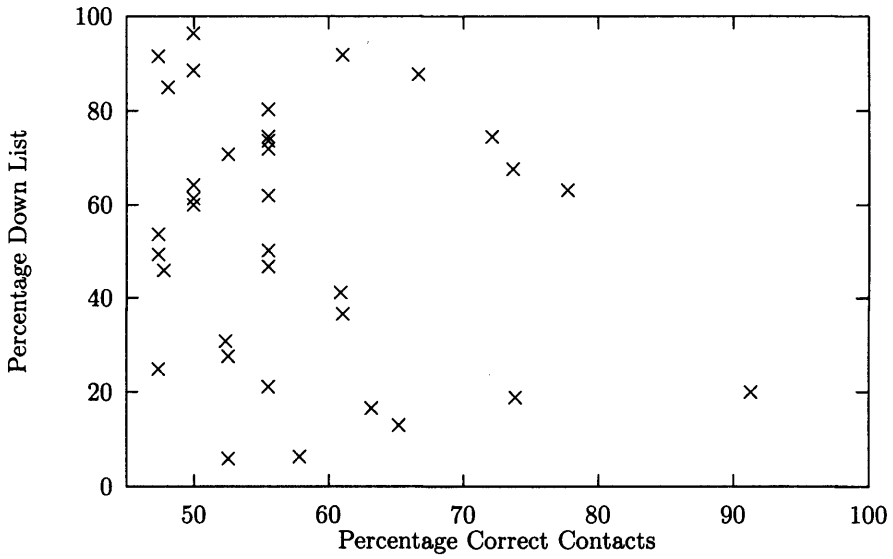


Figure 2.4: Relationship of FTDOCK *ranks to percentage correct contacts.*

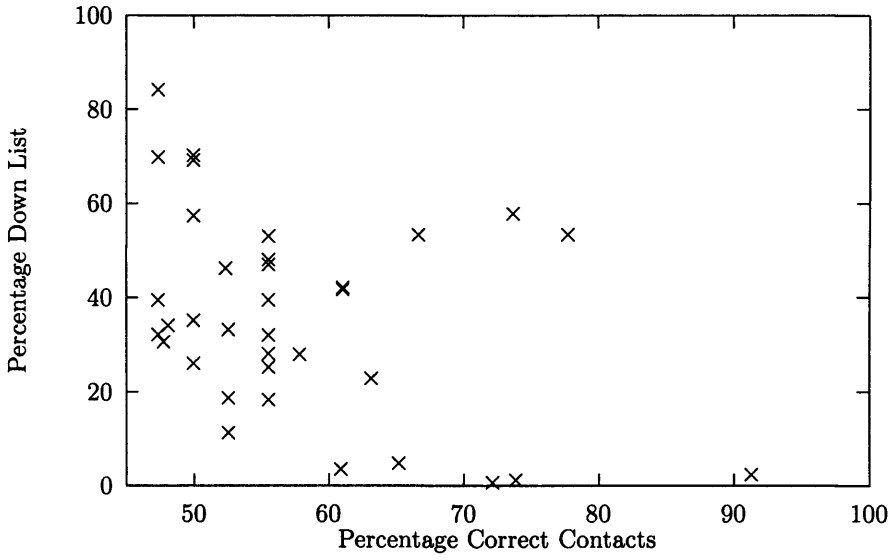


Figure 2.5: Relationship of MULTIDOCK *ranks to percentage correct contacts.*

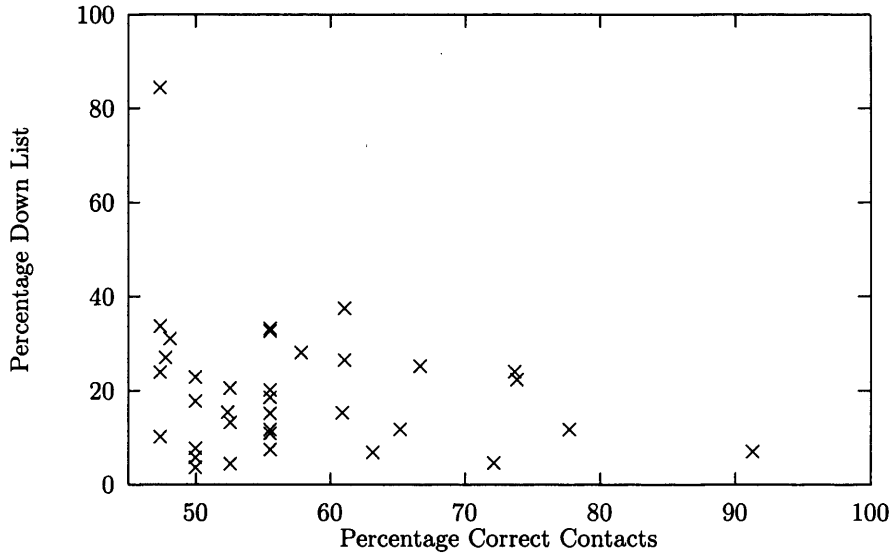


Figure 2.6: Relationship of RPDOCK *ranks to percentage correct contacts.*

2.4 Discussion and Conclusion

The key conclusion of this study was that pair potentials have considerable power in correctly selecting correct dockings from a list of complexes. Although the scoring function also produces false-positives, pair potentials can position a correctly docked complex at or near the top of a score-ranked list. Possibly more importantly, the pair potentials avoid low ranking, extreme false negatives, as compared to the other scoring methods.

This study also shows that of the various models proposed for the random state required in generating pair potentials, it is the mole-fraction method which should be used when using pair potentials across an interface. However, the best strategy for docking need not necessarily apply to the use of pair potentials within a single domain, as is done for threading.

Of the various datasets from which pair potentials were generated, it was clear that the best results came from the intramolecular pairings in a database of non-homologous protein domains. Why this is so is not evident. It could be presumed that pair potentials generated from intermolecular pairings should be a better method, and the reason they were not so here was due to the small size of the datasets available at the time. This is particularly true of the heterodimer dataset, which should most closely represent the propensities of intermolecular pairings in complexes, yet at the time was limited to 11 structures.

The results for the enzyme-inhibitor systems were clearly better than those for antibody-antigen systems. This problem is not unique to pair potentials, and as was discussed above is possibly due to the differing recognition stereo-

chemistry and lower affinity of the antibody-antigen interactions, compared to those for enzyme-inhibitors.[111] However, the results even for the antibody-antigen systems were still significantly more useful than relying only on the shape complementarity as calculated by FTDOCK.

Finally, it has been shown that as part of a combined approach, the large radius of convergence makes pair potentials useful in screening large numbers of structures before more detailed all-atom refinement procedures are used.

Chapter 3

Protein/DNA Docking

3.1 Introduction

The characterisation of the three-dimensional structure of a protein/DNA complex provides major insights into the stereochemistry and biochemistry of recognition and gene expression.[112,113] However, the determination of the structure of the unbound protein and knowledge of the cognate DNA sequence can precede structural knowledge of the complex. Consequently, computational methods can be useful to model protein/DNA associations for structure prediction to probe the stereochemistry of recognition. As a step toward a general approach for protein/DNA docking, this study considers transcriptional repressor complexes[114–123] as an appropriate starting system. There are several experimentally determined structures, revealing a variety of recognition modes. Although they exhibit conformational changes on association, this study aimed to apply to protein/DNA interactions the previously successful strategy used for protein/protein interactions. Those interactions also involve conformational changes, and the algorithms had proved capable of accommodating those changes.

The objective is to start with the crystal or nuclear magnetic resonance structure of an individual repressor. The docking simulation will then be done with a standard B-DNA model containing the recognition sequence. This presents several problems. The number of rotatable bonds in the system makes it at present unfeasible computationally to explore the 6 degrees of associational freedom together with explicit modelling of the conformational changes. Thus one must start by docking the static molecules (*i.e.* rigid body docking) and employ a scoring function that can accommodate some degree of overlap. This softness in the scoring function approximates induced binding conformational changes. The scoring function must also evaluate the electrostatic stability including the cationic screening of the highly charged sugar-phosphate backbone.

These problems presumably have hindered the development of protein/DNA docking protocols. Kaptein and coworkers[124,125] developed a Monte Carlo simulation program (MONTY) to model repressor/DNA interactions. They considered flexibility of the protein side chains together with limited deformations of the DNA and explored docking of unbound repressor to model DNA. The systems considered were repressors that bind with the α-helix/turn/α-helix in the major groove such as the 434 cro protein. The majority of their simulations probed the local specificity of interactions and consequently the repressor was correctly oriented within the DNA major groove and shifted by $\pm$ 2 base pairs. A further simulation was aimed at distinguishing two alternate orientations of the repressor α-helix within the DNA. In general, the MONTY program was capable of retrieving the correct repressor/DNA complex with many native interactions reproduced within the limited space of the search performed.

Campbell[126] explored the specificity of hydrogen-bond recognition in protein/DNA complexes. The bound coordinates of the protein were used and residues with hydrogen-bonding possibilities within or near to the true binding site considered. The protein was docked as a rigid body to a model DNA with limited flexibility. When the study was restricted to consider only those base pairs that have at least two strong hydrogen bonds with the protein, the procedure was able to identify at rank one the biologically correct DNA sequence.

These previous studies did not therefore tackle the overall objective of performing a complete search of protein/DNA binding space starting with both sets of coordinates in their unbound states.

The work described in this chapter is the same as published in the 1998 paper by Aloy *et al.*.[8] The work can be divided into two independent parts. The first was to use FTDOCK to perform a series of rigid body dockings between proteins and DNA fragments. Due to lack of experimental data, although the protein structures were from unbound crystallographic data, the DNA fragments were modelled. In addition, the electrostatics used in FTDOCK had to be further developed in order to be useful. This work was done mainly by Patrick Aloy and Henry Gabb. The second was to screen the results from the first part with pair potentials specific to protein/DNA interactions, and was done by myself, Gidon Moont. The pair potentials were calculated from a database of crystallographic protein/DNA complexes. This database did not include any protein homologous to any of the proteins that were used in the docking evaluation.

3.2 Methods

3.2.1 Repressor/DNA molecules

Eight repressor/DNA complexes (systems) were examined (see 3.3 below), representing the available systems (excluding close homologues of the repressors) in the PDB at the time (May 1998). Five of the repressor structures (CRO, GAL, LAC, LAM, PUR) have the α-helix/turn/α-helix motif, and show major grove recognition on the DNA. Two involve a two-stranded anti-parallel β-sheet recognising bases in the major groove (ARC, MET). One (TRP) has the repressor recognising bases via the minor groove, but also interacts with the major groove and the DNA backbone. This means that the method was tested on three different binding modes.

Unbound coordinates were used for the repressors apart from LAM, where only Cα coordinates were available and so the bound coordinates were used instead. The structures for the DNA sequences corresponding to those in the complexes were constructed starting from standard B-DNA[127] geometry, and then energy minimised using the JUMNA program.[128]

3.2.2 Rigid body docking

A global search of rigid body docking was performed by FTDOCK as described previously. The grid size was set at $128 \times 128 \times 128$, resulting in grid cell sizes ranging from 0.51Å to 0.87Å. In all the systems, the DNA was defined as being the static molecule. The surface was set at 1.2Å, rather than the previously used 1.5Å. The rotational angle step for the mobile repressor molecule was set at 12°, as opposed to the previously used 15°, yielding 12,661 non-degenerate rotations.

The correct treatment of electrostatics had previously been found to be essential to successful docking of protein/protein complexes using FTDOCK. For these simulations, the Coulombic electrostatic field of the DNA (the static molecule) was evaluated for each grid cell, while the repressor (the mobile molecule) had its charged atoms discretised among the closest 8 grid cells (Figure 1.6). Initial work used the AMBER charge set,[21] but gave poor results on trials with CRO. There was too much noise from partially charged Carbon atoms, causing problems such as masking the charge groups that actually contribute to specificity. A specific charge set, shown in Table 3.1, was therefore developed to calculate the electrostatic field of the DNA. Sequence specific recognition tends to occur through the bases rather than the sugar-phosphate backbone. However, the phosphate groups are highly charged, and can so mask the partially charged atoms within the helix grooves. This was overcome by a distance cut-off for calculating the field. The field strength contribution from a charged atom was

Atom type	Charge
All phosphorous atoms	-0.25*e*
The purine N7	0.10*e*
The pyrimidine O2	-0.25*e*
cytosine-N4	0.25*e*
thymine-O4	-0.25*e*
guanine-N2	0.25*e*
guanine-O6	-0.25*e*
adenine-N6	0.25*e*

Table 3.1: Charges assigned by FTDOCK *to the DNA, used for electrostatic complementarity calculations.*

calculated only for grid cells within 2Å of that atom, using a sigmoidal dielectric function.[17]

The use of these values effectively dampened the phosphate charges while exaggerating the partial charges of the chemical groups in the helix grooves, thus enhancing sequence recognition by the repressor proteins.

In the previous work for protein/protein interactions, partial charges were assigned to the main chain, but only fully charged side chains were considered. For this work, the protein charges were as before (Table 1.1), except that Asn, Gln, and His, known to be important in DNA recognition, were also considered. The additional charges are shown in Table 3.2

All the above values and parameters used by FTDOCK, the surface thickness, rotational angle and electrostatic charges, were developed by modelling the CRO system, starting with the unbound repressor and modelled DNA. These parameters were then applied when modelling the docking of the remaining seven systems.

The top 4,000 structures were kept from each experiment, and examined

Residue	Atom	Charge
Asn	OD1	-0.25*e*
Asn	ND2	0.25*e*
Gln	OE1	-0.25*e*
Gln	NE2	0.25*e*
His	ND1	0.25*e*
His	NE2	0.25*e*

Table 3.2: Additional charges assigned by FTDOCK *to the protein repressors, used for electrostatic complementarity calculations when docking to DNA.*

for good solutions. Each run took approximately one day of computational time on a single Silicon Graphics R10000 processor. By the use of a Silicon Graphics parallel Challenge machine, with its own parallel fast Fourier transform routines, a run took about 7 hours on four R10000 processors.

3.2.3 Geometric filters

The structures generated by the global search were filtered by distance constraints. Each amino acid had an effective side-chain length, L, (ranging from 0.5Å for Gly to 6.0Å for Arg), as in the previous work. Distances, D, were calculated between the Cα atoms of the amino acids and the nucleotides (base glycosidic N) of the DNA bases. The initial filter (filter 0) removed dockings that were artifacts of the repressor docking to the terminus of the DNA fragment. A docking was excluded if either end of the fragment had both nucleotides at that end with a distance to a Cα, D, less than the corresponding L for that amino acid.

The next filter (filter 1) was based on the DNA footprinting information. In most cases this information would be available before a docking simulation was attempted. The two central base pairs of the footprint were identified from biochemical references (see 3.3). A docking was passed by the filter if there was at least one amino acid for which $D < L+4.5$Å to any one of the four nucleotides of the two central base pairs.

The last filter (filter 2) considered that there may be information defining which amino acids on the repressor interact with DNA. This would typically be obtained from phylogenetic studies or from mutagenesis. A list of these residues was obtained from the available literature (see 3.3). A docking was passed by the filter if any one of those residues satisfied $D < L+4.5$Å to any nucleotide in the DNA fragment.

3.2.4 Quality of predicted complexes

Two measures were available to evaluate the agreement between the predicted dockings and the experimental structure; root mean square deviation (RMSD) of the atomic positions, and percentage correct contacts (%CC) across the interface. To calculate the RMSD values is a standard procedure, and was calculated using the `profit` program. The calculations were limited to using the C_α atoms in the repressors and C1' atoms in the DNA fragments. To calculate the %CC values it was first necessary to define the interface regions. This was done by finding all the amino acid / nucleotide pairs in the experimental structure which had at least one non-hydrogen atom-atom distance < 5Å. For each of these pairs the C_α-C1' distance was measured in both the experimental structure and the predicted docking, and if the difference in the distances was <

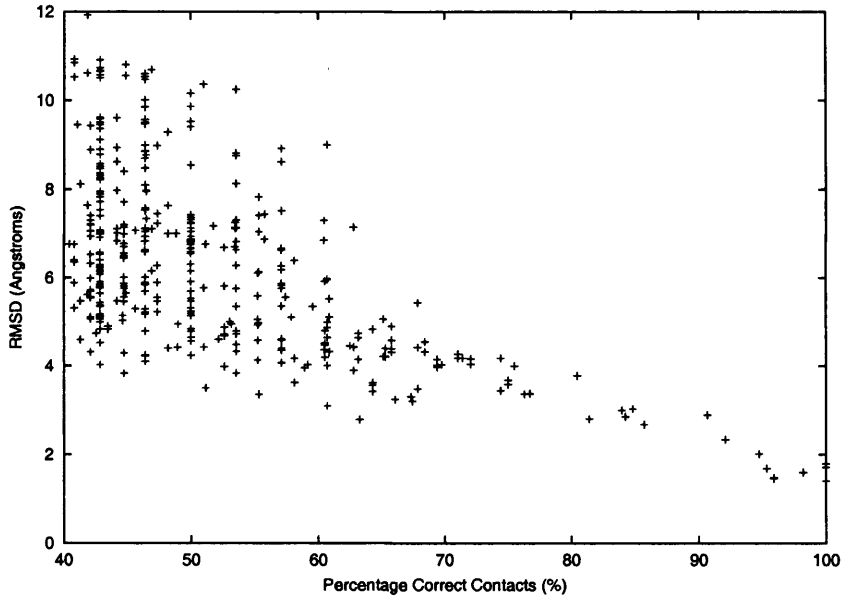


Figure 3.1: Relationship of RMSD values to Percentage Correct Contact values.

4Å then a correct pairing was considered to have been modelled in the predicted docking. %CC gives the number of correct pairings as a percentage of the total pairs found in the experimental structure.

Figure 3.1 shows a plot of these two values calculated for all the predicted dockings that passed through filter 0. As can be seen, there is a very good relationship between %CC above 65% and RMSD values, with a RMSD of 5.5Å or lower guaranteed. However, Figure 3.2 does not show a simlarly good relationship for RMSD values of 5.5Å or lower. For this reason, it was decided that a predicted docking with a %CC value of 65% or higher would be considered 'good'.

The highest value for %CC attainable from rigid body docking was evaluated for each test system. This was done by seperately superimposing the model DNA and the unbound repressor onto the experimental structure using the C_α atoms in the repressor and C1' atoms in the DNA. For six of the eight test systems this yeilded a %CC value of 100%. For MET and GAL the highest attainable values were found to be 91% and 96% respectively.

3.2.5 Pair potentials

Protein/DNA complexes were identified from the Nucleic Acid Database (NDB)[129] (*http://ndbserver.rutgers.edu/*). From the list of entries, non-identical peptide

chains complexed with at least ten nucleotides were selected. A non-redundant set of repressor/DNA complexes were developed by taking the best resolved structure from repressor homologues with more than 25% identity over the entire sequence. In addition, the eight test systems and their homologues (> 25% identity) were excluded from the non-redundant set. The PDB codes of the resulting 20 complexes are given below (3.3).

An empirical amino acid / nucleotide pairing score was then derived. The distance between the C_β (C_α for Gly) to the base glycosidic N between each amino acid and each nucleotide was calculated within each of the 20 complexes. The number of pairs of amino acid type a and nucleotide type b (p_{ab}) having a distance less than a given cut-off, d_{cut}, were then counted. In order to derive a probability for any such pairing, a random model is required. Two models for a random state were considered, both involving the quasi-chemical approximation that assumes that the amino acids and nucleotides are not parts of connected polymers.[94] The first model is a molar-fraction random state that is based purely on composition. Let n_a and n_b be the total occurrences in the whole dataset of amino acids of type a and nucleotides of type b, and P the total number of all pairings, *i.e.*

$$P = \sum_{a=1}^{a=20} \sum_{b=1}^{b=4} p_{ab}$$

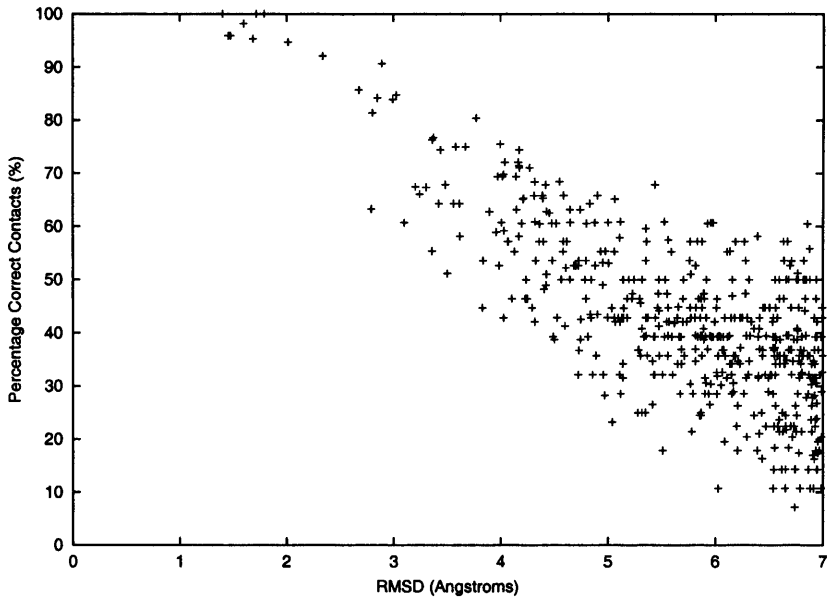


Figure 3.2: Relationship of Percentage Correct Contact values to RMSD values.

Then the molar-fraction expected pairings are given by

$$e_{ab} = P \times \frac{n_a}{\sum_{a=1}^{a=20}} \times \frac{n_b}{\sum_{b=1}^{b=4}}$$

The second model is based on the observed pairings and is proportional to the tendency of a given amino acid to make a pair with any of the 4 nucleotides, and the tendency of a given nucleotide to make a pair with any of the 20 amino acids. The expected pairings are given by

$$e_{ab} = P \times \frac{\sum_{a=1}^{a=20} p_{ab}}{P} \times \frac{\sum_{b=1}^{b=4} p_{ab}}{P}$$

From these expected values, a log-odds score for a pairing is given by

$$S_{a,b} = \log_{10}(\frac{p_{a,b}}{e_{a,b}})$$

A $S_{a,b}$ value was only calculated if $p_{ab} > 0$ and $e_{ab} \geq 5$, to ensure against nonsense values created by small numbers. A total score for a complex was obtained by summing the $S_{a,b}$ values for all the amino acid / nucleotide pairs with a distance less than the distance cut-off, d_{cut} (raw score). Values for scoring a complex were also calculated by dividing this value by the numbers of pairs counted (pair normalised). In addition, a sparse form of the $S_{a,b}$ values was considered, in which interactions (pairs) involving any hydrophobic residue type were excluded (*i.e.* only using C, D, E, G, H, K, N, Q, R, S, and T). We chose to evaluate the complexes using a log odds ratio rather than applying Boltzmann's principle and converting the ratio to a potential mean force, as the validity of this approach has been questioned,[93,94]

3.3 Structural data

The structural data is shown as: complex name (capitalised); repressor name; PDB code of complex (reference); PDB code of unbound repressor (reference); region of peptide used; DNA sequence generated, with the two bases used in filter 1 underlined; peptide residues used in filter 2. The regions and filters use the notation chain(residue-code)residue-number. In the case of only one chain being present, the notation will shorten to (residue-code)residue-number.

1. ARC; arc repressor; 1par;[114] 1arr;[130] chain C, D(P)8–(E)48; ATAGTAGAGTG; C(Q)9, D(Q)9, C(R)13, D(R)13;

2. CRO; cro repressor-operator; 3cro;[131] 2cro;[116] all; AAGTACAAACTTT; (K)27;

3. GAL; CD2-GAL4 DNA binding domain; 1d66;[117] 125d;[132] A(E)8–(K)43; CCGGAGG; A(Q)9, A(R)15, A(K)17, A(K)18, A(K)20, A(C)21, A(K)23;

4. LAC; lactose operon repressor; 1lbg;[118] 1lqc;[133] A(P)3–(P)49; AATTGTGAGCG; A(Y)17, A(Q)18, A(N)25, A(H)29;

5. LAM; LAM phage repressor-operator N-terminal domain; 1lmb;[134] 1lrp;[134] chain 4, and used bound form of the protein since only C_α available for the unbound protein; GGCGGTGATAT; 4(K)4, 4(N)55;

6. MET; met repressor-operator; 1cma;[135] 1cmc;[135] chains A and B; TTAGACGTCT; A(K)23, B(K)23, A(T)25, B(T)25;

7. PUR; pur R repressor-operator; 1pnr;[122] 1pru;[122] A(T)3–(S)46; ACGAAAA; A(K)5, A(S)14, A(T)16, A(T)17, A(H)29, A(N)23, A(R)26;

8. TRP; trp repressor-operator; 1tro;[136] 2wrp;[137] G(S)5–(A)105; TGTACTAGTTAAC; G(Q)68, G(R)69, G(L)71, G(K)72, G(G)78, G(I)79, G(A)80, G(T)81, G(T)83, G(G)85;

The PDB codes of repressor/DNA complexes used to generate the potentials were; 3mht, 2bop, 1lat, 1zaa, 1ber, 1hcr, 1pdn, 1per, 1vol, 1ihf, 1fjl, 1apl, 1pue, 1bhm, 1ign, 1tsr, 1ytb, 1nfk, 1rva, 1eri.

3.4 Results

3.4.1 Rigid body docking and distance constraints

Each of the test systems was run through FTDOCK in the standard fashion, using surface complementarity to rank the some 10^{10} dockings, and with the electrostatics acting as a binary filter. The top 4000 dockings were stored and then put through each of the three filters in turn. Table 3.3 shows the results after each of the three filters with ranks calculated using the surface complementarity score from FTDOCK (as normal). The ranks calculated using the electrostatic scores calculated by FTDOCK are also shown (though these are in fact the ranks within the top 4000 dockings as ranked by surface complementarity, not the rankings from the 10^{10} dockings). Each of the lists of dockings were also scored by the empirical pair potentials, and the ranks calculated from those scores are in the rightmost column. After Filter 0 a good docking is ranked at 140 or better for seven out of the eight systems using the surface complementarity score. These good dockings have %CC ranging from 65% to 85% with corresponding RMSDs between 4.3 and 3.0 Ångstroms. In the list there are several good dockings and the best for each system has

Complex	No. of solutions	No. of good solutions *i.e.* with %CC > 65	Rank of first good solution evaluated by shape complementarity [rank(%CC/RMSD(Å))]	electrostatics [rank (%CC/RMSD(Å))]	empirical pairing score [rank (%CC/RMSD(Å))]
Filter 0 - disallow repressor at ends of DNA					
ARC	2972	14	140(69/4.1)	69(75/4.0)	1(69/4.0)
CRO	3010	2	28(85/3.0)	1815(85/3.0)	220(80/3.8)
GAL	2941	7	55(75/3.6)	298(75/3.6)	2(75/3.6)
LAC	3299	7	88(72/4.0)	960(65/5.9)	302(77/3.4)
LAM	3175	8	38(84/3.0)	32(98/1.6)	4(98/1.6)
MET	2938	none	no solution	no solution	no solution
PUR	2876	33	11(68/4.3)	129(100/1.8)	30(92/2.3)
TRP	2854	9	15(65/4.2)	35(81/2.8)	17(67/3.2)
Filter 1 - use information about DNA bases with filter 0					
ARC	1232	11	91(69/4.1)	42(75/4.0)	1(69/4.0)
CRO	570	2	12(85/3.0)	387(85/3.0)	121(80/3.8)
GAL	1470	4	37(75/3.6)	220(75/3.6)	2(75/3.6)
LAC	800	6	30(72/4.0)	561(65/5.9)	133(77/3.4)
LAM	889	6	22(84/3.0)	10(98/1.6)	4(98/1.6)
MET	1017	none	no solution	no solution	no solution
PUR	1444	13	9(68/4.3)	101(100/1.8)	28(92/2.3)
TRP	564	6	4(65/4.2)	12(81/2.8)	1(67/3.2)
Filter 2 - use information about DNA bases and repressor residues with filter 0					
ARC	219	11	22(69/4.1)	6(75/4.0)	1(69/4.0)
CRO	11	1	3(85/3.0)	7(85/3.0)	9(80/3.8)
GAL	789	4	26(75/3.6)	133(75/3.6)	2(75/3.6)
LAC	270	6	13(72/4.0)	188(65/5.9)	117(77/3.4)
LAM	141	6	3(84/3.0)	5(98/1.6)	4(98/1.6)
MET	40	none	no solution	no solution	no solution
PUR	732	13	2(68/4.3)	55(100/1.8)	28(92/2.3)
TRP	104	6	2(65/4.2)	7(81/2.8)	1(67/3.2)

Table 3.3: Rank of Solutions, Starting With Unbound Structures.
After each of the three filters, the first column gives the complex, the second column the number of solutions left in the list of the top 4000 complexes generated from FTDOCK (N), and the third column gives the number of good solutions, i.e. with %CC > 65. The final three columns give the rank of the first coorect solution followed by its %CC and RMSD(Å). Rankings were calculated using shape complementarity from FTDOCK, electrostatic score from FTDOCK, and the empirical score for nucleotide/amino acid pairings (EPS). No solution with %CC > 65 was generated for MET.

%CC between 75% and 100%, corresponding to RMSDs between 3.6 and 1.4 Ångstroms (Table 3.4).

The MET docking fails to find a solution with %CC $\geq$ 65%. This in part stems from the inability to model the interaction of the DNA with a repressor loop that moves towards the DNA on binding. This loop movement limits to 91% the maximum attainable value for %CC resulting from an optimal superposition of the bound complex with the unbound repressor and model DNA (Table 3.4). Rigid body methods are therefore unable to model the surface contact between this mobile loop and the DNA.

Filter 1 is based on the central two base pairs and provides a substantial discrimination. The number of dockings to be examined in a ranked list to find a good docking is 91 or less when ranking by the surface complementarity score (excluding MET), and for six out of the eight systems is less than 40. The further

	Best possible complex from		
	unbound coordinates (superposition) %CC/RMSD(Å)	unbound coordinates (FTDOCK) %CC/RMSD(Å)	bound coordinates (FTDOCK) %CC/RMSD(Å)
ARC	100/1.1	100/1.7	100/0.5
CRO	100/1.0	85/3.0	100/0.3
GAL	96/2.1	75/3.6	100/0.6
LAC	100/1.0	77/3.4	N/A
LAM	100/1.1	100/1.4	100/0.4
MET	91/3.3	60/5.3	100/0.2
PUR	100/1.5	100/1.8	100/0.6
TRP	100/1.0	98/2.2	77/2.5
TRP + WAT	-	-	100/0.9

Table 3.4: Agreement between Model-Built and X-Ray Structures.
For each complex, and for TRP together with bound waters (TRP + WAT), the correct % correct contact (%CC) / RMSD (Å) are given for three models of the complex. First the best possible model complex that could be generated from unbound coordinates is given, based on optimal superposition of the unbound components onto the bound components of the complex. The next column is that of the best model generated by FTDOCK, *starting with the unbound components. The last column is that of the best model generated by* FTDOCK, *starting with the bound components. N/A denotes that* FTDOCK *was not run for LAC with the bound coordinates, due to only a* C_α *trace existing for the repressor in the bound complex.*

Complex	RMSD DNA All(Å)	RMSD DNA C1(Å)	RMSD PROT All(Å)	RMSD PROT C_α(Å)
ARC	2.3	1.8	1.9	1.1
CRO	2.2	1.7	1.6	0.6
GAL	2.3	1.8	3.0	2.0
LAC	3.2	2.8	-	2.3
LAM	2.3	1.8	-	-
MET	2.7	2.5	2.6	2.2
PUR	2.0	1.3	2.1	1.4
TRP	2.0	1.6	2.2	1.9

Table 3.5: RMSD for Superimposed Bound and Unbound Molecules. The superpositions were performed separately for the DNA and for the protein molecules.

specification of one repressor residue interacting with the DNA (Filter 2) yields a list of no more than 26 dockings to be examined (excluding MET) and for four systems a list of the top three solutions ranked by surface complementarity would include a good docking.

Figures 3.3 through to 3.10 show a superposition of the native complex with the highest ranked good docking, represented by the repressor C_α trace and the DNA as a phosphate backbone with schematic bases. Figures 3.11 through to 3.18 show the same superpositions, but with just the repressors with their side chains drawn. For MET the best available model is shown. In all the systems, including MET, the model reproduces the principal recognition mode such as a helix or a β-sheet fitting into the major groove. No consistent errors such as the molecules being always too distant or too close were apparent. The figures also highlight the extent of conformational change to the molecules on association. In particular, the DNA is substantially distorted from ideal geometry in several of the systems. In addition, a combination of rigid body shifts and the change in conformation results in some, though far from all, of the side chains showing substantial changes in position between the predicted and the X-ray structures. The conformational change on association for each system is quantified in Table 3.5, which details the RMSD between the experimental complex and model DNAs, and the RMSD between the experimental complex and unbound repressors. It was coping with these conformational changes that presented the challenge in developing a viable computational strategy.

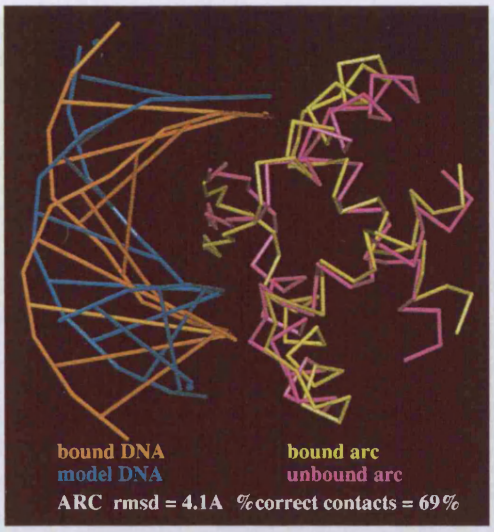


Figure 3.3: Superposition of native and 1st correctly modelled complexes for ARC : arc repressor.

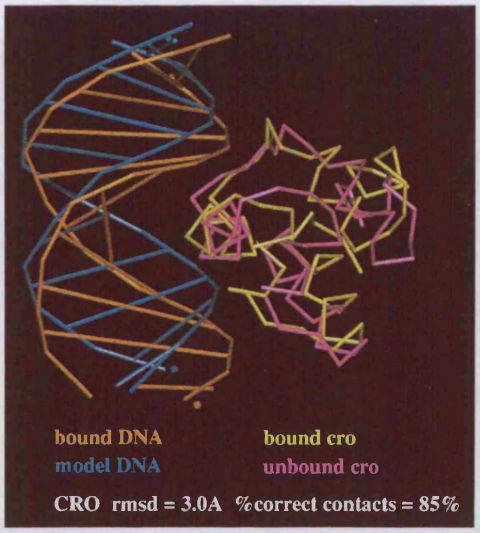


Figure 3.4: Superposition of native and 1st correctly modelled complexes for CRO : cro repressor-operator.

Key to Figures 3.3 and 3.4.
Superposition of native and predicted complexes for repressor (C_α trace) and DNA (phosphate backbone with lines for base pairs). The first correct modelled complex in Table 3.3 (column 4) is shown. See Section 3.2.4 for explanation of fitting values. Diagrams were generated by PREPI (http://www.sbg.bio.ic.ac.uk/prepi/).

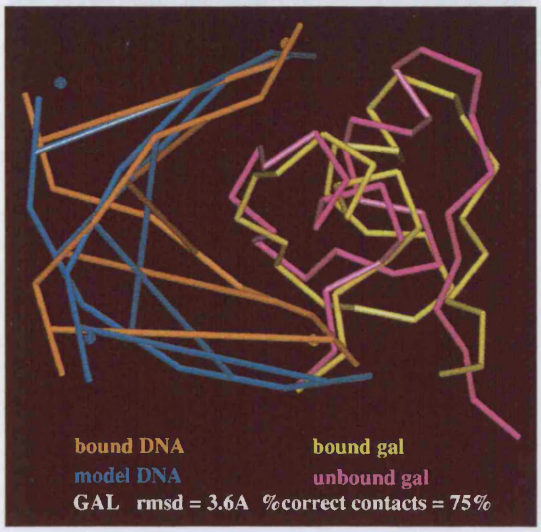


Figure 3.5: Superposition of native and 1st correctly modelled complexes for GAL : CD2-GAL4 DNA binding domain.

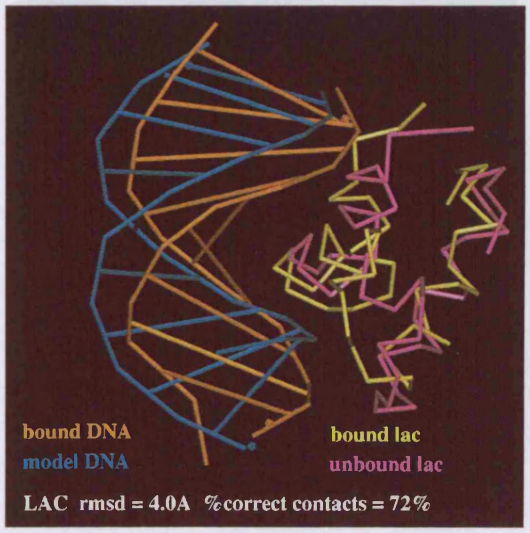


Figure 3.6: Superposition of native and 1st correctly modelled complexes for LAC : lactose operon repressor.

Key to Figures 3.5 and 3.6.
Superposition of native and predicted complexes for repressor (C_α trace) and DNA (phosphate backbone with lines for base pairs). The first correct modelled complex in Table 3.3 (column 4) is shown. See Section 3.2.4 for explanation of fitting values. Diagrams were generated by PREPI.

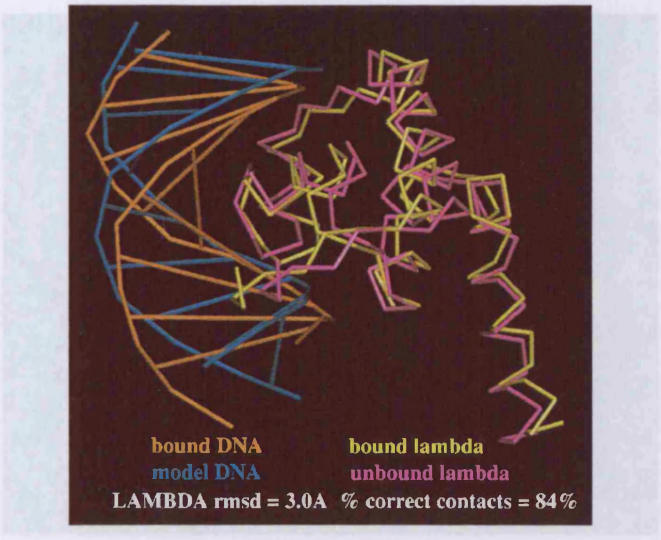


Figure 3.7: Superposition of native and 1st correctly modelled complexes for LAM : LAM phage repressor-operator N-terminal domain.

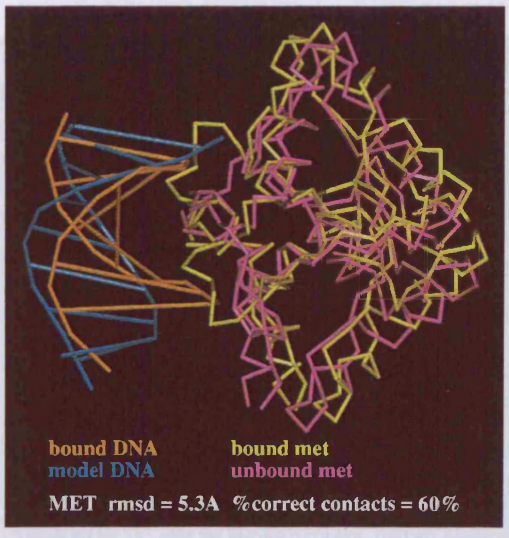


Figure 3.8: Superposition of native and 1st correctly modelled complexes for MET : met repressor-operator.

Key to Figures 3.7 and 3.8.
Superposition of native and predicted complexes for repressor (C_α trace) and DNA (phosphate backbone with lines for base pairs). The first correct modelled complex in Table 3.3 (column 4) is shown for LAM, and the best prediction for MET. See Section 3.2.4 for explanation of fitting values. Diagrams were generated by PREPI.

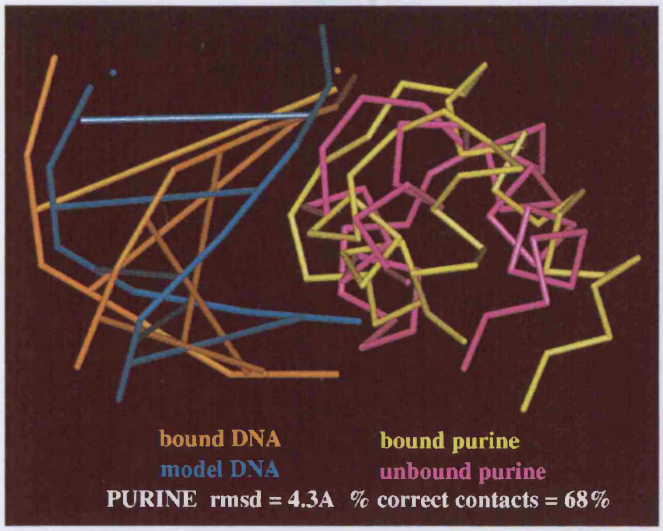


Figure 3.9: Superposition of native and 1st correctly modelled complexes for PUR : pur R repressor-operator.

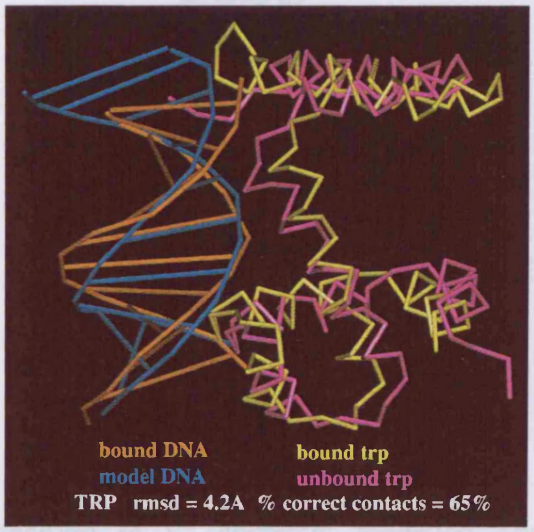


Figure 3.10: Superposition of native and 1st correctly modelled complexes for TRP : trp repressor-operator.

Key to Figures 3.9 and 3.10.
Superposition of native and predicted complexes for repressor (C_α trace) and DNA (phosphate backbone with lines for base pairs). The first correct modelled complex in Table 3.3 (column 4) is shown. See Section 3.2.4 for explanation of fitting values. Diagrams were generated by PREPI.

Figure 3.11: Superposition of 1st correctly modelled and best modelled complexes for ARC : arc repressor.

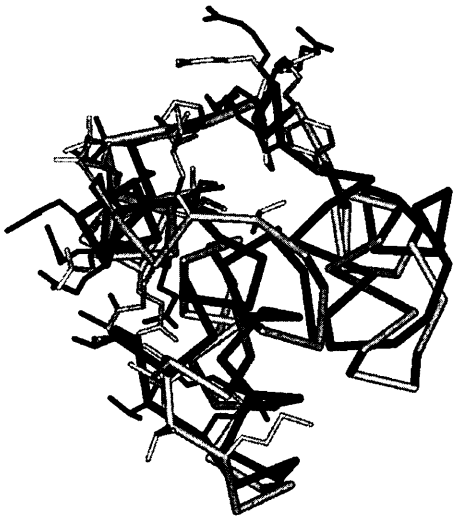


Figure 3.12: Superposition of 1st correctly modelled and best modelled complexes for CRO : cro repressor-operator.

Key to Figures 3.11 and 3.12.
Superposition of native (black) and predicted (grey) complexes for repressor (C_α trace with side chains). The first correct modelled complex in Table 3.3 (column 4) is show along with the best prediction for MET. Diagrams were generated by PREPI.

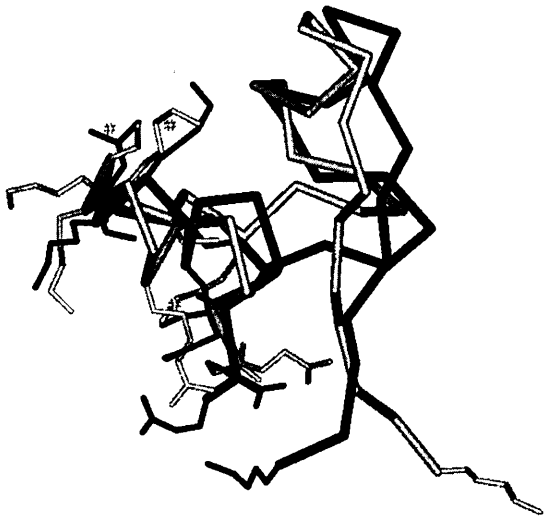


Figure 3.13: Superposition of 1st correctly modelled and best modelled complexes for GAL : CD2-GAL4 DNA binding domain.

Figure 3.14: Superposition of 1st correctly modelled and best modelled complexes for LAC : lactose operon repressor.

Key to Figures 3.13 and 3.14.
Superposition of native (black) and predicted (grey) complexes for repressor (C_α trace with side chains). The first correct modelled complex in Table 3.3 (column 4) is show along with the best prediction for MET. Diagrams were generated by PREPI.

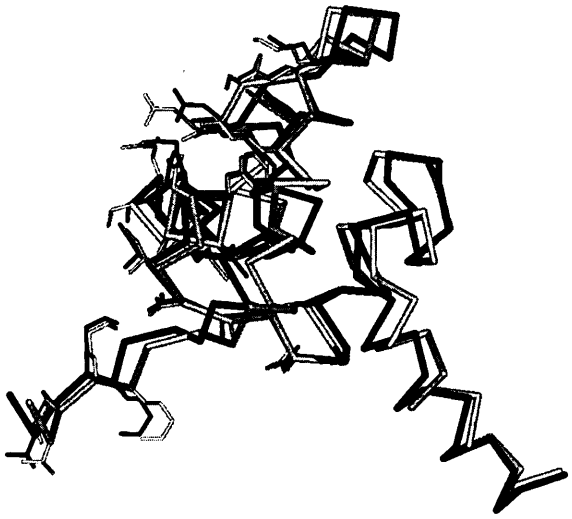


Figure 3.15: Superposition of 1st correctly modelled and best modelled complexes for LAM : LAM phage repressor-operator N-terminal domain.

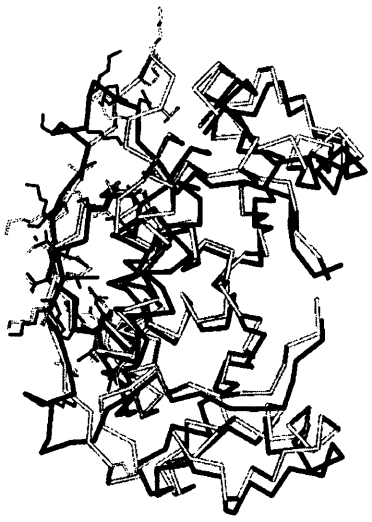


Figure 3.16: Superposition of 1st correctly modelled and best modelled complexes for MET : met repressor-operator.

Key to Figures 3.15 and 3.16.
Superposition of native (black) and predicted (grey) complexes for repressor (C_α trace with side chains). The first correct modelled complex in Table 3.3 (column 4) is show along with the best prediction for MET. Diagrams were generated by PREPI.

Figure 3.17: Superposition of 1st correctly modelled and best modelled complexes for PUR : pur R repressor-operator.

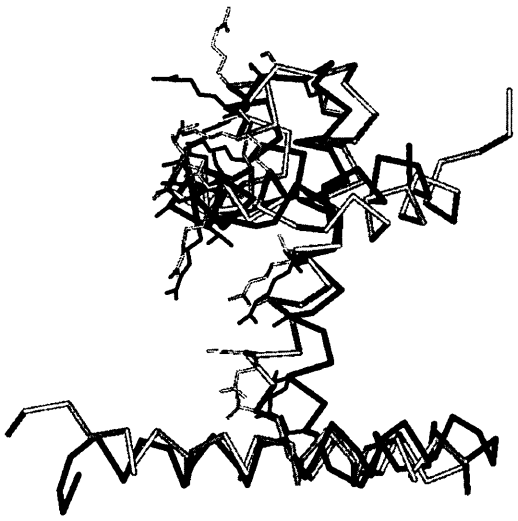


Figure 3.18: Superposition of 1st correctly modelled and best modelled complexes for TRP : trp repressor-operator

Key to Figures 3.17 and 3.18.
Superposition of native (black) and predicted (grey) complexes for repressor (C_α trace with side chains). The first correct modelled complex in Table 3.3 (column 4) is show along with the best prediction for MET. Diagrams were generated by PREPI.

3.4.2 Empirical scoring of amino acid / nucleotide pairings

The use of empirical pairing scores to identify a good solution from false positives was evaluated on dockings that were allowed after filter 1. To decide on the best method for screening the lists of models, eight different methods of scoring dockings by an empirical log-odds score matrix for amino acid / nucleotide pairings were considered. The values for the distance cut-off, d_{cut}, explored were from 10 to 20 Å in 1 Å steps. For each approach we took the values for the number of structures (N) that would need to be examined in the score ranked list to identify a single good docking for all M out of the 7 systems (M = 1–7). MET was excluded as there was no good docking. There is a trade off between the number of dockings N (how few alternatives need to be examined) and M (how many systems can yield a good docking in the top N dockings). We considered that when the procedure is practically used for experimental design, N can be no more than 5, and accordingly identified the optimal parameters to yield the maximum M for $N \leq 5$. The best approach used was the un-normalised score with a sparse matrix using molar-fraction expected and d_{cut} of 12 Å, corresponding to N=4 and M=4.

Table 3.3 presents the results of ranking by the empirical score in the final column. After filter 1, for the seven systems that can be considered (*i.e.* excluding MET), in four (ARC, GAL, LAM, TRP) a good solution was ranked four or better, and for ARC and TRP a good solution was the top rank. In PUR the solution was at a middle rank (28) whereas for CRO and LAC more than 100 dockings would have to be examined before a good one would be found. In four systems, pairing scores improved the ranking of the first good docking, compared to the ranking by shape complementarity. In the remaining three the ranking was poorer.

Table 3.3 also gives the results of re-ranking by the pair potentials after filter 0. For three systems (ARC, GAL, LAM), good dockings were in the top 5 ranked scores, which is a marked improvement over ranking by shape complementarity. Again, for CRO and LAC, the ranking by pair potentials was poorer than by shape complementarity. PUR and TRP gave similar ranking from the two scores. This shows that just using FTDOCK followed by ranking by the empirical pair potential score can yield a very small list of dockings, one of which is a good docking, but with only a success rate of three out of eight. Nevertheless, this level of accuracy can be useful to suggest subsequent experiments. The results of screening by the empirical score after filter 2 are also given in Table 3.3. These show little improvement over the ranking after filter 1.

The magnitude of the empirical score provides a guide to the confidence that can be placed in a high ranking docking. For six of the eight systems the maximum empirical pair potential score was < 30. However, for ARC and LAM,

when the dockings were ranked by pair potential score (after filters 0 and 1), the highest ranked good docking had a score > 35. This suggests that when the answer in a study is unknown, if the highest values of the pair potential score are > 30, then one can have confidence that the list will have a good solution at a high rank. However, the converse is not true – a list with scores < 30 can still have a good solution at a high rank.

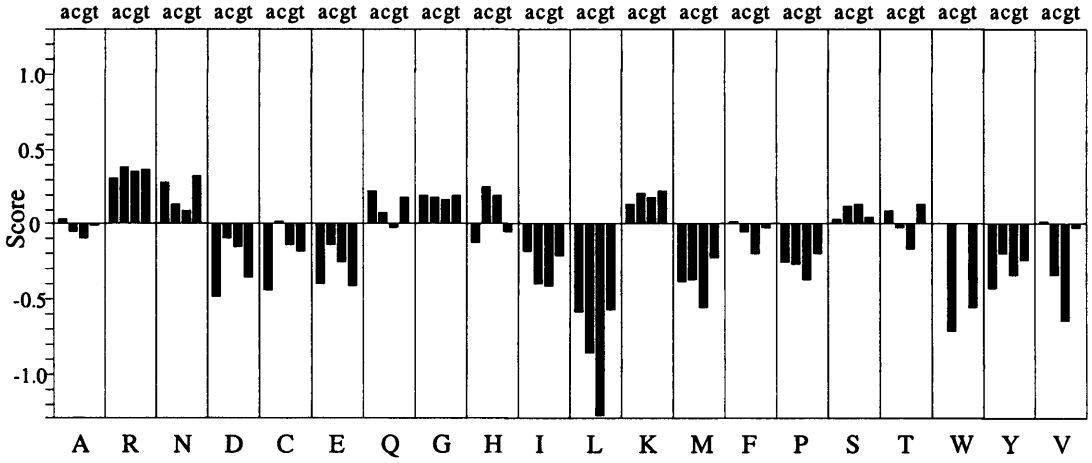


Figure 3.19: Empirical amino acid / nucleotide pairing scores.
Values of S_{ab} for amino acid pairing with the four nucleotides (acgt), derived using molar expected frequencies with a distance cut-off of 12Å. All amino acid scores are shown although only the sparse matrix was used in scoring dockings. due to absence of data there are no values for TRP (W) with bases c and g.

The empirical score for all amino acid / nucleotide pairings (molar fraction and d_{cut} of 12Å) are shown in Figure 3.19. Residues with favourable scores for interactions with nucleotides are, as expected, the positively charged Arg, His and Lys, the amides Asn and Gln, Ser and Thr with a hydroxyl group (but not Tyr), and Gly, which allows the close approach of the main-chain of the protein to the DNA. Some amino acids show a discrimination between AT and GC bases. In particular, the preference of Gln for AT could reflect the opportunity for the amide to make hydrogen bonds with both the acceptor and a donor of the base pair via atoms accessible in the major groove. Lustig and Jernigan[138] have previously derived an empirical scale of amino acid / nucleotide interactions from a series of zinc fingers interacting with DNA. Their values differ from ours. For example, in their values, Arg and Lys form unfavourable interactions with some bases. These differences stem from their use of only zinc fingers and their use of an expected frequency based on contact fraction.

3.4.3 False positives

To obtain further information on the modelling implemented in this strategy, we visually examined, using a graphics package, the structures of the five top ranked dockings after filter 1, when ordered by either shape complementarity or by empirical pair potential score. We identified three "native-like" distinguishing features in the false positive dockings (Table 3.6). *Shifted* is when the false positive repressor is docked in the same mode to the DNA helix as in the experimental complex, but the site of the interaction was shifted along the DNA helix. The second, *Rotated*, is when the false positive is rotated by roughly 180° about an axis perpendicular to a plane that defines the interface region compared with the correct solution (*e.g.* an α-helix that lies in the major groove still sits in the groove, but runs in the opposite direction). The third is referred to as *correct Key residues*, as several of the key repressor residues that recognise the DNA in the experimental complex still point to the DNA in the false positive, but the repressor is in a different position with respect to the DNA. Combinations of the above features are possible, and it is of course possible for a false positive to display none of these features.

Table 3.6 shows that most of the false positives ranked by shape complementarity do not show any of the native-like features. In contrast, the empirical pair potential score tends to maintain more of the native-like features in the top ranked dockings. In particular, of the six possible false positives at the top rank, five maintain the native-like key interactions.

Four representative false solutions are illustrated in Figures 3.20 through to 3.23. The DNA coordinates of the false positive docking were superposed on the crystallographic DNA coordinates. Each picture then depicts the predicted coordinates of the repressor with the X-ray repressor/DNA complex. The predicted model of LAM (ranked 3 by shape complementarity) is *shifted* with *correct key residues*. The docking for MET (ranked 4 by shape complementarity) does not have any native-like features. This predicted model is similar to that incorrectly proposed for the interaction of the met repressor operator with DNA after the structure of the repressor was solved but before the complex was experimentally determined.[135] The proposal was based on the α-helix/turn/α-helix motif inserted into the major groove. Our study confirms that the proposed model was highly favourable when assessed by shape complementarity. However, when ranked by the pair potential score, this model was placed at rank 1,011. Figures 3.22 and 3.23 also show the predicted structures for GAL and LAM at the top rank when ranked by the pair potential score. Both models were determined to be *rotated* with *correct key residues*.

Complex	Rank 1			Rank 2			Rank 3			Rank 4			Rank 5		
	S	R	K	S	R	K	S	R	K	S	R	K	S	R	K
Surface Complementarity															
ARC			*					+		+		*	+		
CRO			*					+		+				+	*
GAL					+			+				*			
LAC	+									+					
LAM				+	+	*				+		*		+	
MET	+			+		*			*						
PUR		+									+				
TRP										correct					
Empirical Pair Potentials															
ARC	correct			correct			+		*	+		*	correct		
CRO		+	*									*			
GAL				correct					*		+	*			*
LAC	+		*		+	*	+		*						
LAM					+	*	+		*	correct			correct		
MET						*									
PUR	+		*	+		*	+		*						
TRP	correct				+	*	+		*	+		*		+	*

Table 3.6: Analysis of False-Positive Solutions.
The first five ranked solutions after Filter 1, evaluated by shape complementarity and by the Empirical Pair Potentials Score, are reported. A "+" under "S" denotes a solution shifted along the DNA groove, and "+" under "R" a solution with about 180° rotation. Asterisk under "K" denotes that the solution had some of the key repressor residues correctly positioned interacting with the DNA. Solid black background denotes a correct prediction. Grey shading shows the four complexes whose structures are shown in Figures 3.20, 3.21, 3.22, and 3.23.

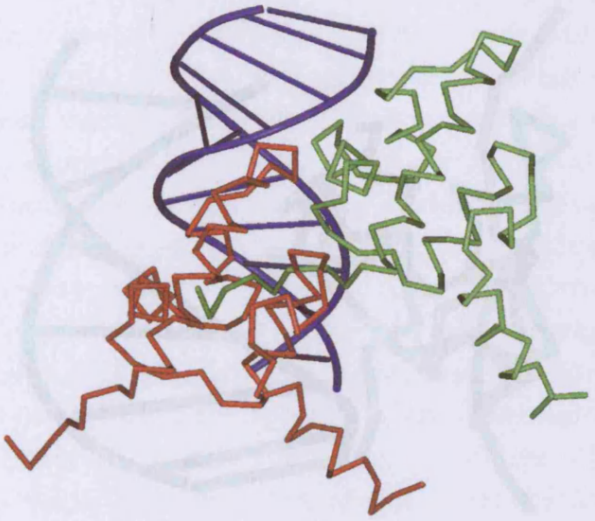


Figure 3.20: Superposition of a false positive and native complexes for LAM: rank 3 (after filter 1) by Surface Complementarity.

Figure 3.21: Superposition of a false positive and native complexes for MET: rank 4 (after filter 1) by Surface Complementarity.

Key to Figures 3.20 and 3.21.
Structures of false-positive modelled structures. The false-positive location of the repressor (red) is shown on top of the X-ray DNA (blue) and repressor complex (green). Diagrams were generated by PREPI.

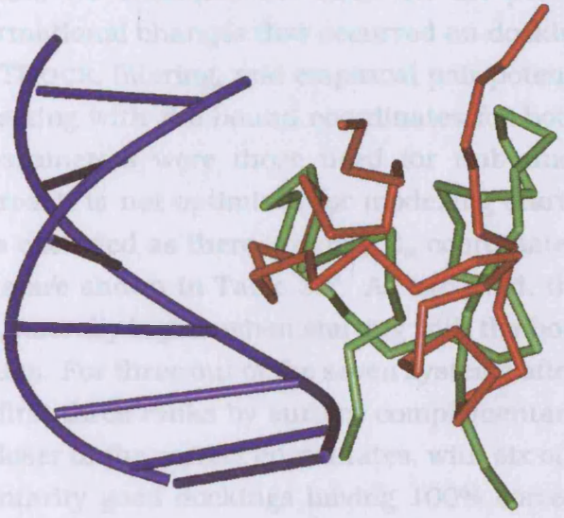


Figure 3.22: Superposition of a false positive and native complexes for GAL: rank 1 (after filter 1) by Empirical Pair Potential Score.

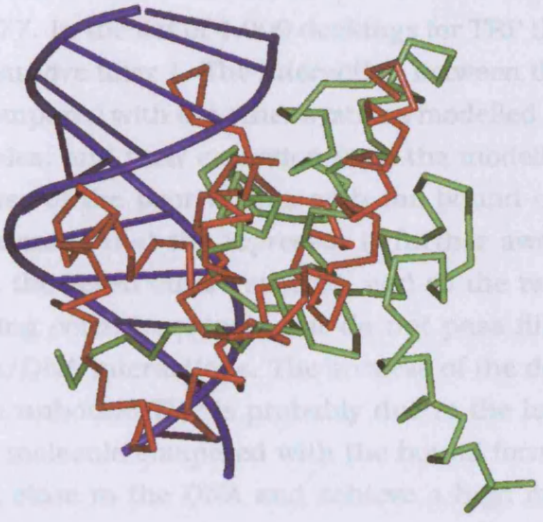


Figure 3.23: Superposition of a false positive and native complexes for LAM: rank 1 (after filter 1) by Empirical Pair Potential Score.

Key to Figures 3.22 and 3.23.
Structures of false-positive modelled structures. The false-positive location of the repressor (red) is shown on top of the X-ray DNA (blue) and repressor complex (green). Diagrams were generated by PREPI.

3.4.4 Control - bound complexes

To assess the algorithm, we evaluated the effect on the performance of our approach of the conformational changes that occurred on docking. The docking procedure involving FTDOCK, filtering, and empirical pair potential scoring was repeated, this time starting with the bound coordinates for both the repressor and the DNA. The parameters were those used for unbound docking, and consequently the approach is not optimised for modelling starting with bound coordinates. (LAC was excluded as there are only C_α coordinates for the bound repressor.) The results are shown in Table 3.7. As expected, the ranks for the first good docking are generally higher when starting with the bound rather than the unbound coordinates. For three out of the seven systems after filter 0, a good docking is within the first three ranks by surface complementarity. In addition, the solutions are far closer to the correct coordinates, with six of the top ranking by surface complementarity good dockings having 100% correct contacts and RMSD < 2.0Å. When starting from bound components, there were no systems that failed to generate at least one good docking.

The MET complex can now be modelled with a structure that has an RMSD of 0.9Å. However, this docking is at a low rank of 268, showing that there are many alternatives with superior shape complementarity. The role of electrostatics in the MET complex will be discussed below.

The poorest result was for TRP, with the first correct solution after filter 0 at rank 638 with %CC of 77. In the list of 4,000 dockings for TRP their are two good dockings, but neither survive filter 1. The interaction between the TRP repressor and DNA is unusual compared with the other systems modelled, as it is mediated by many water molecules, and their exclusion from the modelling was thought to be the possible cause of the poor results with the bound coordinates. The effect of the bound waters is that the repressor is further away from the DNA in this system than in the seven others studied, and so the two good dockings can be scored as having correct contacts but do not pass filter 1 as they do not form direct protein/DNA interactions. The success of the docking algorithm when starting with the unbound TRP is probably due to the less extended side chain in the unbound molecule compared with the bound form, which enables unbound TRP to pack close to the DNA and achieve a high measure of shape complementarity.

The modelling was repeated adding the 31 crystallographic waters that were in van der Waals contact with any repressor interface amino acid. The aim was to explore the effect of water on the evaluation of shape complementarity by FTDOCK, and so the waters were considered as van der Waals particles without charge. The results are a marked improvement in the ranking of the first good docking after filter 0 and filter 1 (Table 3.7). Of course in a real predictive study,

Complex	No. of solutions	No. of good solutions *i.e.* with %CC > 65	Rank of first good solution evaluated by shape complementarity [rank(%CC/RMSD(Å))]	electrostatics [rank (%CC/RMSD(Å))]	empirical pairing score [rank (%CC/RMSD(Å))]
Filter 0 - disallow repressor at ends of DNA					
ARC	3147	27	3(100/0.5)	90(100/0.5)	4(100/0.5)
CRO	3110	47	11(100/0.3)	20(100/1.1)	1(100/0.3)
GAL	2535	107	2(100/0.6)	3(100/0.8)	6(96/3.1)
LAC	N/A	N/A	N/A	N/A	N/A
LAM	3023	27	2(100/1.2)	2(100/0.4)	2(77/3.5)
MET	3371	8	268(100/0.9)	1(100/0.2)	188(100/0.2)
PUR	2793	31	58(100/1.6)	14(89/2.4)	16(100/0.9)
TRP	3365	2	638(77/2.5)	800(77/3.3)	78(77/3.3)
TRP + WAT	3371	4	1(84/2.1)	249(98/1.2)	55(98/1.2)
Filter 1 - use information about DNA bases with filter 0					
ARC	1318	22	3(100/0.5)	57(100/0.5)	4(100/0.5)
CRO	705	34	2(100/0.3)	5(100/1.1)	1(100/0.3)
GAL	1481	100	2(100/0.6)	3(100/0.8)	6(96/3.1)
LAC	N/A	N/A	N/A	N/A	N/A
LAM	674	23	2(100/1.2)	1(100/0.4)	2(77/3.5)
MET	197	2	86(100/0.9)	1(100/0.2)	24(100/0.2)
PUR	1016	16	32(100/1.6)	5(89/2.4)	16(100/0.9)
TRP	727	none	no solution	no solution	no solution
TRP + WAT	718	3	1(84/2.1)	85(98/1.2)	55(98/1.2)
Filter 2 - use information about DNA bases and repressor residues with filter 0					
ARC	201	21	1(100/0.5)	7(100/0.5)	4(100/0.5)
CRO	25	12	1(100/0.3)	1(100/1.2)	1(100/0.3)
GAL	713	100	2(100/0.6)	2(100/0.8)	6(96/3.1)
LAC	N/A	N/A	N/A	N/A	N/A
LAM	138	23	1(100/1.2)	1(100/0.4)	2(77/3.5)
MET	5	2	2(100/0.9)	1(100/0.2)	2(100/0.2)
PUR	469	16	16(100/1.6)	5(89/2.4)	16(100/0.9)
TRP	140	none	no solution	no solution	no solution
TRP + WAT	150	3	1(84/2.1)	16(98/1.2)	16(98/1.2)

Table 3.7: Rank of Solutions, Starting With Bound Structures.

After each of the three filters, the first column gives the complex, the second column the number of solutions left in the list of the top 4000 complexes generated from FTDOCK (N), and the third column gives the number of good solutions, i.e. with %CC > 65. The final three columns give the rank of the first correct solution followed by its %CC and RMSD(Å). Rankings were calculated using shape complementarity from FTDOCK, electrostatic score from FTDOCK, and the empirical score for nucleotide/amino acid pairings. TRP + WAT refers to simulations with the trp repressor and bound waters. N/A denotes that FTDOCK was not run for LAC with the bound coordinates, due to only a C_α trace existing for the repressor in the bound complex.

knowledge of these waters would not be available. The results show that when bound waters are included for TRP, FTDOCK can model bound-protein/bound-DNA for all seven systems. This is basic to any confidence in predictive modelling starting from unbound coordinates.

The bound FTDOCK solutions after each of the three filters were also ranked by their empirical pair potential scores (Table 3.7), and the results were comparable to the ranking from shape complementarity. For some systems, ranking with the pair potential scores was worse for the bound than with the unbound.

A further question was which was the greater problem for our approach; the conformational changes in the repressor or in the DNA. To study this, FTDOCK was run for unbound repressor with bound DNA, and for bound repressor with model DNA. After filter 1, there was no systematic trend showing that modelling is better starting with the bound coordinates for the repressor or for the DNA (data not shown). The results could not be explained in terms of the RMSD values between unbound (or model) and bound coordinates, as shown in Table 3.5.

3.4.5 Role of electrostatics

To evaluate the role of electrostatics, FTDOCK was also run without the electrostatic binary filter, so only shape complementarity was used when ranking the some 10^{10} dockings in order to provide the top 4,000 as output. The ranking of the first good docking was about twice the rank from when the electrostatic filter was on (data not presented), showing that choosing only dockings with favourable electrostatics is an important filter in these docking simulations.

Tables 3.3 and 3.7 also show the consequence of ranking the FTDOCK dockings by electrostatic score rather than by shape complementarity, for each of the three filters. (Note that as the dockings generated by FTDOCK are the top 4,000 by shape complementarity, this set could exclude a very favourable electrostatic score that has a poor shape complementarity.) For most systems starting from unbound coordinates, ranking by electrostatics yielded poorer results than by shape complementarity. This confirms our strategy of using electrostatics only as a binary filter. The results when starting from the bound coordinates show that, in general, ranking by electrostatics is comparable to ranking by shape complementarity. This suggests that for most complexes, recognition is a combination of good shape complementarity and favourable electrostatics. However, the ranking of the best attainable docking for MET (%CC of 60 for unbound coordinates) improves markedly using electrostatics, for both the unbound and bound systems. After filter 0 starting with the bound coordinates, a model for MET is top rank by electrostatics, with an RMSD of

0.2Å. This suggests that for MET, specificity is determined primarily by the favourable electrostatic interactions.

3.5 Discussion and Conclusion

The results show that computer modelling can generate from unbound coordinates a limited set of repressor/DNA dockings, one of which is close to the experimental complex. This was, as far as is known, the first reported study that considered several systems starting with unbound repressor and model DNA, and systematically performed a global search to yield a restricted list of dockings that in the vast majority of cases included at least one docking close to the experimental complex. In the absence of any biological data, the standard procedure is to run FTDOCK, that evaluates shape complementarity and requires favourable electrostatics, followed by ranking using the empirical score for amino acid / nucleotide pairing. This can yield a list of less than five dockings, one of which is good, in three out of eight of the case systems studied. Additional use of DNA footprinting increases this to four out of the eight systems. For further improvement, knowledge of which amino acids on the repressor interact with the DNA is required. With this knowledge, using the ranking just from FTDOCK, a good docking is found in the top 30 ranks for seven out of the eight systems. These different results show the applicability of the algorithm with different levels of available biological data.

The empirical scoring of amino acid / nucleotide pairings were successful in removing false solutions from the list of dockings generated by FTDOCK. This would be useful to reduce the number of dockings to be examined by a subsequent computationally more intensive step, such as refining the structures of the dockings by allowing flexibility in the DNA and the protein side chains. Such refinement procedures have been developed by Kaptein's group.[124,125] There are also several approaches for deriving empirical pairing scores,[87,93,94] and further studies could improve the discrimination of these scores in screening dockings of protein/DNA complexes.

From this study, it can be seen that the level of discrimination, and the accuracy of the good dockings (%CC and RMSD), is useful enough to prioritise further experimental studies in a genuine unkown system. For example, mutagenesis could be used to probe structure / function relationships or to design protein based regulators of transcription. More generally, this study shows that in the absence of gross conformational change on association, it is viable to predict computationally the structure of protein/DNA complexes.

Chapter 4

Integrated Docking System

4.1 Introduction

The initial investigation into using empirical residue level pair potentials (Chapter 2)[7] was successful enough to warrant further study. The two main aims of the work in this chapter were to use a pair potential derived from protein interfaces, and to test the method on a wider range of systems. Both aims were made possible by the increased numbers of structures in the PDB. A further aim was to rewrite the software so as to create a better integrated system.

Empirical residue level pair potentials were used to screen possible dockings of protein/protein complexes. A possible docking is defined as any model structure of a protein/protein complex. A correct docking is defined as a possible docking which meets two criteria. First, it must have not more than a 7.0Å RMSD for the C_α atoms of the smaller component of the complex, the larger component having been superposed on its C_α atoms, from the known experimental structure of the complex. Second, it must exhibit at least 25% of the pairs across the interface that exist in the known experimental structure. A pair is defined as when two residues on different sides of the interface have at least one atom within a distance cut-off. This distance is determined to be the same as the optimal value used for defining the pair potential function (4.5Å). A change from earlier work is the use of the word 'pair' as opposed to using the word 'contact'. It was felt that 'contact' was misleading when a pair of residues can interact without being in contact.

The possible dockings were generated by FTDOCK (version 2.0)[15] , a rigid-body docking program that ranks using shape complementarity and contains an electrostatic filter.

The complexes studied were 4 protease/inhibitors, 5 antibody/antigens, and 9 other complexes that do not fall into either of those categories. All these complexes exist in the PDB, as do their component parts in unbound forms, with the exception of the β-lactamase / β-lactamase inhibitor complex.[69] Starting

from the unbound crystallographic coordinates, FTDOCK was used to generate 10,000 possible dockings for each complex.

The pair potential functions tested were derived from observed intermolecular pairings across the interfaces in sets of non-homologous complexes. We found that the best parameters for the pair potential function was an interatomic cut-off distance of 4.5Å, along with disregarding any residue that did not have at least the equivalent surface accessibility of half a water molecule.

The experiment was run 3 times for each of the 18 complexes. In 14 out of the 18 systems, FTDOCK generated at least one correct docking in all 3 runs. In the remaining 4, FTDOCK generated a correct docking in at least 1 of the 3 runs. After the use of cross validated pair potential functions, a correct docking was found in the top 750 model structures (the top 250 from each of the 3 runs) in 12 out of the 18 complexes. This increased to 13 out of 18 if a biological filter was used where available.

The use of MULTIDOCK,[19] a side-chain refinement algorithm on the filtered lists of 750 model structures resulted in 6 out of the 18 complexes having a correct docking in the top 10 of the ranked complexes, with a further 3 in the top 30.

4.2 Methods

4.2.1 Software

Previous versions of all the software used in this work existed. However, due to both portability issues and a wish to integrate the software, both the rigid-body docking program and the pair potential scoring program were rewritten.

The original version of FTDOCK was written in Fortran77 with parallel capabilities on Silicon Graphics architectures with the appropriate parallel Fourier Transform libraries. The original version of the pair potential scoring program was written in Fortran90 and required a PDB structure as its input. Both programs were rewritten in ANSI compliant C and have successfully been compiled and run on Silicon Graphics machines with IRIX, and on Intel (and AMD) and DEC Alpha processors running Linux.

FTDOCK now uses for its Fourier library the 'Fastest Fourier Transform in the West' (FFTW)[139] which is itself written in C and portable to most platforms. The basic algorithm of FTDOCK has not changed apart from in the two following ways. The first is that the angle sampling was changed to what was considered a fairer sampling. The second is that it is now a command line option to choose the size of the grid cells as opposed to choosing the size of the grid (at compile time) in the original. This means it is easier to treat different size systems equally in the way in which they are discretised and subsequently surfaced by the program.

RPSCORE is the pair potential scoring program, and now reads in output files from FTDOCK directly. (There is still the option of reading in a PDB file, so as to allow for the evaluation of models built by other software.) All other peripheral programs were also rewritten in C to form an integrated package. The only exception is MULTIDOCK, the side-chain refinement program, which has not been updated apart from to compile it for an Intel processor Linux platform.

4.2.2 Test set of protein/protein interfaces

In previous work[7] we generated matrices from intramolecular pairings (*i.e.* pairs within a single protein domain) from a dataset of 385 protein domains. These matrices were used to screen modelled complexes for 9 protein/protein complexes with encouraging results. We also presented results using matrices generated from intermolecular pairings across protein/protein interfaces. An intermolecular dataset would have been expected to give better results due to the better similarity with the intermolecular pairs being evaluated by the matrices. However, due to the small size of the dataset the results were not as good as from using the intramolecular dataset.

With the steadily increasing number of structures in the PDB it is now possible to retrieve a larger dataset of protein/protein interfaces; large enough to be used to generate useful matrices. We selected all PDB structures with more than one chain, and the words 'Complex' or 'Bound' in their 'HEADER' or 'COMPOUND' fields. We then used SCOP (version 1.53) to determine homology criteria for the domains on either side of the interface. An interface (A–B) was considered to be homologous to another interface (C–D) if A was in the same SCOP Superfamily as C, and B was in the same SCOP Superfamily as D, *i.e.* there were only two SCOP Superfamilies represented in the total of four domains from the two interfaces. We did not use homodimeric interfaces as they are not the same form as protein/protein complex interfaces. A homodimer is more permanent and the component parts are not necessarily biologically viable in isolation.

In addition to this homology criteria, we made a restriction for the resolution of the PDB structure to be equal to or better than 2.5Å, and for there to be no nucleic acids in the structure. In the case of there being homologous interfaces, the interface from the best resolution structure was selected. The resulting dataset of 90 interfaces is shown in Table 4.1.

4.2.3 Interface residue level pair potential matrix generation

The matrices were generated from the interface dataset by counting the frequency of pairs of residue types i and j. From previous work [7] we already established that the random model should be calculated from the residue

PDB	Side 1	Side 2	PDB	Side 1	Side 2	PDB	Side 1	Side 2
1A14	LH	N	1DPJ	A	B	1QAV	A	B
1A2X	A	B	1DUZ	A	B	1QKZ	LH	A
1A2Y	AB	C	1DX5	AM	I	1QLA	AB	CF
1A4Y	A	B	1EAI	A	C	1QMZ	A	B
1ACB	E	I	1EAY	A	C	1QO0	A	DE
1AK4	A	D	1EER	A	BC	1QO3	AB	CD
1AVA	A	C	1EFN	A	B	1QUQ	A	B
1AVW	A	B	1EFU	A	B	1SBB	A	B
1AY7	A	B	1EG9	A	B	1SGP	E	I
1AZS	AB	C	1EUV	A	B	1SLU	A	B
1B0N	A	B	1EV2	A	E	1SMP	A	I
1BGX	HL	T	1FAK	H	I	1STF	E	I
1BLX	A	B	9FAK	LH	T	1TAF	A	B
1BVN	P	T	1FLE	E	I	1TCO	AB	C
1BVY	AB	F	1FLT	VW	X	1TMQ	A	B
1C1Y	A	B	1GC1	C	G	1TX4	A	B
1CDK	A	I	9GC1	LH	G	1UGH	E	I
1CLV	A	I	1GOT	BG	A	1VPP	V	X
1CSE	E	I	1IBR	A	B	1WEJ	LH	F
1CXZ	A	B	1ICF	AB	I	1YAG	A	G
1D2Z	A	B	1IKN	AC	D	1YCS	A	B
1D3B	A	B	1JSU	A	B	2BTC	E	I
1D4V	A	B	9JSU	AB	C	2JEL	LH	P
1D5M	A	C	1LPB	A	B	2PCC	A	B
1D6R	A	I	1MDA	HL	A	2TRC	BG	P
1D8D	A	B	1NSG	A	B	3BTH	E	I
1DAN	HL	TU	1OAK	LH	A	3SIC	E	I
1DEV	A	B	1OSP	LH	O	4CPA	E	I
1DHK	A	B	1PDK	A	B	4HTC	HL	I
1DIO	AB	G	1PYT	AB	CD	7CEI	A	B

Table 4.1: Dataset of 90 interfaces used to generate pair potential matrices. The PDB code is provided along with the chainIDs for the two component sides of the interface.

frequency (mole-fraction) as opposed to using pairing propensities (contact-fraction). *i.e.*

$$e_{i,j} = P_{Total} \times \frac{n_i}{N} \times \frac{n_j}{N}$$

$$P_{Total} = \sum_{i=1}^{i=20} p_i$$

$$p_i = \sum_{j=1}^{j=20} p_{i,j}$$

$$N = \sum_{i=1}^{i=20} n_i$$

where n_i and n_j are the total occurrences of each residue, $p_{i,j}$ is the number of pairs made across the interfaces between residue types i and j, and $e_{i,j}$ is the expected number of pairs made between residue types i and j. Also established from the previous work,[7] was the best definition of a pair. A pair is considered to exist if any atom from one residue is within a cut-off distance of any atom in another residue across the interface. (We have not tried using an atom level pair potential, as previous work[7] showed them to be of less value than residue level potentials.)

It should be noted that the ranges over which $p_{i,j}$ is summated to calculate P_{Total} is important. Although in effect the same pairs are counted twice, this compensates for the fact that $p_{i,j}$ is a symmetrical value, *i.e.* a single pair between residue types i and j causes an increment to both the counts of $p_{i,j}$ and $p_{j,i}$ (except when i is the same as j). All this ensures that the total of the expected values is equal to the total of the observed number of pairs (P_{Total}).

The score function is the same in the previous work, *i.e.*

$$s_{i,j} = ln(\frac{p_{i,j}}{e_{i,j}})$$

$$S_{Total} = \sum^{allpairs} p_{i,j}$$

There are two ways in which it is sensible to limit the counts of the occurrences of a residue type n_i; either to only those residues found on the

surface of the domains, or to only those residues found to make at least one pair across an interface. There is also the option of no restriction. This would mean counting n_i for all the residues, including core, in the structures. In work by Glaser[140] in 2001, the latter restriction is made. We, however, consider this restriction to bias the potentials towards an ability to select the correct orientation of an interface that is already established, as opposed to an ability to select docking interfaces from the surfaces of the two molecules. By not counting residues that do not appear in the interfaces, the potential loses any information about the propensity of a residue to be in an interface at all (regardless of what it pairs with). We therefore restricted the residue count only by a minimum relative surface accessibility (MRSA) value (calculated by DSSP).

There were two parameters therefore for which optimal values needed to be established; the cut-off distance and the MRSA of the residues to be considered in the calculations. We tested MRSA values of 0 (no restriction), 5, 10, 15, and 20, where a value of 10 is equivalent to an accessibility of one water molecule. The distance cut-off value was varied between 3 and 7Å at half Angstrom steps.

4.2.4 Test systems

In order to properly evaluate a docking algorithm it is necessary to find complexes in the PDB which also have their component parts in the PDB in an unbound form. By docking these unbound components, and comparing the modelled structures to the known crystallographic structure, the algorithm can be properly evaluated.

It has been a problem in the past to find a sufficiently large set of such examples in order to evaluate a docking algorithm. Previous studies have either shown only the few possible results from the examples available at the time, or have otherwise used bound components, either for one half of the complex, or for the total. However, with the increased size of the PDB we consider we have now got to the stage that there are sufficient examples of complexes with unbound components available to validly test an algorithm on these alone.

In order to find our test systems of complexes with unbound components, a Perl script was used to do as much as possible in an automated fashion. The use of SCOP (version 1.53) also sped the process up considerably. The method was as follows:

1. Identify all PDB coordinate sets with a resolution of 3Å or better. Only these PDB coordinate sets are being referred to below.

2. Use SCOP to make a list of those PDB coordinates that consist of more than one domain. This means the coordinates potentially constitute a complex. The keyword "COMPLEX" or "BOUND" is not always clearly present in a

PDB file that is a complex. By presuming these multidomained coordinate sets are complexes until the end of the procedure, we reduce dramatically the number of PDB files that need to be examined by hand.

3. Check the above structures to determine how many SCOP Superfamilies are present. If there is only one, then the structure was removed from the list. This removed homodimers.

4. For each of the list from above:

   (a) for each domain, find all other PDB coordinate sets that contain the same domain. These coordinates are potentially unbound conformations of the domain in the potential complex.

   (b) stop unless at least two domains in the potential complex have been matched to other PDB coordinate sets. Since we want both components of the complex in unbound conformations, this requires at least two domains to have been matched. There is no guarantee at this stage of the automated process that even if the potential complex is a real complex, that the domains matched constitute the domains involved in the interface. This is determined by hand at the end.

   (c) for each match above, attempt a sequence match to the appropriate potential complex domain. A match was considered to have occurred when at least half the potential complex domain sequence was exactly matched.

   (d) stop unless at least two domains are still matched, for the same reason as in (b), now with more information.

   (e) if the potential complex is matched across all of its domains by a single other PDB coordinate set, remove those matches since that other coordinate set would also be the same complex (presuming the potential complex is an actual complex).

   (f) stop unless at least two domains are still matched, for the same reason as in (b) and (d), now with even more information.

   (g) report on those PDB coordinate sets that have reached this stage, along with the matches. These are potential complexes with potentially unbound conformations available.

When we ran this the numbers were whittled down from just over 10 thousand PDB structures with domains in SCOP 1.53, to 180 possibles. The last stage had to be performed by hand. To determine between a multidomained structure and a complex, looking to see if the word "COMPLEX" or such was in the PDB headers was not always enough. SCOP proved useful in unclear

cases. The search resulted in 17 systems. A further system was known; beta-lactamase/BLIP, from a previous blind trial.[69] Table 4.2 shows the PDB files of the complexes and their unbound components. Figure 4.1 shows the reason for the inclusion of two antibody/lysozyme complexes. As can be seen, the two antibodies bind to the lysozyme in very different ways.

Description	PDB codes (chains or sections)		
	Complex	Component 1	Component 2
TRYPSIN / AMYLOID BETA-PROTEIN PRECURSOR INHIBITOR DOMAIN (APPI)	1BRC[141] (e: i:)	1BRA[142] (-)	1AAP[141] (a:)
ALPHA-CHYMOTRYPSINOGEN / HUMAN PANCREATIC SECRETORY TRYPSIN INHIBITOR	1CGI[25] (e: i:)	2CGA[143] (a:)	1HPT[144] (-)
KALLIKREIN A / BOVINE PANCREATIC TRYPSIN INHIBITOR	2KAI[27] (a: b: i:)	2PKA[27] (a: b:)	5PTI[145] (-)
SUBTILISIN BPN' / STREPTOMYCES SUBTILISIN INHIBITOR	2SIC[146] (e: i:)	1SUP[103] (-)	3SSI[146] (-)
EXTRACELLULAR DOMAIN OF TISSUE FACTOR / INHIBITORY FAB (5G9)	1AHW[147] (a:1-108 b:1-117 c:107-211)	1FGN[147] (l:1-108 h:1-117)	1BOY[148] (107-213)
HUMANIZED ANTI-LYSOZYME FV / LYSOZYME	1BVK[149] (a: b: c:)	1BVL[149] (b: a:)	3LZT[150] (-)
MONOCLONAL ANTIBODY FAB D44.1 / LYSOZYME	1MLC[31] (a:1-108 b:1-118 e:)	1MLB[31] (a:1-108 b:1-118)	3LZT[150] (-)
ANTI-LYSOZYME ANTIBODY HYHEL-63 / LYSOZYME	1DQJ[151] (a: b: c:)	1DQQ[151] (a: b:)	3LZT[150] (-)
IGG1 FAB FRAGMENT / HORSE CYTOCHROME C	1WEJ[152] (l:1-107 h:1-112 f:)	1QBL[152] (l:1-107 h:1-112)	1HRC[153] (-)
YEAST CYTOCHROME C PEROXIDASE (CCP) / YEAST ISO-1-CYTOCHROME C	2PCC[154] (a: b:)	1CCA[155] (-)	1YCC[154] (-)
BARNASE (G SPECIFIC ENDONUCLEASE) / BARSTAR MUTANT (C40A,C82A)	1BGS[156] (a: e:)	1A2P[156] (a:)	1A19[157] (a:)
BETA-LACTAMSE / BETA LACTAMASE INHIBITOR PROTEIN	tem.blip[69] (a: b:)	1TEM[158] (-)	blip[69] (-)
RIBONUCLEASE INHIBITOR / RIBONUCLEASE A	1DFJ[159] (i: e:)	2BNH[160] (-)	7RSA[161] (-)
ACETYLCHOLINESTERASE / FASCICULIN-II	1FSS[162] (a: b:)	1VXR[163] (a)	1FSC[162] (-)
V-1 NEF PROTEIN / WILD TYPE FYN SH3 DOMAIN	1AVZ[164] (a: c:)	1AVV[164] (-)	1SHF[165] (a:)
HUMAN URACIL-DEOXYRIBONUCLEIC ACID GLYCOSYLASE / PROTEIN INHIBITOR	1UGH[166] (e: i:)	1AKZ[167] (-)	1UGI[168] (a:)
RAS / RASGAP	1WQ1[169] (g: r:)	1WER[170] (-)	5P21[171] (-)
HPT DOMAIN / CHEY	1BDJ[172] (a: b:)	3CHY[173] (-)	2A0B[174] (-)

Table 4.2: Dataset of 18 systems used as the test set.

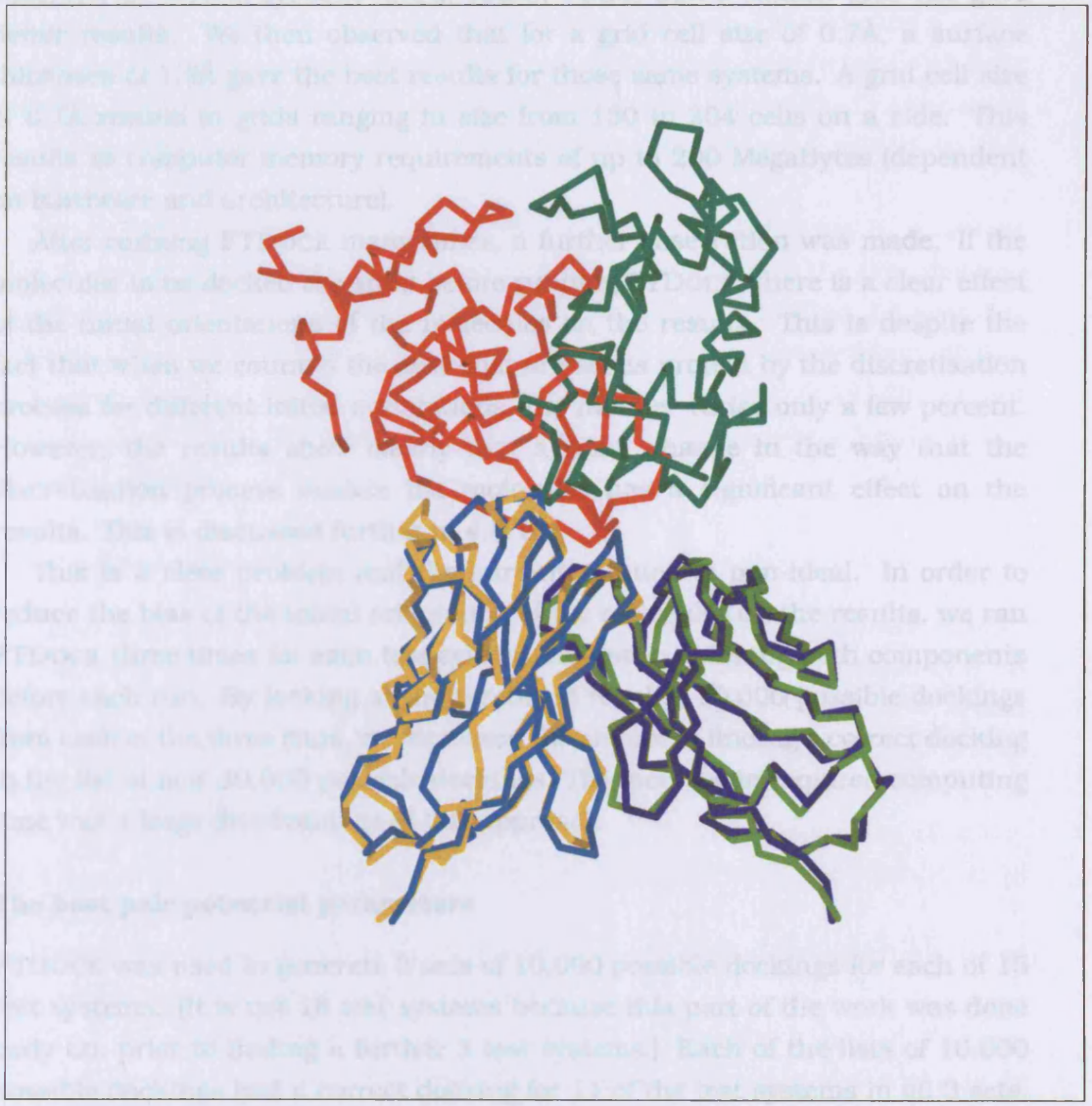


Figure 4.1: Superposition of two antibody/lysozyme complexes. The antibody is in the lower part of the picture, with the Heavy chain of the left. The lysozyme of 1mlc is on the top left (red), and that of 1bvk on the top right (green).

4.3 Results

4.3.1 Best use of FTDOCK

FTDOCK[15] was run on all 18 test systems. Although no systematic study was done for all 18 systems to determine the best value for the grid size, it was observed for several systems (2KAI, 1MLC, 1BGS) that a smaller grid size gave better results. We then observed that for a grid cell size of 0.7Å, a surface thickness of 1.3Å gave the best results for those same systems. A grid cell size of 0.7Å results in grids ranging in size from 130 to 204 cells on a side. This results in computer memory requirements of up to 200 MegaBytes (dependent on hardware and architecture).

After running FTDOCK many times, a further observation was made. If the molecules to be docked are spun before running FTDOCK, there is a clear effect of the initial orientations of the molecules on the results. This is despite the fact that when we counted the cells that are set as protein by the discretisation process for different initial orientations, the number varies only a few percent. However, the results show clearly that a small change in the way that the discretisation process models the molecules has a significant effect on the results. This is discussed further in 4.5.1.

This is a clear problem and the current solution is non-ideal. In order to reduce the bias of the initial orientation of the molecules on the results, we ran FTDOCK three times for each test system, randomly spinning both components before each run. By looking at the combined results, 10,000 possible dockings from each of the three runs, we increased the chance of finding a correct docking in the list of now 30,000 possible dockings. The increase in required computing time was a large disadvantage of this approach.

The best pair potential parameters

FTDOCK was used to generate 3 sets of 10,000 possible dockings for each of 15 test systems. (It is not 18 test systems because this part of the work was done early on, prior to finding a further 3 test systems.) Each of the lists of 10,000 possible dockings had a correct docking for 11 of the test systems in all 3 sets. The remaining 4 test systems had a correct docking in 2 of the 3 sets. We then ranked the total of 45 sets by cross-validated pair potential matrices. These were generated separately for each of the 15 test systems. For each test system, any interface in the dataset of 90 interfaces (Table 4.1) was removed if any of its components shared a SCOP Superfamily with any component of the test system. This resulted in 69 to 90 interfaces being left, which were then used to generate the cross-validated matrix for that test system. Each of these 15 cross-validated matrices were generated for each set of parameters that were investigated.

The performance of the parameters was evaluated by the number of test systems having a good model placed in the top 200 ranks in at least one of the three sets of 10,000 possible dockings. A good model was here defined as being one with 3Å or less RMSD over C_α atoms for the whole complex from the experimental structure. The value of 200 was decided upon for computational reasons. Any final stage of screening the possible dockings is currently computationally intensive, and it was felt that 600 complexes was a sensible number for such an algorithm to manage in a reasonable time. (We later changed this to 250.) Table 4.3 shows the values over the range of parameters. As can be seen, several pairs of parameters give equally good results by this criterion.

We therefore changed the criteria by changing the definition of a correct docking to being a possible docking with 5Å or less RMSD from the experimental structure. Table 4.4 shows the subsequent results. From this it can be seen that a distance cut-off of 4.5Å and a minimum relative surface accessibility (MRSA) of 5% gave the best results by this new criterion, with a total of 14/15 systems having a good model. This determined the use of those parameters for evaluation of pair potentials.

4.3.2 The pair potential matrix

The parameters of a distance cut-off of 4.5Å and a minimum relative surface accessibility (MRSA) of 5 are biologically sensible values. A distance of 4.5Å reflects the maximum distance at which an atomic interaction is likely across an interface. The requirement of having a residue not be totally buried is also sensible. Although such totally buried residues would rarely contribute to the count of pairs across an interface, the removal of these residues has an effect on the scores, since the expected values are calculated from the abundance of residue types.

Table 4.5 and Figure 4.2 respectively show the non cross validated matrix as a graphical representation and as numerical values.

It can be seen that the charged residues act in a largely 'classical' manner, *i.e.* like repel and opposites attract. The exception is Lys–Arg which shows no less preference to being in a pair than the expected calculation. The other broad observation is that large hydrophobics like to be paired.

These two observations show that our matrix is fulfilling two roles. The first is showing that the matrix, by reflecting features we know to be true about electrostatic interactions, is also hopefully including features we do not know about other specific residue/residue pairing preferences. The second is that a given residue having consistently positive scores in the matrix is showing a preference to being in the interface, a fact we know to be true of large

	Distance cut-off (Å)								
MRSA (%)	**3.0**	**3.5**	**4.0**	**4.5**	**5.0**	**5.5**	**6.0**	**6.5**	**7.0**
00	0	2	9	8	8	9	8	9	8
05	2	4	9	9	7	8	8	8	7
10	2	4	9	8	7	8	8	7	7
15	4	4	8	9	6	9	7	7	7
20	3	3	8	8	6	9	7	7	6

Table 4.3: Parameter Justification for Pair Potentials Matrices: 3Å as a good prediction.

	Distance cut-off (Å)								
MRSA (%)	**3.0**	**3.5**	**4.0**	**4.5**	**5.0**	**5.5**	**6.0**	**6.5**	**7.0**
00	8	10	12	12	11	13	12	13	11
05	11	12	13	14	10	12	13	11	11
10	11	11	13	13	11	12	12	11	11
15	12	11	12	13	11	13	10	11	10
20	11	11	13	12	11	13	10	11	10

Table 4.4: Parameter Justification for Pair Potentials Matrices: 5Å as a good prediction.

Key to Tables 4.3 and 4.4: The numbers in the grids show the number of test systems (out of 15) where a good model was found in at least one of the three sets of 10,000 possible dockings, ranked in the top 200 by the cross-validated pair potential generated with the given distance cut-off and MRSA.

hydrophobics. Thus our matrix is providing both the ability to select the correct binding areas of protein surfaces, and then the more specific residue/residue pairing preferences can provide information about the correct orientation of interfaces.

As a result of not using homodimers in our dataset, there is no anomalously high cystine–cystine score that needs to be dealt with as in previous work.

Arginine may have slightly more preference to pair than the chemically similar Lysine due to the larger charged terminus side chain. This would mean that the exact position of an Arginine side chain would have less effect than for that of Lysine.

Table 4.6 shows the raw counts of residue/residue pairs across all the interfaces. This shows that the size of the dataset being used to generate the matrix is statistically sensible. The addition of a single individual interface will not cause a significant change to these values. However, the addition of a single individual interface will also change the calculations of the expected values,

	D	E	K	R	A	V	F	P	M	I	L	W	Y	N	C	Q	G	H	S	T
Aspartic Acid	-0.7	-0.3	0.2	0.5	-0.4	-0.3	-0.0	-0.2	0.0	-0.2	-0.2	0.0	0.2	-0.1	-0.5	-0.0	-0.2	0.1	0.0	-0.2
Glutamic Acid	-0.3	-0.7	0.2	0.4	-0.4	-0.1	-0.1	-0.3	0.1	-0.3	-0.3	0.0	0.1	-0.0	-0.1	-0.1	-0.4	0.1	-0.1	-0.2
Lysine	0.2	0.2	-0.8	-0.0	-0.5	-0.2	-0.1	-0.5	0.1	-0.2	-0.1	0.1	0.3	-0.2	-0.0	-0.2	-0.2	-0.0	-0.2	-0.2
Arginine	0.5	0.4	-0.0	-0.1	-0.2	-0.0	0.3	0.0	0.4	0.1	0.1	0.4	0.4	0.2	0.2	0.2	0.0	0.1	0.2	-0.0
Alanine	-0.4	-0.4	-0.5	-0.2	-0.5	-0.0	0.1	-0.6	0.1	-0.1	0.0	0.1	0.1	-0.2	-0.1	-0.2	-0.3	0.0	-0.1	-0.4
Valine	-0.3	-0.1	-0.2	-0.0	-0.0	-0.0	0.2	-0.1	0.4	0.2	0.2	0.4	0.3	-0.1	-0.1	0.1	-0.1	-0.1	-0.2	-0.1
Phenylalanine	-0.0	-0.1	-0.1	0.3	0.1	0.2	0.3	0.1	0.6	0.6	0.4	0.3	0.4	-0.1	0.2	0.2	-0.1	0.3	0.1	0.0
Proline	-0.2	-0.3	-0.5	0.0	-0.6	-0.1	0.1	-0.6	0.4	-0.4	-0.1	0.3	0.2	0.0	-0.3	-0.1	-0.3	-0.2	-0.1	-0.1
Methionine	0.0	0.1	0.1	0.4	0.1	0.4	0.6	0.4	0.0	0.3	0.4	0.0	0.5	-0.0	0.0	0.2	0.3	0.0	-0.0	0.1
Isoleucine	-0.2	-0.3	-0.2	0.1	-0.1	0.2	0.6	-0.4	0.3	-0.2	0.3	0.4	0.3	-0.1	0.0	0.1	-0.2	0.2	-0.1	-0.1
Leucine	-0.2	-0.3	-0.1	0.1	0.0	0.2	0.4	-0.1	0.4	0.3	-0.1	0.2	0.3	-0.3	0.1	0.0	-0.1	0.1	-0.2	-0.1
Tryptophan	0.0	0.0	0.1	0.4	0.1	0.4	0.3	0.3	0.0	0.4	0.2	0.0	0.6	0.0	0.0	0.2	0.1	0.0	0.1	0.0
Tyrosine	0.2	0.1	0.3	0.4	0.1	0.3	0.4	0.2	0.5	0.3	0.3	0.6	0.1	0.3	0.2	0.3	0.1	0.5	0.1	0.1
Asparagine	-0.1	-0.0	-0.2	0.2	-0.2	-0.1	-0.1	0.0	-0.0	-0.1	-0.3	0.0	0.3	-0.2	-0.3	0.1	-0.2	-0.1	-0.2	-0.2
Cystine	-0.5	-0.1	-0.0	0.2	-0.1	-0.1	0.2	-0.3	0.0	0.0	0.1	0.0	0.2	-0.3	0.0	0.2	0.1	0.0	0.1	-0.2
Glutamine	-0.0	-0.1	-0.2	0.2	-0.2	0.1	0.2	-0.1	0.2	0.1	0.0	0.2	0.3	0.1	0.2	-0.3	0.1	0.0	-0.1	-0.2
Glycine	-0.2	-0.4	-0.2	0.0	-0.3	-0.1	-0.1	-0.3	0.3	-0.2	-0.1	0.1	0.1	-0.2	0.1	0.1	-0.4	0.1	-0.2	-0.2
Histidine	0.1	0.1	-0.0	0.1	0.0	-0.1	0.3	-0.2	0.0	0.2	0.1	0.0	0.5	-0.1	0.0	0.0	0.1	0.0	0.1	0.1
Serine	0.0	-0.1	-0.2	0.2	-0.1	-0.2	0.1	-0.1	-0.0	-0.1	-0.2	0.1	0.1	-0.2	0.1	-0.1	-0.2	0.1	-0.6	-0.1
Threonine	-0.2	-0.2	-0.2	-0.0	-0.4	-0.1	0.0	-0.1	0.1	-0.1	-0.1	0.0	0.1	-0.2	-0.2	-0.2	-0.2	0.1	-0.1	-0.6
totals	-2.5	-2.6	-2.6	3.1	-3.5	0.2	3.8	-2.8	4.0	0.8	0.9	3.3	5.4	-1.6	-0.4	0.6	-2.1	1.3	-1.6	-2.5

Table 4.5: Best Pair Potential Matrix

The values of the non cross validated pair potential matrix, generated using the best parameters of distance cut-off 4.5Å, MRSA of 5%.

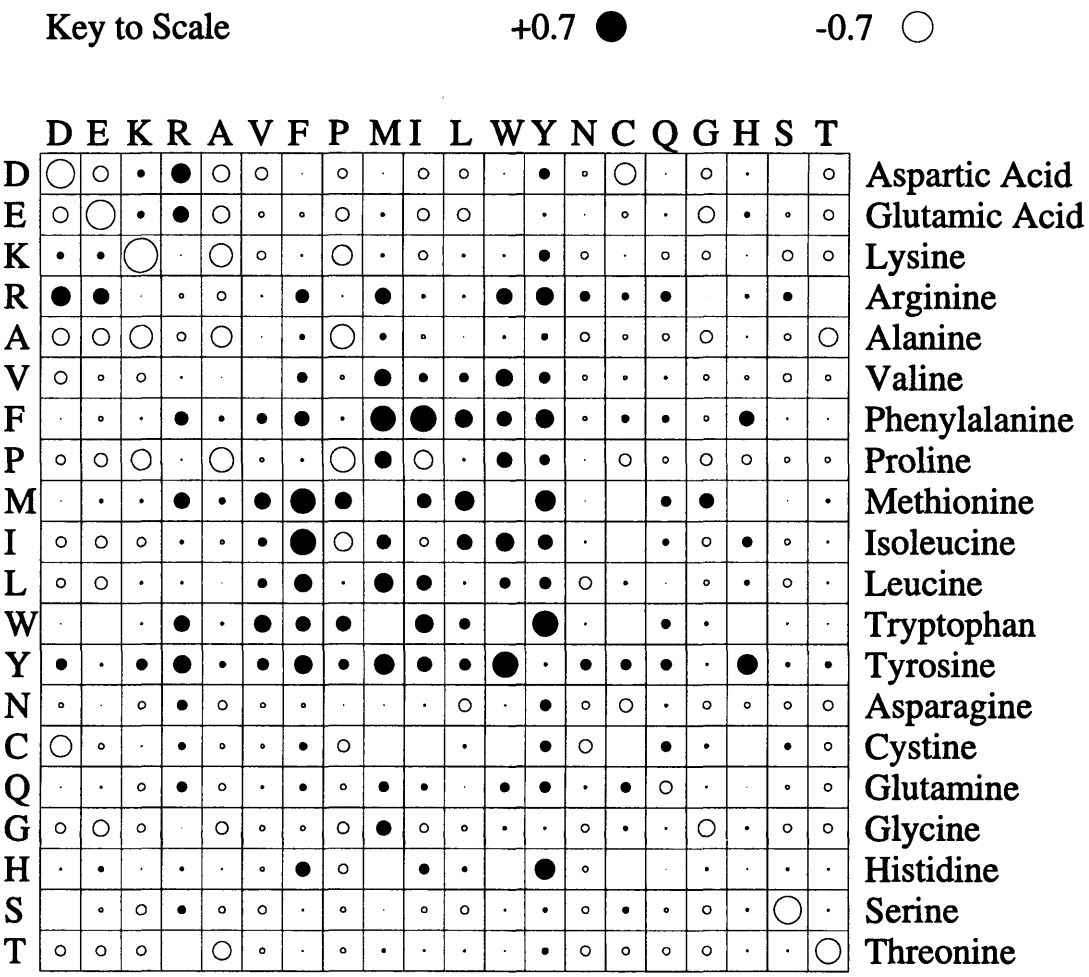


Figure 4.2: Best Pair Potential Matrix

A graphical representation of the non cross validated pair potential matrix, generated using the best parameters of distance cut-off 4.5Å, MRSA of 5%.

used in the generation of the score matrix. In the case where the interface is part of a large domain, the number of residues, not in the interface, counted for the calculations could change more markedly. This could cause a significant change to the calculations of the expected values (not shown).

Figure 4.3 shows that the composition of residues types in the dataset is broadly similar to that of the whole of SCOP 1.53 (*i.e.* PDB without homologues), from which the dataset was extracted.

4.4 Measurements of quality of models

In any field of research, it is important for different studies to agree on which metrics to use to measure the results of various approaches to the problem. Unfortunately there is no such clarity when it comes to measuring the success

	D	E	K	R	A	V	F	P	M	I	L	W	Y	N	C	Q	G	H	S	T
Aspartic Acid	11	24	74	120	17	20	23	23	12	16	30	13	52	30	4	34	30	22	56	27
Glutamic Acid	24	13	83	111	19	32	20	21	16	16	28	14	39	43	10	35	25	29	49	31
Lysine	74	83	9	44	13	26	22	14	15	18	44	15	56	28	12	25	36	19	34	30
Arginine	120	111	44	29	23	31	42	36	24	31	50	26	67	60	16	53	47	21	76	41
Alanine	17	19	13	23	12	33	27	9	14	19	43	13	35	21	8	21	23	18	34	15
Valine	20	32	26	31	33	31	32	24	22	35	64	25	43	24	8	34	31	12	29	28
Phenylalanine	23	20	22	42	27	32	24	22	21	52	60	13	37	15	9	24	19	20	30	24
Proline	23	21	14	36	9	24	22	8	22	8	33	22	39	32	5	20	22	9	32	27
Methionine	12	16	15	24	14	22	21	22	5	14	32	14	20	9	10	14	27	17	13	14
Isoleucine	16	16	18	31	19	35	52	8	14	10	62	19	36	20	7	28	19	19	23	24
Leucine	30	28	44	50	43	64	60	33	32	62	43	22	54	20	15	37	39	26	36	42
Tryptophan	13	14	15	26	13	25	13	22	14	19	22	8	29	11	7	15	17	10	16	13
Tyrosine	52	39	56	67	35	43	37	39	20	36	54	29	21	42	13	38	37	33	44	39
Asparagine	30	43	28	60	21	24	15	32	9	20	20	11	42	21	5	34	28	11	31	22
Cystine	4	10	12	16	8	8	9	5	10	7	15	7	13	5	1	15	17	14	20	8
Glutamine	34	35	25	53	21	34	24	20	14	28	37	15	38	34	15	13	42	14	32	23
Glycine	30	25	36	47	23	31	19	22	27	19	39	17	37	28	17	42	23	25	39	31
Histidine	22	29	19	21	18	12	20	9	17	19	26	10	33	11	14	14	25	8	26	21
Serine	56	49	34	76	34	29	30	32	13	23	36	16	44	31	20	32	39	26	15	48
Threonine	27	31	30	41	15	28	24	27	14	24	42	13	39	22	8	23	31	21	48	12
totals	638	658	617	948	417	584	536	428	335	476	780	322	774	507	204	551	577	374	683	520

Table 4.6: Raw observations of residue-residue pairings.

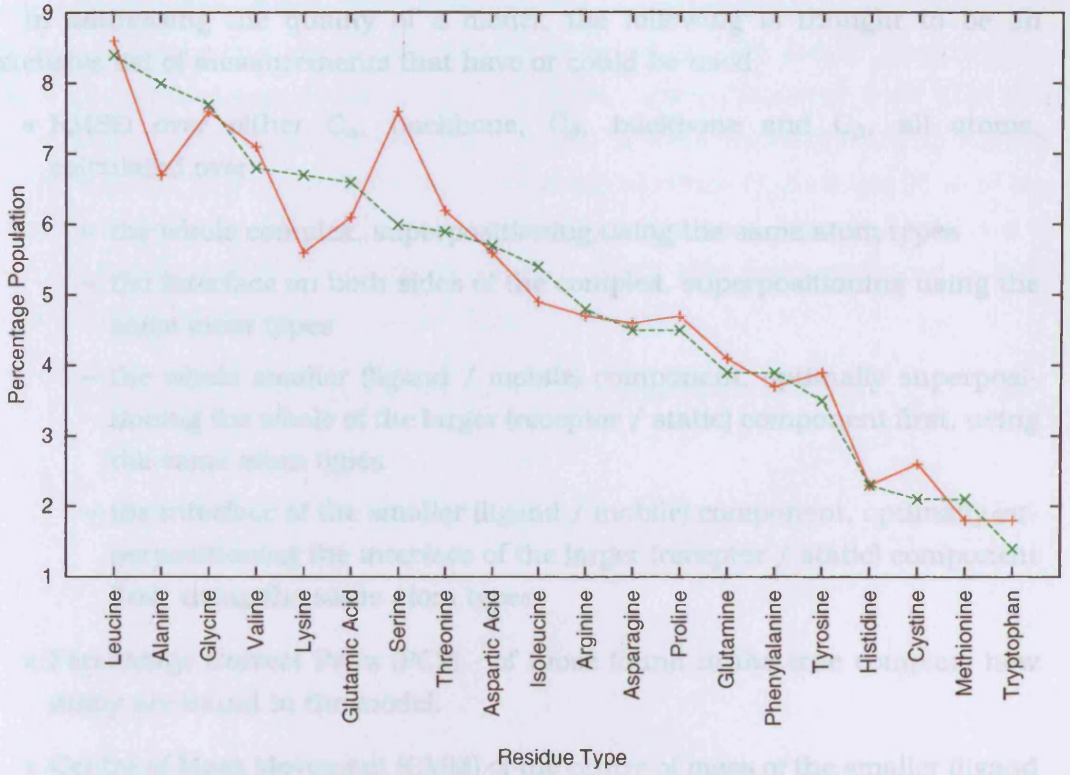


Figure 4.3: Comparison of populations of residue types. The comparison is between the residue type populations in SCOP 1.53 (steadily decreasing line) and the dataset of 90 complexes shown in Table 4.1.

of protein/protein docking algorithms. Even the blind trials have used different ways of measuring success.

There are two issues that need to be addressed. The first is how to measure the quality of a given model complex. The other is the number of models an algorithm may propose, stating that it thinks at least one of them is of good quality, and still be useful.

To address the second issue first, it is probably of little use to an end user to provide more than ten possible models of how two proteins dock. Even this may be considered too many, and recognising that a single good model is the at present unrealistic ideal, it is enough to say that the fewer the better. What an algorithm should be able to state is a certain probability of generating a good model in the top N models, where N needs to be no more than 10. It is however evident that any given study will want to differentiate between the types of protein/protein complexes when providing such statistics. The reason for this is that all algorithms to date have for example had more success at docking serine protease / inhibitor complexes than antibody / antigen complexes, and providing statistics across all systems with no discrimination will often

downgrade the successes of an algorithm.

In addressing the quality of a model, the following is thought to be an extensive list of measurements that have or could be used.

- RMSD over either C_α, backbone, C_β, backbone and C_β, all atoms, calculated over
  - the whole complex, superpositioning using the same atom types
  - the interface on both sides of the complex, superpositioning using the same atom types
  - the whole smaller (ligand / mobile) component, optimally superpositioning the whole of the larger (receptor / static) component first, using the same atom types
  - the interface of the smaller (ligand / mobile) component, optimally superpositioning the interface of the larger (receptor / static) component first, using the same atom types
- Percentage Correct Pairs (PCP) - of those found in the true complex, how many are found in the model.
- Centre of Mass Movement (CMM) of the centre of mass of the smaller (ligand / mobile) component from the correct position

There are two further questions with respect to the calculations of RMSD values. The first is the matter of what the model is being superposed onto and being compared to. Most studies use the true crystallographic complex. However, some use a "best possible model", where the unbound components have been individually optimally superposed onto the true crystallographic complex. The second question is what constitutes an interface. In this study the interface was defined as consisting of those residues which had at least one atom within 10Å of any atom in the other component of the true complex. This is not an uncommon value to choose as a distance cut-off, although there are no reasons against justifying the use of a smaller value.

For the analysis of the results of FTDOCK, a large number of the above measurements were calculated. The RMSD values were calculated only from the true crystallographic complex, and only using either C_α atoms or all atoms. This resulted in 8 RMSD values for each model. The other two values, namely percentage correct pairs (PCP) and centre of mass movement (CMM) were also calculated. With a total of 510,000 models from a total of 51 runs (17 systems each run three times - 1BDJ not included), reasonable analysis could be done on the relationship between these different values.

The first thing to note is the difference between calculations using all atoms or just C_α atoms. For calculations across both components of the

complex, whether it was the whole complex or the interface only, the differences were always within ± 1Å. For calculations where the larger component was first optimally superposed, and the RMSD calculation was over the smaller component, the value differed more, up to ± 2.5Å. Although these values are large enough to make a difference in stating whether a given model is of a good quality or not, it is small enough to be able to use to compare previous studies where different sets of atoms were used, given that the superpositioning and calculations were over the same regions.

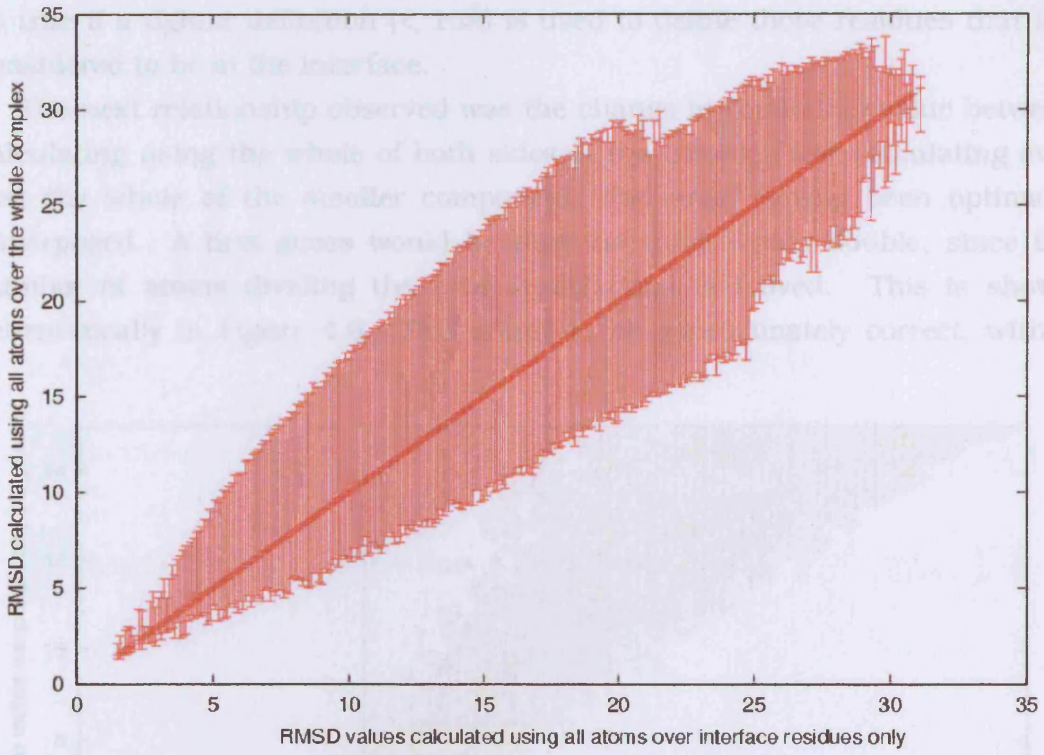


Figure 4.4: Comparison between RMSD (Å) value calculations.
The two calculations being compared are the RMSD values (in Angstroms) calculated using the whole complex, and RMSD values (in Angstroms) calculated using just interface residues. RMSD calculations are for all atoms. The graph shows the range of RMSD values calculated using the whole complex that can exist for a given calculation of RMSD using only the interface residues. For example, a calculation of 10Å using only interface residues can correspond to a calculation of between 7Å and 18Å for the whole complex, depending on the structure. The same can be shown in the other direction. For a calculation of 10Å using the whole complex, a calculation using only the interface residues can yield values between 5Å and 16Å. An interface residue is one that possesses at least one atom within 10Å of any atom on the other side of the interface. Data is from the evaluation of 510,000 model complexes generated by FTDOCK.

In comparing values calculated over different regions there is a substantially larger change. Comparing the methods of superposing just the interface as opposed to the whole complex, the highest increase is 8Å and the lowest decrease is 10Å, both using C_α and all atoms. However, as can be seen in Figure 4.4, the change is related to the value. At low, good quality, RMSD values, the region of calculation changes the values very little, though more often than not an interface calculation gives a higher value. From this it can be seen that, as above, different studies using different measures can still be crudely compared, at least when discussing the number of good quality models. This may not be as true if a tighter definition (< 10Å) is used to define those residues that are considered to be in the interface.

The next relationship observed was the change in the RMSD value between calculating using the whole of both sides of the complex and calculating over just the whole of the smaller component, the larger having been optimally superposed. A first guess would be that the value would double, since the number of atoms dividing the total translations is halved. This is shown schematically in Figure 4.6. This seems to be approximately correct, with a

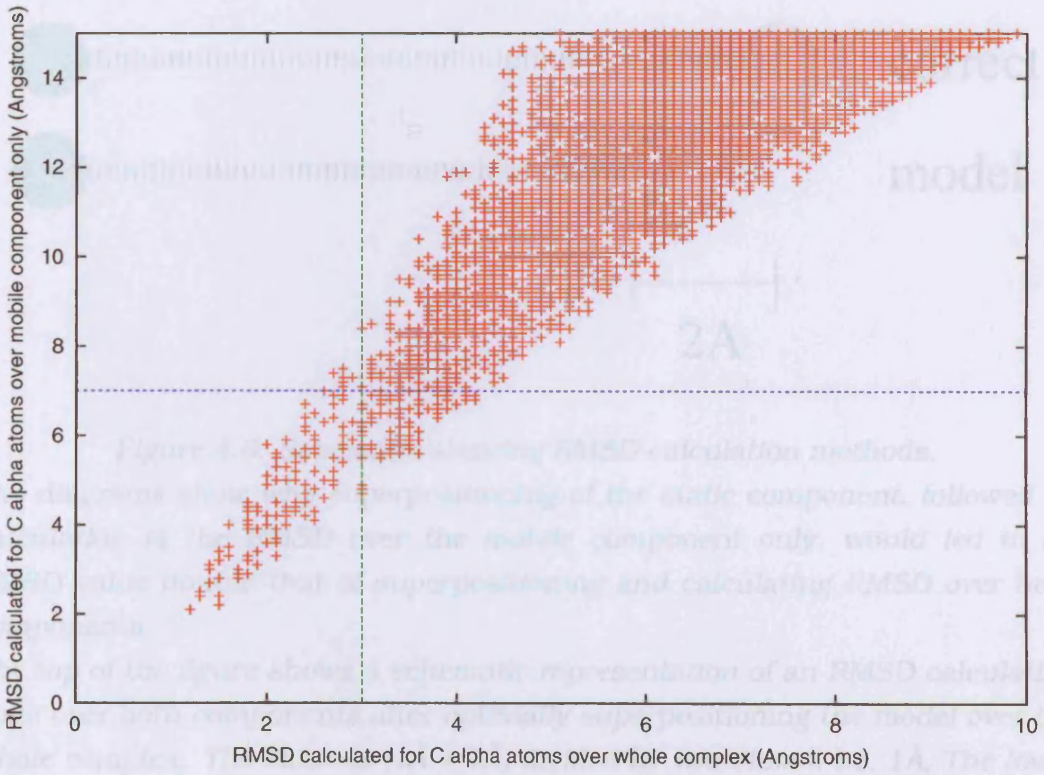


Figure 4.5: Further Comparison between RMSD (Å) value calculations. Comparison between use of RMSD calculated using either a whole model complex, or using just the whole of the mobile component, the static component having been optimally superposed on the native static component.

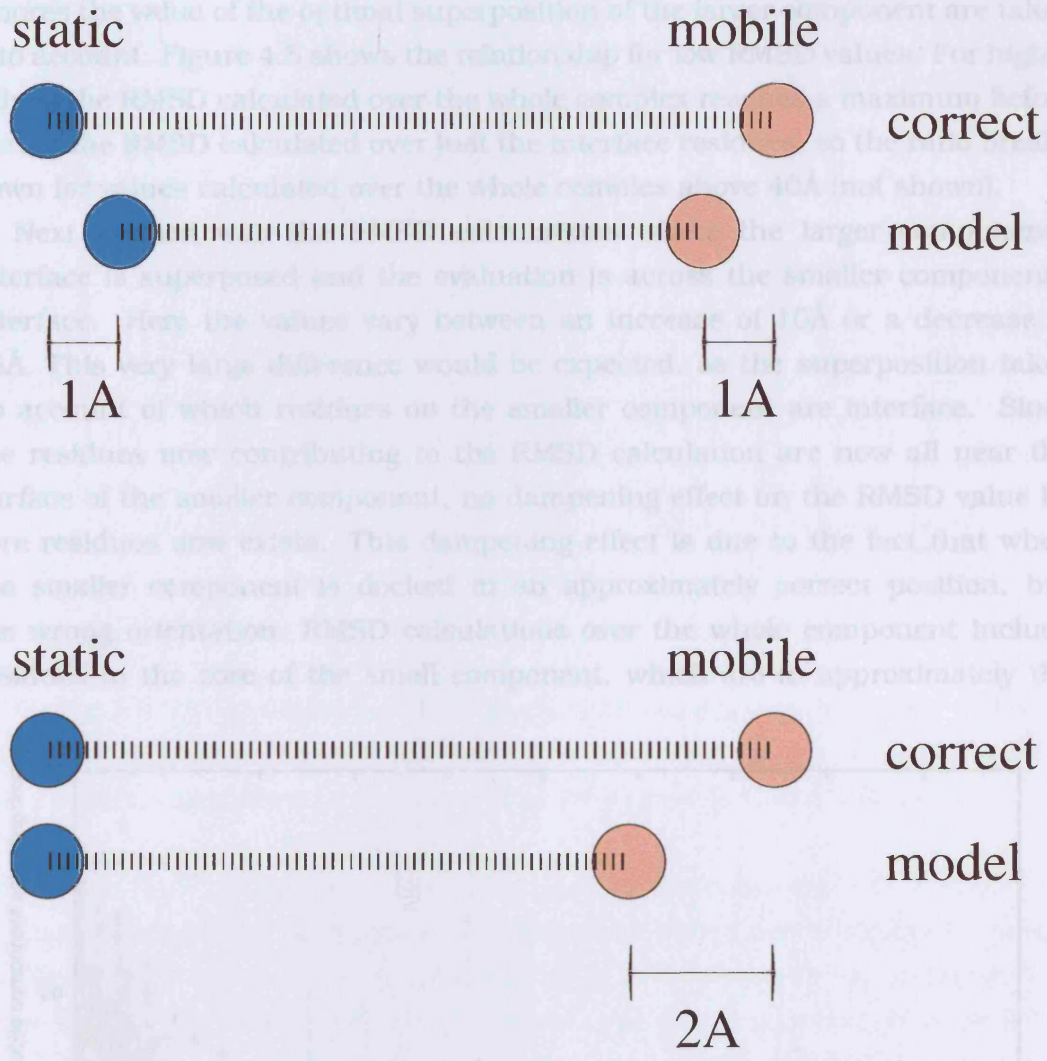


Figure 4.6: Schematic showing RMSD calculation methods.
The diagrams show why superpositioning of the static component, followed by calculation of the RMSD over the mobile component only, would led to an RMSD value double that of superpositioning and calculating RMSD over both components.
The top of the figure shows a schematic representation of an RMSD calculation done over both components after optimally superpositioning the model over the whole complex. The value is (1Å + 1Å) divided by two atoms, i.e. 1Å. The lower figure shows a representation of an RMSD calculation done over the mobile component with the static component already optimally superposed. The value here is simply 2Å.

certain amount of variation, once that fact that the smaller component often has significantly less than half the total atoms, and the fact that the calculation ignores the value of the optimal superposition of the larger component are taken into account. Figure 4.5 shows the relationship for low RMSD values. For higher values the RMSD calculated over the whole complex reaches a maximum before that of the RMSD calculated over just the interface residues, so the ratio breaks down for values calculated over the whole complex above 40Å (not shown).

Next studied was the RMSD calculations where the larger component's interface is superposed and the evaluation is across the smaller component's interface. Here the values vary between an increase of 10Å or a decrease of 25Å. This very large difference would be expected, as the superposition takes no account of which residues on the smaller component are interface. Since the residues now contributing to the RMSD calculation are now all near the surface of the smaller component, no dampening effect on the RMSD value by core residues now exists. This dampening effect is due to the fact that when the smaller component is docked in an approximately correct position, but the wrong orientation, RMSD calculations over the whole component include residues in the core of the small component, which are in approximately the

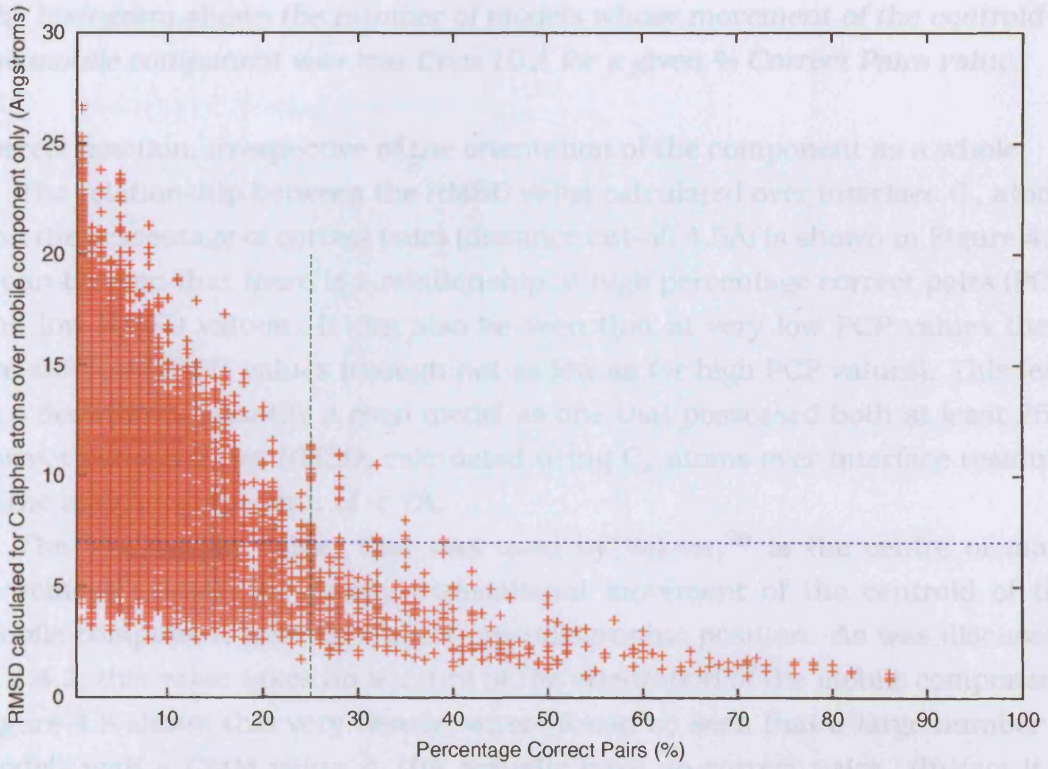


Figure 4.7: Comparison between RMSD (Å) calculated using the mobile component and % Correct Pairs.
The two lines are at $x = 25$ and $y = 7$

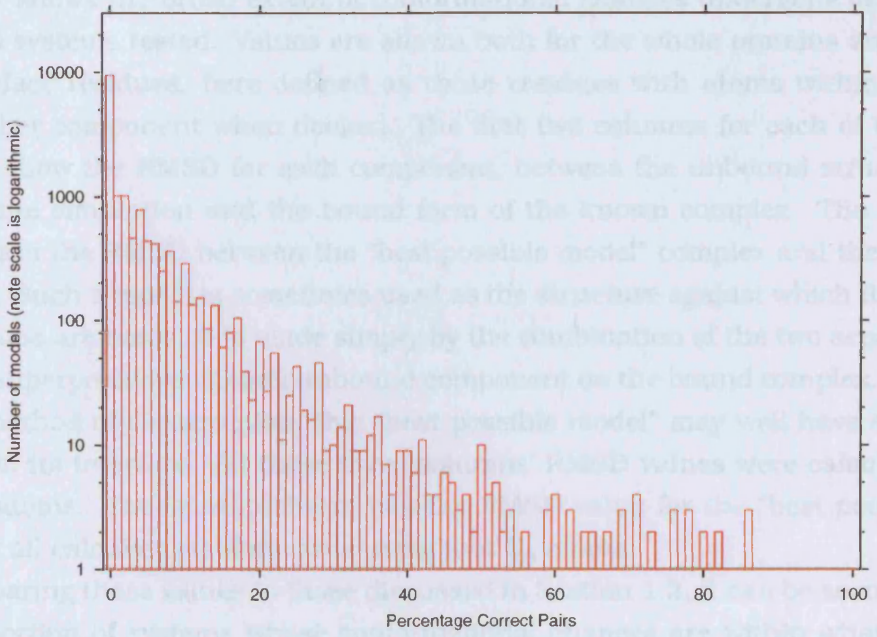


Figure 4.8: Comparison between Centroid Movement and % Correct Pairs. The histogram shows the number of models whose movement of the centroid of the mobile component was less than 10 Å for a given % Correct Pairs value.

correct position, irrespective of the orientation of the component as a whole.

The relationship between the RMSD value calculated over interface C_α atoms and the percentage of correct pairs (distance cut-off 4.5Å) is shown in Figure 4.7. It can be seen that there is a relationship at high percentage correct pairs (PCP) and low RMSD values. It can also be seen that at very low PCP values there are still low RMSD values (though not as low as for high PCP values). This lead to a decision to quantify a good model as one that possessed both at least 25% correct pairs, and an RMSD, calculated using C_α atoms over interface residues of the mobile component, of < 7Å.

The last quality value, that was used by Vakser,[36] is the centre of mass movement (CMM), the purely translational movement of the centroid of the mobile component from its correct crystallographic position. As was discussed in 1.4.2, this value takes no account of the orientation of the mobile component. Figure 4.8 shows this very clearly, where it can be seen that a large number of models with a CMM value $<$ 10Å actually have no correct pairs. (In fact it is 9452 models of a total 16019 models that had such a CMM value $<$ 10Å from all the experiments, *i.e.* 59%.) 10Å was the limit used by Vakser as defining a good quality model.[36]

4.4.1 Conformational changes

Table 4.7 shows the broad extent of conformational changes undergone in each of the 18 systems tested. Values are shown both for the whole proteins and for the interface residues, here defined as those residues with atoms within 10Å of the other component when docked. The first two columns for each of these two sets show the RMSD for each component, between the unbound structure used in the simulation and the bound form of the known complex. The third column has the RMSD between the "best possible model" complex and the true complex. Such a model is sometimes used as the structure against which RMSD calculations are made. It is made simply by the combination of the two separate optimal superpositions of each unbound component on the bound complex. Due to this method of construction, this "best possible model" may well have steric clashes in its interface. All these three columns' RMSD values were calculated over all atoms. The fourth column has the RMSD value for the "best possible model" if all calculations were done using just C_α atoms.

Comparing these values to those discussed in Section 1.2, it can be seen that the proportion of systems whose conformational changes are within what can be regarded as experimental crystallographic variance is about half, the same as in the study by Betts and Sternberg.[3] However, the results below (4.4.2) show an unclear relationship between the extent of conformational change and the success of FTDOCK.

4.4.2 Results of global scans

Table 4.8 shows the results for FTDOCK, along with the re-ranking of the models by pair potentials cross-validated for each system. The first thing that is clear is the variance in the quality and number of good models between the different systems. All the test systems generated at least 3 good models out of the 30,000 models generated. However, the ranking of some of those models, by both surface complementarity and pair potentials was simply too poor to be of any real use. On the other hand, some systems have good models ranked very highly, particularly by the pair potential scores, and this is without any biological filter.

There is in general a large improvement from the surface complementarity ranks to the pair potential ranks. What is also still true, as it was in the work shown in Chapter 2 (and in the 1999 paper by Moont *et al.*[7]), is that the number of false negatives, *i.e.* good models ranked badly, is significantly less for the pair potential ranks than the surface complementarity score ranks (data not shown).

As can be seen in Table 4.8, only one system, in one of the three runs, has a good model ranked highly by the surface complementarity scores, namely the Ribonuclease Inhibitor / Ribonuclease A complex (1DFJ) at rank 3. The pair

Test	Whole Structure [RMSD(Å)]				Interface [RMSD(Å)]			
System	Static	Mobile	Complex	Complex (C_α)	Static	Mobile	Complex	Complex (C_α)
1BRC	0.8	1.2	0.9	0.4	0.9	1.3	1.0	0.4
1CGI	1.9	2.5	2.0	1.5	2.8	2.4	2.7	2.0
2KAI	1.2	1.2	1.2	0.7	1.4	1.1	1.3	0.6
2SIC	0.6	1.3	0.9	0.4	0.6	1.1	0.8	0.4
1AHW	0.9	1.3	1.0	0.7	0.9	1.2	1.1	0.7
1BVK	1.4	1.9	1.6	1.0	1.2	2.0	1.6	1.2
1MLC	1.0	1.2	1.1	0.7	0.9	0.9	0.9	0.6
1DQJ	1.2	1.5	1.3	0.7	1.0	1.4	1.2	0.8
1WEJ	0.8	1.2	1.0	0.3	0.7	1.3	1.0	0.3
2PCC	0.8	1.1	0.9	0.4	1.1	0.9	1.0	0.4
1BGS	0.9	1.1	1.0	0.5	0.9	1.0	0.9	0.4
BLIP	1.0	1.0	1.0	0.5	1.2	1.1	1.1	0.5
1DFJ	1.8	1.2	1.7	1.4	1.6	1.3	1.5	1.0
1FSS	1.0	1.3	1.1	0.6	1.3	1.5	1.4	0.7
1AVZ	1.2	1.3	1.2	0.6	1.3	1.4	1.3	0.6
1UGH	0.9	1.5	1.1	0.5	0.9	1.5	1.2	0.5
1WQ1	1.3	1.0	1.2	0.8	1.4	1.2	1.3	0.8
1BDJ	1.8	1.1	1.5	0.9	2.3	1.3	1.9	1.0

Table 4.7: Conformational changes between unbound and bound structures.
Values are RMSD (in Å) between unbound and bound structures. Calculations are over all atoms, except where C_α is specified.

Test System	Run 1				Run 2				Run 3				Total Good Structures
	Surface Complementarity		Pair Potentials		Surface Complementarity		Pair Potentials		Surface Complementarity		Pair Potentials		
	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	
1BRC	631	5.3/58	20	4.3/78	565	3.8/76	4	5.7/60	206	1.9/86	8	1.6/86	33
1CGI	failed				89	4.8/31	21	6.6/33	failed				3
2KAI	201	6.5/38	286	3.2/40	2936	2.7/59	307	2.7/55	160	6.8/28	218	3.2/51	17
2SIC	620	6.9/29	225	4.8/72	8526	4.7/46	715	4.7/46	2215	2.1/54	291	2.1/54	15
1AHW	288	2.2/64	240	5.9/37	780	3.6/50	600	6.9/35	240	5.1/45	123	5.1/45	12
1BVK	579	4.5/45	144	6.9/45	368	4.0/40	194	4.4/62	7103	5.4/37	686	5.3/45	14
1MLC	5514	3.8/27	209	3.8/27	2375	5.0/34	364	5.0/34	failed				3
1DQJ	6227	6.1/26	518	6.1/26	failed				2867	5.5/29	654	5.7/31	5
1WEJ	200	4.3/48	95	5.5/65	68	6.7/28	45	7.0/51	6893	5.4/25	28	4.5/28	13
2PCC	979	6.7/36	771	6.5/36	9404	5.6/44	2165	5.6/44	3768	5.9/52	205	5.9/52	6
1BGS	542	4.9/36	33	4.9/36	892	5.9/56	76	4.0/56	1993	6.7/29	7	4.3/59	18
BLIP	185	4.2/45	33	3.8/30	400	3.1/36	63	3.1/36	5620	3.1/47	419	3.1/47	13
1DFJ	3	5.6/25	3316	5.4/28	failed				85	4.6/35	2821	4.6/35	3
1FSS	5251	3.4/43	1289	3.4/45	47	5.7/26	78	5.7/26	257	5.2/26	45	4.9/33	13
1AVZ	1198	5.2/25	1243	5.2/25	1421	6.9/50	557	6.9/50	4680	5.1/28	2481	6.8/53	10
1UGH	1692	6.1/25	911	6.1/25	6747	5.3/25	441	5.3/25	587	5.5/27	458	5.5/27	6
1WQ1	7994	5.5/30	1414	5.5/30	776	5.1/47	5362	5.1/47	3404	6.6/30	2672	4.8/44	6
1BDJ	2394	5.3/56	7618	4.8/43	8407	6.7/50	3388	6.4/50	2988	5.1/50	7649	5.1/50	9

Table 4.8: Top ranks of correct dockings. RMSD values in Ångstroms.

Each set of 4 columns shows the top rank for each of the 3 runs - both by surface complementarity score and by cross validated pair potential score. The last column showing the total number of correct dockings over all 3 runs. A correct docking is a structure whose RMSD value is $\leq$ 7Å over C_α atoms for the interface residues of the mobile molecule (the static molecule having been superposed with the correct structure over all its C_α atoms), and which has at least 25% of the correct pairs (PCP $\geq$ 25) that exist in the correct structure. An interface residue is a residue with at least one atom within 10Å of any atom on the other side of the interface. A pair is defined as between two residues spanning the interface, where at least one atom in one residue is within 4.5Å of any atom in the other residue on the other side of the interface.

potentials ranks two systems, in at least one run, in the top 10, the Trypsin complex (1BRC) and the Barnase / Barstar complex (1BGS). Both of these two systems are often used in docking studies, and most algorithms model them well. The success of modelling these systems correctly is therefore more an indication that the FTDOCK algorithm is not doing anything too wrong, rather than an indication that it is advancing the progress of protein/protein docking.

The reasons for being able to generate good models for some test systems and not others, and the subsequent ranking of good models, tend to be unclear and varied. There is a case for example that the α-Chymotrypsinogen complex (1CGI) is the least successful of the serine protease systems, at least in terms of the total number of good models generated, because of the large conformational changes. However, in the case of the three antibody/lysozyme complexes, it is the system with the highest conformational changes (1BVK) that in fact was the most successful at generating a large number of good models.

4.4.3 Effect of biological filtering

Table 4.9 shows the results from Table 4.8 after a biological filter has been applied. Information for biological filtering existed for 16 of the 18 systems, the two without being the last two in the tables, for which there is therefore no change between from Table 4.8. The biological filter for the enzyme/inhibitor systems was that the inhibitor should be in contact with at least one of the catalytic triad residues. For the antibody/antigen systems, the antigen had to be in contact with at least one residue on the H3 or L3 CDRs of the antibody. Other systems biological filters were based on literature. It should be noted that these biological filters would not always be known in the absence of an experimentally determined structure for the complex.

It can be seen that there is still in general a large improvement from the surface complementarity ranks to the pair potential ranks.

On investigation, it could be seen that the ranks changed differently for surface complementarity compared to pair potentials, when a biological filter was applied. Fewer models ranked highly by pair potentials were removed by the biological filter, compared to models ranked highly by surface complementarity. This was particularly true of antibody/antigen systems. In the antibody / horse cytochrome C system (1WEJ), there is no improvement at all for the best model rank by pair potentials for any of the three runs.

Figure 4.9 shows the different effects of filtering the two different rankings. The lines show the % of models remaining after filtering (in statistical bins of 100 models), when ordered by that rank. The green lines show the % for the pair potential ranks, the red lines show the % for the surface complementarity ranks. Figure 4.9 shows lines for 1WEJ, as well as the antibody / tissue factor

Test System	Run 1				Run 2				Run 3				Total Good Structures
	Surface Complementarity		Pair Potentials		Surface Complementarity		Pair Potentials		Surface Complementarity		Pair Potentials		
	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	Rank	RMSD(Å)/PCP	
1BRC	41	5.3/58	13	4.3/78	39	3.8/76	1	5.7/60	10	1.9/86	1	1.6/86	33
1CGI	failed				8	4.8/31	4	6.6/33	failed				3
2KAI	43	6.5/38	81	3.2/40	302	2.7/59	41	2.7/55	21	6.8/28	35	3.2/51	17
2SIC	36	6.9/29	9	4.8/72	235	4.7/46	12	4.7/46	101	2.1/54	12	2.1/54	15
1AHW	126	2.2/64	233	5.9/37	276	3.6/50	443	6.9/35	85	5.1/45	112	5.1/45	12
1BVK	249	4.5/45	144	6.9/45	133	4.0/40	189	4.4/62	2267	5.4/37	646	5.3/45	14
1MLC	1117	3.8/27	162	3.8/27	564	5.0/34	251	5.0/34	failed				3
1DQJ	1509	6.1/26	482	6.1/26	failed				968	5.5/29	604	5.7/31	5
1WEJ	87	4.3/48	95	5.5/65	36	6.7/28	45	7.0/51	1918	5.4/25	28	4.5/28	13
2PCC	526	6.7/36	355	6.5/36	4914	5.6/44	1013	5.6/44	1979	5.9/52	84	5.9/52	6
1BGS	314	4.9/36	26	4.9/36	503	5.9/56	48	4.0/56	1150	6.7/29	7	4.3/59	18
BLIP	4	4.2/45	7	4.9/41	270	4.9/37	5	4.9/37	84	3.1/47	10	3.1/47	8
1DFJ	3	5.6/25	1061	5.4/28	failed				37	4.6/35	933	4.6/35	3
1FSS	401	3.4/43	108	3.4/45	6	5.7/26	10	5.7/26	19	5.2/26	7	4.9/33	12
1AVZ	708	5.2/25	1046	5.2/25	805	6.9/50	484	6.9/50	2489	5.1/28	1754	6.8/53	10
1UGH	295	6.1/25	134	6.1/25	941	5.3/25	64	5.3/25	110	5.5/27	53	5.5/27	6
1WQ1	7994	5.5/30	1414	5.5/30	776	5.1/47	5362	5.1/47	3404	6.6/30	2672	4.8/44	6
1BDJ	2394	5.3/56	7618	4.8/43	8407	6.7/50	3388	6.4/50	2988	5.1/50	7649	5.1/50	9

Table 4.9: Top ranks of correct dockings after filtering. RMSD values in Ångstroms.

Each set of 4 columns shows the top rank for each of the 3 runs after filtering - both by surface complementarity score and by cross validated pair potential score. The last column showing the total number of correct dockings over all 3 runs after filtering. A correct docking is a structure whose RMSD value is $\leq$ 7Å over C_α atoms for the interface residues of the mobile molecule (the static molecule having been superposed with the correct structure over all its C_α atoms), and which has at least 25% of the correct pairs (PCP $\geq$ 25) that exist in the correct structure. An interface residue is a residue with at least one atom within 10Å of any atom on the other side of the interface. A pair is defined as between two residues spanning the interface, where at least one atom in one residue is within 4.5Å of any atom in the other residue on the other side of the interface.

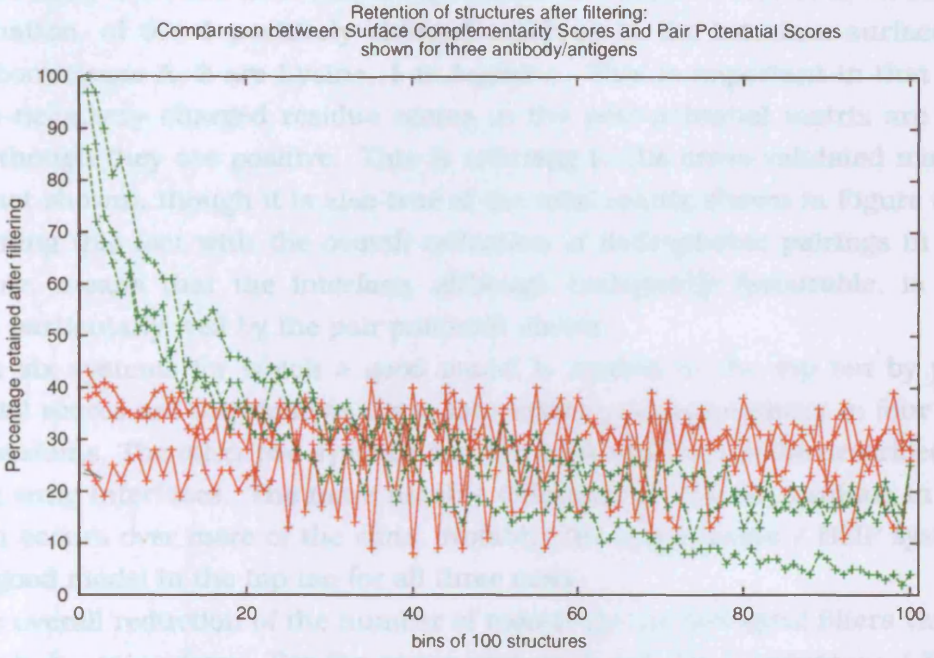


Figure 4.9: Complexes not discarded after filtering.
The x-axis shows 100 bins of 100 model complexes corresponding to the 10000 ranked models produced by FTDOCK *before biological filtering. The steeply dropping green lines show that models ranked high (in the first several bins) by pair potential score are rarely discarded. The steady red lines show that the ranks provided by surface complementarity do not exhibit this feature. The behaviour is shown here for three systems; 1WEJ, 1AHW, and 1MLC.*

(1AHW) and one of the antibody / lysozyme systems (1MLC). It can be seen that the green lines have near 100% values at the highest ranks, dropping rapidly for lower ranks. The red lines do not show this feature.

It can be seen that after filtering, surface complementarity scores can place a good model in the top 10 ranks for five different systems. Each of these systems have quite a snug interface, in that the two components of the complex come together in a lock and key type fit. This is in comparison to the type of interface an antibody / antigen complex exhibits, which is significantly more open, and for which surface complementarity cannot rank a good model well. Since FTDOCK primarily relies on surface complementarity to screen the millions of possible relative positions and orientations of the two components, it is also reasonable to assume that the overall success of FTDOCK is affected in the same manner.

There is only one system for which surface complementarity ranks are more

successful than pair potentials, namely the Ribonuclease Inhibitor / Ribonuclease A complex (1DFJ). The interface here is predominantly electrostatic in nature, and the inhibitor is strongly negative, while the ribonuclease is approximately balanced in overall charge in the interface.[175] However, on closer examination, of the 4 positively charged residues in the interface surface of the Ribonuclease A, 3 are Lysine, 1 is Arginine. This is important in that the Lysine–negatively charged residue scores in the pair potential matrix are not large, though they are positive. This is referring to the cross-validated matrix (data not shown), though it is also true of the total matrix shown in Figure 4.2. Combining this fact with the overall reduction of hydrophobic pairings in the interface, means that the interface, although biologically favourable, is not scored particularly well by the pair potential matrix.

The six systems for which a good model is ranked in the top ten by pair potential scores are the same as those for surface complementarity in four out of six systems. The other two systems are also ones which could be described as having snug interfaces. The more notable difference is that the ranking in the top ten occurs over more of the runs. Notably, the β-lactamase / BLIP system has a good model in the top ten for all three runs.

The overall reduction of the number of models by the biological filters varied greatly between systems. For the serine proteases and the β-lactamase / BLIP system, the number of models left after the filter was less than 1000, a reduction to less than 10% of the list generated by FTDOCK. However, for some other systems the reduction was not even to half the list generated by FTDOCK, though all systems did lose at least 40% of models (data not shown).

4.4.4 Combined results with MULTIDOCK

The combined approach, as used in Chapter 2 (and in the 1999 paper by Moont *et al.*[7]), was again used, though with some changes. The top 250 models (see section 4.3.1) as ranked by pair potentials were taken from each of the three runs. This was done with the lists that had been filtered, and also with those that had not been filtered. MULTIDOCK was then run for each of the 750 models from each system, and the resulting models were ranked according to the energy function. The results can be seen in Table 4.10, both for the unfiltered and filtered experiments.

The most impressive improvement is for the antibody / tissue factor system (1AHW), which now has moved to rank 1, irrespective of whether a filter was used or not (though Figure 4.9 shows that this is maybe not surprising). In total there are four systems when unfiltered, 6 when filtered, which have a good model in the top ten ranks, and in both cases three of these are in the top 3 (though not exactly the same three).

Test System	Unfiltered Rank	Unfiltered RMSD(Å)/PCP	Filtered Rank	Filtered RMSD(Å)/PCP
1BRC	3	4.2 / 86	4	4.2 / 86
1CGI		failed		failed
2KAI	45	4.5 / 57	18	3.3 / 59
2SIC		failed	8	3.7 / 53
1AHW	1	6.4 / 43	1	6.4 / 43
1BVK	213	6.0 / 50	218	6.0 / 50
1MLC		failed		failed
1DQJ		failed		failed
1WEJ	22	5.0 / 42	22	5.0 / 42
2PCC	6	6.8 / 56	7	6.8 / 56
1BGS	108	2.6 / 54	106	2.6 / 54
BLIP	1	4.4 / 50	2	4.4 / 50
1DFJ		failed		failed
1FSS	45	4.4 / 45	1	3.5 / 67
1AVZ		failed		failed
1UGH		failed	11	5.4 / 28
1WQ1		failed		failed
1BDJ		failed		failed

Table 4.10: MULTIDOCK *results.*

The change from rank 1 to rank 2 for the β-lactamase / BLIP system may seem strange. However, all it is showing is that the filtering, though removing false positives as ranked by pair potentials, in fact introduced other false positives, one of which was ranked by MULTIDOCK at rank 1 in place of the good model.

There is in fact a problem with MULTIDOCK which would be hard for any algorithm to overcome. If the initial rigid body model, good or not, has too many steric clashes, then MULTIDOCK crashes, unable to complete the algorithm. It is known that this does remove some good models (not shown).

4.4.5 False positives

Table 4.11 shows the quality of the models ranked top by the various algorithms. All these ranks are without filtering. Although none of the top rank models as ranked by either surface complementarity or pair potentials are of any use, a small change can be seen in the quality. The top model by surface complementarity scores have an RMSD below 20Å a total of 6 times over the three runs. None of these 6 is below 10Å. In comparison, pair potential ranks

Test System System	Run 1 Surface Complementarity RMSD(Å)/PCP	Run 1 Pair Potentials RMSD(Å)/PCP	Run 2 Surface Complementarity RMSD(Å)/PCP	Run 2 Pair Potentials RMSD(Å)/PCP	Run 3 Surface Complementarity RMSD(Å)/PCP	Run 3 Pair Potentials RMSD(Å)/PCP	MULTIDOCK RMSD(Å)/PCP
1BRC	43.1 / 0	15.4 / 8	18.1 / 0	24.0 / 0	27.8 / 0	29.0 / 0	17.1 / 0
1CGI	49.3 / 0	8.7 / 20	23.8 / 11	44.0 / 0	44.1 / 0	14.1 / 15	14.4 / 5
2KAI	35.1 / 0	17.1 / 10	22.0 / 0	14.1 / 0	34.5 / 0	20.8 / 0	18.5 / 12
2SIC	28.1 / 0	34.4 / 0	44.2 / 0	45.5 / 0	33.5 / 0	33.8 / 0	19.9 / 12
1AHW	45.3 / 0	31.1 / 0	33.4 / 0	28.8 / 0	21.4 / 1	21.2 / 0	6.4 / 43
1BVK	15.0 / 7	21.2 / 12	33.9 / 0	16.6 / 5	53.0 / 0	14.7 / 2	20.2 / 0
1MLC	35.4 / 0	24.2 / 0	16.7 / 4	35.5 / 0	24.3 / 0	29.7 / 0	25.7 / 0
1DQJ	40.6 / 0	21.3 / 3	53.5 / 0	24.6 / 0	24.5 / 4	24.3 / 0	18.0 / 0
1WEJ	44.1 / 0	19.8 / 0	27.0 / 0	12.7 / 5	49.2 / 0	19.2 / 0	19.1 / 14
2PCC	49.0 / 0	62.6 / 0	17.5 / 0	24.2 / 0	17.9 / 4	53.2 / 0	62.4 / 0
1BGS	42.2 / 0	13.3 / 6	44.4 / 0	18.0 / 11	19.1 / 2	45.4 / 0	22.3 / 0
BLIP	31.7 / 0	30.1 / 0	23.1 / 1	38.1 / 0	28.5 / 0	24.9 / 0	4.4 / 50
1DFJ	27.6 / 5	27.6 / 1	46.8 / 0	32.5 / 8	29.4 / 0	27.2 / 0	7.7 / 8
1FSS	46.7 / 0	48.8 / 0	39.6 / 0	42.5 / 0	55.2 / 0	48.7 / 0	48.2 / 0
1AVZ	20.9 / 3	35.0 / 0	20.5 / 0	35.7 / 0	27.1 / 0	22.9 / 0	39.1 / 0
1UGH	32.7 / 0	23.5 / 3	38.6 / 0	29.2 / 0	42.0 / 0	30.0 / 3	41.2 / 0
1WQ1	52.0 / 0	51.3 / 0	61.6 / 0	37.1 / 0	35.5 / 2	57.8 / 0	24.5 / 0
1BDJ	32.7 / 0	29.5 / 0	40.8 / 0	30.0 / 0	30.4 / 0	45.9 / 0	31.4 / 0

Table 4.11: Top ranks of the docking algorithms. RMSD values in Ångstroms.

Each set of 2 columns shows the RMSD and PCP of rank 1 for each of the 3 runs - both by surface complementarity score and by cross validated pair potential score. The last column shows the RMSD and PCP of the rank 1 after MULTIDOCK. *No filtering is used for these results. The RMSD value is in Å calculated over C_α atoms for the interface residues of the mobile molecule, the static molecule having been superposed with the correct structure over all its C_α atoms. The Percentage Correct Pairs (PCP) value is the % of the correct pairs exhibited in the predicted structure that exist in the correct structure. An interface residue is a residue with at least one atom within 10Å of any atom on the other side of the interface. A pair is defined as between two residues spanning the interface, where at least one atom in one residue is within 4.5Å of any atom in the other residue on the other side of the interface.*

have an RMSD below 20Å a total of 12 times, one of which is as low as 8.7Å.

The main change is when you look at the results for the combined approach with MULTIDOCK. Now there are models with an RMSD below 20Å in half the systems, including all of the serine proteases. Two of which are good models. Although a model with an RMSD below 20Å is in itself of no use, it may be useful as a starting point for a non global search algorithm. Even though using FTDOCK and MULTIDOCK is not fast, an algorithm that introduced flexibility would be even more computationally intensive, and knowledge of the approximate correct position would be of great use.

4.5 Discussion

4.5.1 Initial orientation of the molecules

As was discussed above (section 4.3.1), there is a large and problematic effect of the initial orientations of the component molecules on the results of FTDOCK. In order to assess exactly how large this problem is we repeatedly randomly spun both individual components and complexed models, and plotted histograms of the resulting surfacing and surface complementarity values.

In order to see how much the initial orientation affected the surfacing algorithm, the trypsin variant 1BRA was spun 10,000 times. The average number of cells set to core was 14330, with a standard deviation of 84. The average number of cells set to surface was 49826, with a standard deviation of 78. These standard deviations may seem small against the average value, yet small initial changes can result in large final differences.

4.5.2 Decoy sets

We generated a set of decoys for each system from our runs of FTDOCK. A decoy set is a list of model structures, containing both good and bad models. Such a set can then provide a quick way of testing new energy functions. A good set of decoys for a sizable group of systems can also then provide a standardised test set for use in the wider community.

There are 100 structures in each set of decoys, one set for each of the 18 systems, with the exception of BLIP. The first is the experimental crystal structure. The next 3 are the best good models, selected from the combination of the three runs of FTDOCK. The remaining 96 are decoys, grouped into bins of 3 for each 1 Ångstrom RMSD range. The lowest range is from 9 to 10 Ångstroms, the highest from 40 to 41, making a total of 32 bins. These decoys were selected by binning the much longer list of all the models generated over three runs, and then choosing as decoys the 3 in each bin with the highest surface complementarity scores. This binning schema is illustrated in Figure 4.10.

It was also considered that we should test our pair potential matrix on our own decoy set. We have already established the success of a correctly cross-validated pair potential score. Here we used the non-cross-validated matrix, the one which has been distributed with the software. Table 4.12 shows the results.

It can be seen that the native crystal structure is placed at rank one in 4 systems, and in the top ten for a total of 14 systems. A good model is ranked top once (and in the case of it being ranked 2 it is because the native structure is at rank 1), and in the top ten for a total of 12 systems. The very poor performance of the HPT domain /CheY (1BDJ) system may be partially due to the large conformational change on association, as shown in Table 4.7. However, it is not the only complex with such high values.

4.6 Conclusion

The work done for this chapter showed improvement on the previous work. A larger set of test systems was used, and showed that the method could be applied to a substantively wider variety of biological forms than was shown before. A larger dataset of interfaces allowed for the successful generation of a pair potential from the same environment as it was to be applied to, and so avoids assumptions about the similarity or not of interfaces to domain cores.

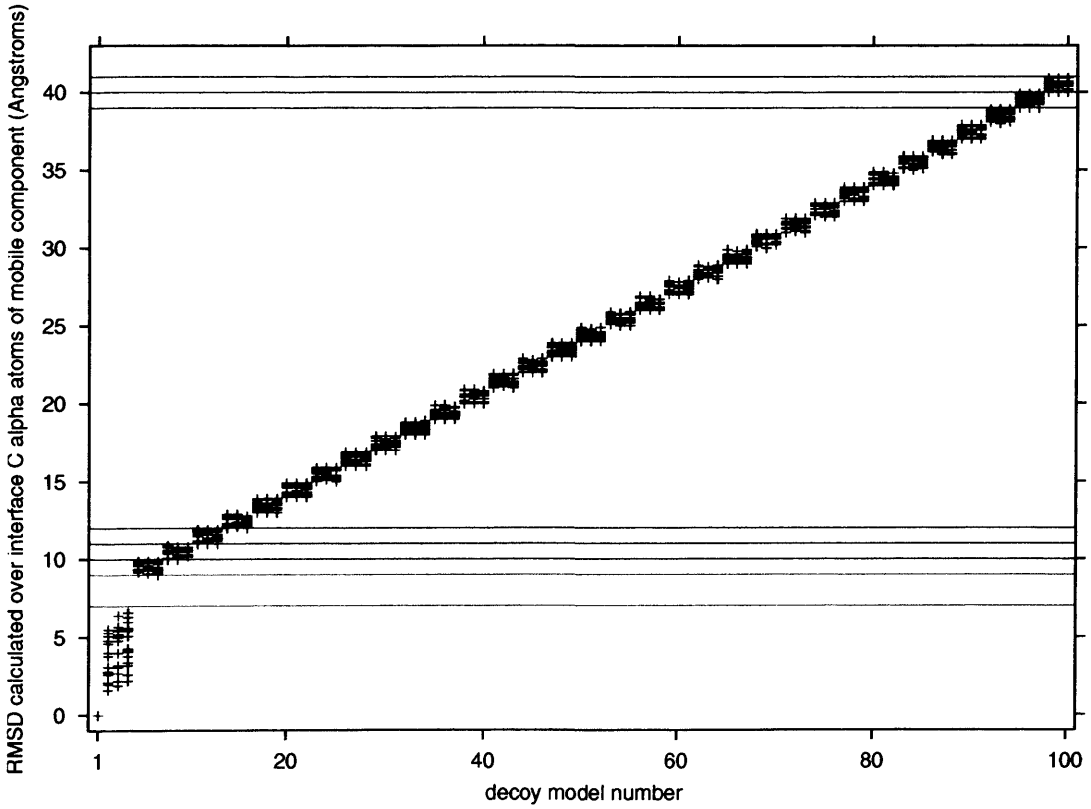


Figure 4.10: Distribution of RMSD values for the decoy sets.

The final results using a non-cross-validated matrix on our decoy sets gave results which are not only very good, but never fail entirely for the native structure. This is always in the top quartile, with the exception of 1BDJ. There is further evidence of pair potentials of giving false negatives less commonly than surface complementarity in the way that the biological filters retain models ranked highly by the pair potentials.

MULTIDOCK, although not developed any further for this work, has still shown itself to be a very useful tool. It is particularly good at picking out near native models. However, it is a slow algorithm and so is in its present form only practical to use as a final step in any methodology.

However, in the course of reimplementing the FTDOCK program, a problem came to light stemming from the fact that the initial orientation of the molecules effects the discretisation models. This problem would indicate that the initial global search step may be the weakest point in the overall strategy described in this chapter.

Test System	Rank of... native	good model best	good model worst
1BRC	5	1	4
1CGI	1	5	23
2KAI	8	3	17
2SIC	1	2	4
1AHW	9	7	15
1BVK	26	28	45
1MLC	7	5	28
1DQJ	4	19	33
1WEJ	3	5	16
2PCC	1	4	20
1BGS	2	3	6
1DFJ	3	43	62
1FSS	1	9	16
1AVZ	3	12	69
1UGH	22	6	23
1WQ1	23	28	53
1BDJ	33	68	85

Table 4.12: Results of using non-cross-validated pair potential matrix on decoy sets.

Chapter 5

Conclusions

The work included in this thesis, and published in journals[7,8] has shown that knowledge based pair potentials can be useful in differentiating good models in protein/protein and protein/DNA docking algorithms. This thesis has also shown the ability of one docking method, the combined approach in Chapter 4, in tackling a large set of dissimilar protein/protein complexes.

Knowledge Based Potentials

The number of docking algorithms and protocols has increased quickly in the last few years, encouraged by the CAPRI competitions. Some of the other algorithms have incorporated knowledge based pair potentials, though none of them have used the exact method as described in Chapter 4.

The major way in which the pair potentials of Glaser[140] and Gray[67] differ is in how to count the residues. Both agree with our method in using a mole fraction calculation for the expected pairings. Gray uses the same dataset of interfaces as Glaser. However, it is not clear if the volume term used by Glaser is used by Gray.

In Section 4.2.3 it was argued that all the surface residues should be counted in order to calculate the expected pairing values. If only the residues in the interface are counted, then the potential loses information about where on the surface it is better to associate. Both Glaser[140] and Gray[67] use only residues in the interface (defined by the areas of protein surface whose accessible surface area changes on association). The algorithm reported by Gray[67] does compensate for this by a separate residue environment term in the potential function. The success of that algorithm could imply that two pair potentials could be used. The first would incorporate information about where on the surface it is better to associate (as used in Chapter 4). The second would give more information about the correct orientation, given that the position on the surface was correct.

Zhou has used a knowledge based potential in round 4 of CAPRI.[176,177]

The form of the potential is more of a model of physical processes based on structural data, than a statistical potential as in this work. Interestingly, the derivation from an ideal gas model, effectively uses a mole fraction calculation when modelling the expected interactions. The end form of the potential is quite different.

The other knowledge based potential used is the atomic contact energies (ACE), developed by Zhang[41] and applied to screening lists of docking models by Camacho[178] and Chen.[40] ACE uses 18 atom types rather than residues. The dataset used is of protein domains. The potential used across an interface is the same potential as used for say protein folding, with the part of the potential describing interactions along the peptide chain ignored. This is reported as not being a problem due to the known ability to discriminate protein/protein interfaces.[41] The calculation of the reference state (expected values) involves a contact fraction part.

Initial Orientation

Two of the more successful algorithms to date in the CAPRI competitions have been ZDOCK (Weng) and SmoothDock (Camacho) (see sections 1.4.2, 1.5.3). Due to the Fourier grid method natures of these methods, and the required discretisation of the structures, they are both susceptible to problems associated with initial orientation (Section 4.5.1). The reason for the higher success rate of these methods against other Fourier grid methods, including FTDOCK, is therefore of interest.

ZDOCK uses a finer rotational scan, 6° as opposed to 12° for FTDOCK. It is unclear what the rotational sampling is for DOT, though it is probably 10° (based on the value used for the ClusPro server). Both these values will result in more rotational sampling than the 12° default of FTDOCK. There is also the chance factor, and even the large number of targets in CAPRI as opposed to previous competitions is maybe not enough to yet rule out statistical variances. The FTDOCK procedure purposefully spins the components before running, and this is a random process, so there is always the chance that none of the three orientations happens to cause a favourable discretisation for the global scans.

Biological knowledge, intervention, and automation

A large number of the groups reporting on the first two rounds of CAPRI in the special issue of Proteins[74] wrote that they used some form of manual intervention. This either took the form of using visualisation tools to check the seeming sensibility of the models they were going to submit, or they used general biological knowledge to discount a model. There is a danger in doing this in that general ideas of what is sensible biologically can only be based on specific

knowledge from previous cases. Even when the new problem is homologous to a system which is known, the small differences can change the form of the binding interface. Using a strict constraint to be like the homologous case may be a mistake. This was reported by, for example, Smith and Sternberg[11] for Target 07 and Camacho and Gatchell[75] for Target 04.

ClusPro is a full automation of the first three steps of the SmoothDock algorithm (everything bar the refinement step). A server has been set up and is running at *http://nrc.bu.edu/cluster/* (April 2004). It succeeded in generating medium (Target 08) and High (Target 12) models in CAPRI. Although the submissions by the Camacho group for these targets were of higher rank, this is a very good result for what is currently the only fully automated method.

Standardisation

There has been a persistent problem when comparing the reported results for different algorithms. Each paper has superimposed different bits of the structure before RMSD calculations, used different atoms for those RMSD calculations (C_α, backbone, all), and defined the interfaces in different ways. It can be hoped that all future papers in the field will adopt the calculations used by CAPRI, and described in Table 1.6, although other measurements could be reported as well.

Further Work

The CAPRI competitions have shown that there is a wide variety of algorithms available now to tackle the problem of computationally modelling protein/protein interactions. The work in this thesis, and CAPRI, has also shown the advantages possible from combining various algorithms. Section 2.3.6 shows that different algorithms can have very different convergence behaviours. It is important that these be known and the knowledge used to order the steps in a procedure correctly.

The available number of targets on which to test algorithms has grown significantly since the work carried out in Chapter 4 was completed. At the end of 2002, just over 30 targets existed with the same criteria that both components existed in an unbound conformation.[179] This has now doubled in just over two years (http://zlab.bu.edu/zdock/benchmark.shtml). It is still noticeable that more than half the targets are either enzyme/inhibitors or antibody/antigens, so limiting the available systems that can be tested to a subset of all the types of biological systems that exist. However, this larger benchmark for an algorithm is of definite benefit. It makes less likely the case of an algorithm being over trained on a limited set of test cases which do not represent the wider biological diversity of interactions.

Computing power has increased dramatically in the last decade, and is now less of a concern. This greatly increases the possibilities of which algorithms to use. Flexibility is a necessity at some stage in any methodology, and the increased computational power puts the algorithms' run times into the realm of hours rather than days. The increased available computer resources also make it more reasonable to introduce flexibility earlier on in procedures. More computing power, in particular cluster architectures, also allows for multiple algorithms to be run and the results brought together in a consensus manner.

Appendix A

Publications

This thesis includes work which has been published in journals, as proceedings, and in books. The following are the publications, in chronological order by publication.

"Modelling repressor proteins binding to DNA"
P. Aloy, G. Moont, H.A. Gabb, E. Querol, F.X. Aviles & M.J.E. Sternberg
Proteins: Structure, Function, and Genetics 33(4):535–549 (1998)

"A computational system for modelling flexible protein-protein and protein-DNA docking"
M.J.E. Sternberg, P. Aloy, H.A. Gabb, R.M. Jackson, G. Moont, E. Querol & F.X. Aviles
Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology (ISMB-98) 6:183–192 (1998)

"Use of Pair Potentials Across Protein Interfaces in Screening Predicted Docked Complexes"
G. Moont, H.A. Gabb & M.J.E. Sternberg
Proteins: Structure, Function, and Genetics 35(3):364–373 (1999)

"Protein-protein docking. Generation and filtering of complexes"
M.J.E. Sternberg, H.A. Gabb, R.M. Jackson & G. Moont
appeared in collection "Protein Structure Prediction: Methods and Protocols (Methods in Molecular Biology Series)" published by Humana Press, pages 399–415, ISBN 0-896-03637-5 (2000)

"Modelling Protein-Protein and Protein-DNA Docking"
M.J.E. Sternberg & G. Moont
appeared in collection "Bioinformatics - From Genomes to Drugs" published by Wiley-VCH, pages 361–404, ISBN 3-527-29988-2 (2002)

Appendix B

Software Manual

The following pages contain the software manual as distributed with the software. This version of the manual differs from the distributed one only in that margin settings have been changed so as to be regulation for this thesis. As a result, some figures have been resized, and the length of the whole thing extends over more pages. There are also some very minor changes to the examples given in the Tutorial section.

The software homepage is http://www.bmm.icnet.uk/docking/. The source code for all programs written by me is available.

Since the software was made public there have been almost 1500 downloads of the RPSCORE part of the program. Since the FTDOCK software is available without registration it is unknown how many times this has been downloaded (though presumably at least as many times).

Openness of algorithms should be encouraged. FTDOCK is still only case of open source code.

B.1 Introduction

3D-Dock is a suite of programs designed to enable computational prediction of protein/protein docking. It does this in several steps, as described in Algorithms (B.2) below. This document is designed to enable the various programs to be run successfully, as well as provide a basic understanding of the underlying algorithms.

Although the suite includes the program MULTIDOCK, it is not covered by this document
(please see *http://www.bmm.icnet.uk/docking/*).

Although this document explains the basics of how the programs work it does not discuss how various parameters or strategies were decided upon. For this information please refer to the published papers in the References at the end of the document.

B.1.1 Key to font usage

To try and make things slightly clearer, different fonts are used in this section to signify different things.

- Normal font is explanation and hence most text.
- `typewriter font is used for program names, things that would be typed on a command line, and things that would be seen when looking in a file.`
- *italics are used for file and directory names*

B.1.2 Requirements

There are several different requirements that have to be met in order to run this suite of programs. These fall into 3 categories; operating system, hardware, and software.

operating system The main programs were written on, and with an aim to running on, a UNIX style operating system. They were actually written on an SGI/IRIX platform, but have also been tested on the easily available Linux, running on an i386 processor. Anything else is not supported, though since the programs are in Perl and C, it is possible that you could compile and run them on something else.

hardware The main limitation to hardware is RAM. `ftdock` uses large amounts of memory, and although you could set the parameters to lower this, standard run of the program will want up to 100 Megabytes of memory. If you do not have this as RAM, the program will be paging constantly and may well take weeks to run.

software You will need a C compiler (though there are unsupported binaries available for SGI and Linux i386), and PERL, version 5.003 or later. The only non-standard C libraries required are those of the fast fourier transform, which you will need to download and compile (see Installation (B.3)).

B.2 Algorithms

This suite of programs is intended to be able to dock two proteins. This means starting from the known structures of two protein subunits of a biological complex known to exist, in unbound conformations, and ending up with a limited set of possible models for the complex. This overall algorithm is here achieved in up to 4 steps.

1. a global scan of translational and rotational space of possible positions of the two molecules, limited by surface complementarity and an electrostatic filter (`ftdock`).
2. an empirical scoring of the possible complexes using residue level pair potentials (`rpscore` (and `rpdock`)).
3. using biological information to screen the possible complexes (`filter`).
4. an energy minimisation and removal of steric clashes on the side-chains of the interface (MULTIDOCK[19]).

The middle two steps are interchangeable in the order in which they are run, and the filter can be run more than once if so desired (see Tutorial). A schematic of the overall approach is shown in Figure B.1.

The `ftdock` algorithm is based on that of Katchalski-Katzir.[14] It discretises the two molecules onto orthogonal grids and performs a global scan of translational and rotational space. In order to scan rotational space it is necessary to rediscretise one of the molecules (for speed the smaller) for each rotation. The scoring method is primarily a surface complementarity score between the two grids, and this is shown in Figure B.2. To speed up the surface complementarity calculations, which are convolutions of two grids, Fourier Transforms are used. This means that the convolutions are replaced with multiplications in Fourier space, and despite having to perform the forward and reverse Fourier Transforms, this decreases the overall computation required. The surface complementarity was the only score used in the original method. The original work on `ftdock` by Gabb[15] found it a useful addition to include an electrostatic filter, and this is again implemented in the current version (though it can be turned off).

The `rpscore` program uses an empirical pair potential matrix to score each possible complex. The pair potentials are at a amino acid residue level. Each potential corresponds to the empirically derived likelihood of a trans-interface pair of two residue types, limited only by a distance cut-off.[7] The present most useful matrix used is generated from 103 non-homologous interfaces found in the PDB with the aid of SCOP 1.50 (*http://scop.mrc-lmb.cam.ac.uk/scop/*). If two interfaces are described as pairings of domains $A - B$ and $C - D$, then a non-homologous interface is defined as being when either A and C, or B and D, are homologous, but not both. Homology is in this case defined as being in the same 'Superfamily' in the SCOP classification tree.

The biological filter is a simple program to screen the complexes by requiring them to have a given chain or residue on one side of the interface within a certain distance of another chain or residue on the other side. The manual (B.5.6) explains this in full.

For the manual and program MULTIDOCK please see *http://www.bmm.icnet.uk/docking/*. This calculates side-chain energy minimisations and removes steric clashes along the interface. It is presently only available as an IRIX5.3 executable.

B.3 Installation

You need to download two files: *fftw-2.1.3.tar.gz*, and once you have registered, *ftdock.tar.gz*.

The first thing is to compile FFTW. You do not need to install it. To do this you should not have to do more than

`gunzip fftw-2.1.3.tar.gz` — gives you file *fftw-2.1.3.tar*
`tar xvf fftw-2.1.2.tar` — makes a directory *fftw-2.1.3* with all the bits inside it.

Change into that directory then

```
./configure -enable-float
make
```

This is all you need to do. There is of course no harm in installing it properly. The reason for using `-enable-float` is to reduce (typically halve) the memory requirements. If you are going to use FFTW for other programs you may need to consider if you want this.

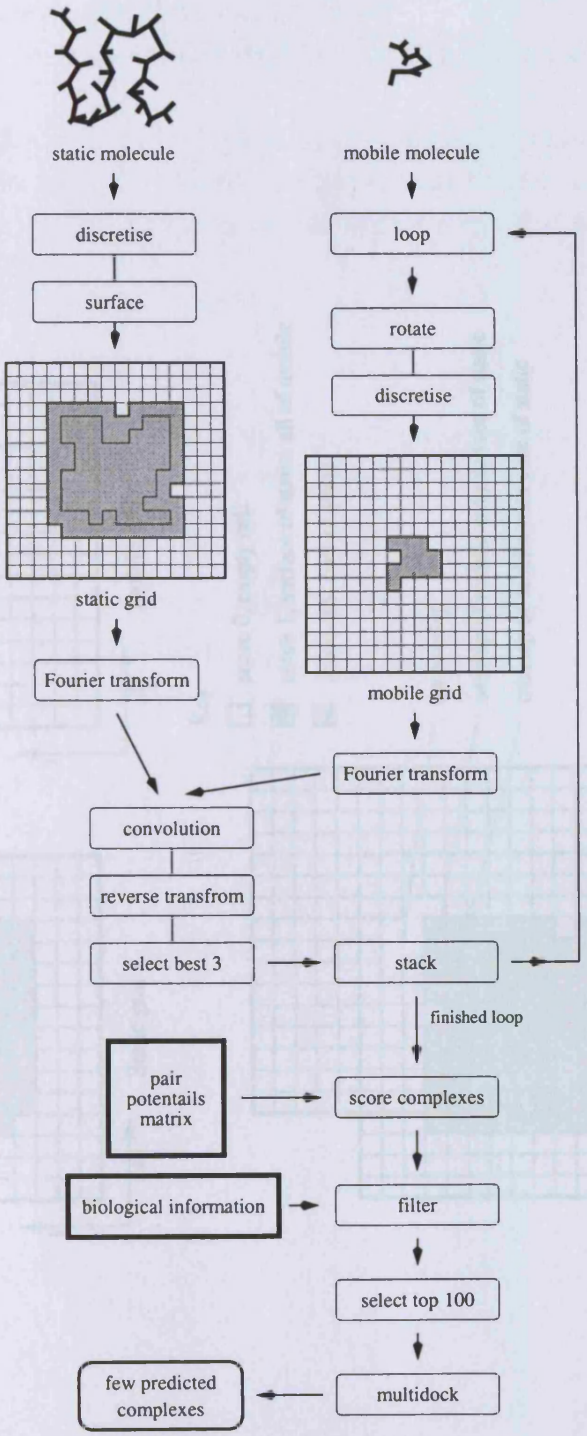


Figure B.1: Flow diagram of overall docking method.

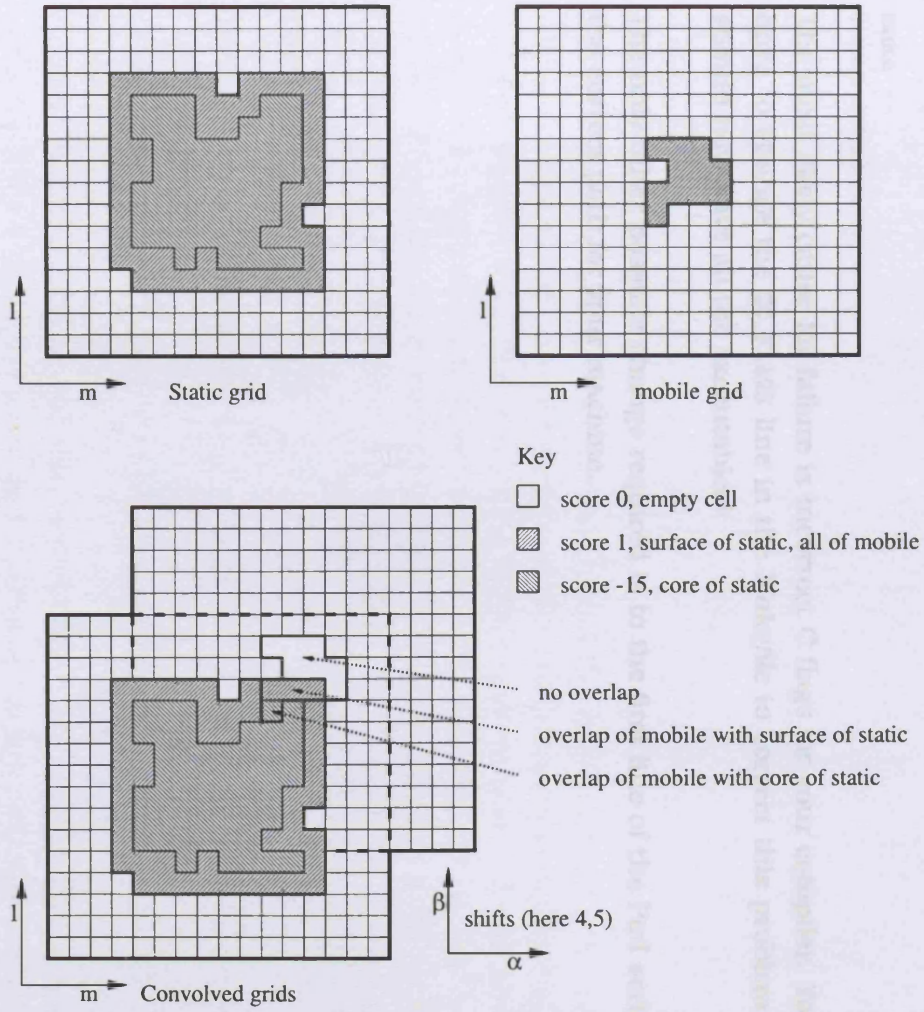


Figure B.2: Grid discretisation of molecules and calculation of surface complementarity.

Once you have done this, you can compile the actual programs

`gunzip ftdock.tar.gz` — gives you file *ftdock.tar*
`tar xvf ftdock.tar` — makes a directory *ftdock* with all the bits inside it.

Change into that directory, and then into the *progs* directory. You will have to edit the *Makefile* to give the correct complete path to the *fftw-2.1.3* directory. (If you have fully installed FFTW, you will need to edit the *Makefile* to put the correct paths in.) Then

```
make
```

The most likely cause for failure is incorrect C flags for your compiler. You will have to change the `CC_FLAGS` line in the *Makefile* to correct this problem. You should now have all the executables!

The only other possible change required is to the first line of the Perl scripts to the correct path for your machine.

B.4 Tutorial

This tutorial will take the example of bovine pancreatic trypsin inhibitor bound to kallikrein A complex. The necessary PDB files are included in the distribution. You will have to give the full paths to the various executables as appropriate. This tutorial uses the minimum number of options for each program. For complete options and further details please see the manuals section (B.5) below.

All this tutorial presumes you are executing all the programs in the same directory, and not changing the names of any files produced.

Preprocessing

Manually edit the PDB files so that you have the components you want to dock. Then

```
preprocess-pdb.perl -pdb file.pdb
```

This will give any number of messages, normally complaining of non-standard residue designations. I do not recommend you use this program indiscriminately for other work as it removes everything but the `ATOM` records of the 20 standard residues it recognises, and it also removes Hydrogens and `OXT` records as well. The output will have the name *file.parsed*. It will also produce a FASTA format file called *file.fasta* which you may find useful.

Global scan

To run the main program type

```
ftdock -static 2pka.parsed -mobile 5pti.parsed > output &
```

I recommend you redirect the standard out for safety reasons. The program is going to take a long while to run, and it will want to write out stuff throughout. If you want to be able to close the shell without crashing the program, you need to do this. In order to see what is going on, the following UNIX command is ideal

```
tail -f output
```

The output you will now have is the file *ftdock_global.dat*, which will contain 10000 records.
(best rank on my run = 1619)

Pair Potential scoring

In order to assign a pair potential score to each record you should type

```
rpscore
```

This very simple command will only work if you also have the file *best.matrix* in the current directory.

The output is *ftdock_rpscored.dat*, which contains the same 10000 records, but reordered by the new score.
(best rank on my run = 65)

Filtering

As is often the case, we have biological information which can reduce the number of possibilities. We want to filter such that the remaining complexes have the inhibitor (chain I) in proximity (distance default is 4.5 Angstroms) to the catalytic triad of the enzyme (chains A and B). This is expressed as

```
filter -constraints A57:I B102:I B195:I
```

Each constraint is treated as an **OR** statement. The designators each side of the colon are of the form chainID then residue number (+ insertion code if defined), and for the whole chain, the residue number is simply missed out. The order is irrelevant, so

```
filter -constraints I:B195 I:A57 B102:I
```

would give the same output. For more explanations see the manuals section (B.5) below.

The output is *ftdock_filtered.dat*, which contains a reduced set of records.
(in my run 900)
(best rank on my run = 12)

In order to have the effect of an **AND** statement, you will have to run the filter program several times.

```
filter -constraints A57:I -out ftdock_filter_A.dat
```

then

```
filter -constraints B102:I -in ftdock_filter_A.dat -out ftdock_filter_B.dat
```

then

```
filter -constraints B195:I -in ftdock_filter_B.dat -out ftdock_filtered.dat
```

(best rank on my run = NA . This can often happen that a too strict series of constraints will loose good results)

Side-chain refinement

For the manual and program MULTIDOCK please see *http://www.bmm.icnet.uk/docking/*. This calculates side-chain energy minimisations and removes steric clashes along the interface. It is presently only available as an IRIX5.3 executable.

B.5 Manuals

B.5.1 `preprocess-pdb.perl`

Due to the nature of PDB files, a preprocessor is used to both clean up and add limited information to the PDB files. The cleaning method is described below. The added information is simply the one letter amino acid codes, and a numerical assignment for each residue type, assigned in alphabetical order (1–20).

what the cleaner does

1. removes all residues that are not one of the twenty standard amino acids or one of the five standard nucleic acids.
2. only keeps atoms it recognises as 'useful' - so removes all Hydrogen atoms. It also removes 'OXT' - terminal Oxygens, simply because their assignment is not always sensible.
3. removes all but the first of an alternative atom indicator entry.
4. checks for the correct number of atoms for that residue, then
   - if too many, checks for doubles of any atom type labels and removes all but first (ie copes with missing Alternate Indicator).
   - if still too many atoms for residue, then checks for atom type validity for that residue type.
   - if still too many, will chuck (remove) that residue.
   - if too few, will do nothing, unless MULTIDOCK is set, in which case it will attempt to replace with a modelled Alanine.

Command line options

`-pdb` PDB style file name
no default

`-nowarn` turns off all but the most severe warnings

`-multidock` this makes the output fit for input into the program MULTIDOCK
this is not for use prior to running ftdock
it will change to model Alanine any residue which does not contain its full complement of (non-Hydrogen) atoms

B.5.2 change-pdb-chain-id.perl

A script to change PDB ChainIDs.

Command line options

-pdb	PDB style file name no default
-old	chain ID that you want to change for a non labelled chain use ' ' no default
-new	replacement chain ID that you want for a non labelled chain use ' ' no default

examples

```
change-pdb-chain-id.perl -pdb 2pka.pdb -old A -new E
change-pdb-chain-id.perl -pdb 1hpt.pdb -old ' ' -new I
```

B.5.3 ftdock

The main global docking program. Due to the rescue abilities, please do not run this in a given directory more than once at any one time.

Command line options

-out	output file name default is *ftdock_global.dat*
-static	larger of the two molecules being docked this PDB style file must be output from preprocess no default
-mobile	smaller of the two molecules being docked this PDB style file must be output from preprocess no default

`-grid` number of grid units in one dimension
this means a grid of 64 has 64^3 grid units in total
this also means that memory requirements go roughly as n^3 of grid size
a grid that results in a grid spacing of more than 1 angstrom is unlikely to be useful
the grid size must be integer and even (to ease Fourier calculations)
no default

`-calculate_grid` the desired size of a single grid unit in angstroms
due to the limitations on the grid size, the actual grid unit will vary slightly (less than ± 0.01) from the given value
default is on with a value of 0.875
to turn off, use `-grid` option

`-angle_step` the maximum planar angle (in degrees) separating any two rotations of the mobile molecule when subtended to the point around which the rotation takes place (geometric centre of the mobile molecule)
default is 12 degrees
will only accept integer values that are integer factors of 180

`-surface` surface thickness in angstroms
default is 1.5

`-internal` internal deterrent value
default is -15.0

`-noelec` electrostatics calculations switch
default is to do the electrostatics, this switch will turn them off

`-keep` number of (best surface complementarity) translations to keep from each rotation
default is 3

`-rescue` if your machine falls over, then just run `ftdock -rescue` in the same directory
do not alter anything between the crash and using this
to make this option available in this very simple form, two files exist in the directory from which you run `ftdock`; namely *scratch_parameters.dat* and *scratch_scores.dat*. This means that you should not run `ftdock` more than once at any given time in the same directory. There is no system at present to prevent this from being done, so be careful.

Understanding the output

```
FTDOCK data file

Global Scan

Command line controllable values
Static molecule                  :: static.parsed
Mobile molecule                  :: mobile.parsed

Global grid size                 ::    110      (default calculated)
Global search angle step         ::     12      (default)
Global surface thickness         ::      1.40   (default)
Global internal deterrent value  ::    -15.00   (default)
Electrostatics                   ::     on      (default)
Global keep per rotation         ::      3      (default)

Calculated values
Global rotations                 ::   9240
Global total span (angstroms)    ::     96.079
Global grid cell span (angstroms) ::     0.873

Data
Type       ID    prvID    SCscore       ESratio        Coordinates          Angles

G_DATA      1        0        173         13.992        22   19   -7      120  96 228
G_DATA      2        0        173          1.941       -27    4   -4      144  96 336
G_DATA      3        0        163         17.331       -23   10  -11      180  84 156
```

Output 1: Example output from `ftdock`

Output 1 shows a typical output file to a run of `ftdock` (default *ftdock_global.dat*) or defined by the `-out` option). All values that can be controlled by the command line (apart from the output file name) are shown at the top of the file. Along with each value is information showing whether it has been chosen or is the default value (apart from for the molecules which are required and have no default

values). These lines must not be changed since the values are used by other programs which use this file for input. In general, it is suggested that any data files produced by any of the programs should not be edited directly, and there should be no need to do so.

Below this are shown a few calculated values. These are purely for the users information and are never used by any of the other programs.

The `G_DATA` lines contain all the information corresponding to each putative complex. The `ID` is that used for the `build` program (see B.5.7 below). The previous ID (`prvID`) is zero in this case as this is the first program. The Surface Complementarity Score (`SCscore`) is the value which determines the order of the file, the highest score having the lowest ID. The electrostatic score ratio (`ESratio`) is there to possibly show varying electrostatic favourability when a group of complexes have the same surface complementarity. It is a ratio as opposed to an absolute value, ranging from 0 (least favourable) to 100 (most favourable). After this come the translational coordinates (x, y, z) expressed as integer grid cell displacements of the mobile molecule's centre from the centre of the static molecule. At the end come the rotational angles $(z_{twist}, \theta, \phi)$ expressed in degrees.

B.5.4 `rpscore`

The residue level pair potentials scoring program.

Command line options

`-in`	input file name default is *ftdock_global.dat*
`-out`	output file name default is *ftdock_rpscored.dat*
`-matrix`	matrix file name default is *best.matrix* this can be found in the data directory

Understanding the output

Output 2 shows a typical output file to a run of `rpscore` (default *ftdock_rpscored.dat* or defined by the `-out` option). All the information from the run of `ftdock` and

```
FTDOCK data file

Global Scan

Command line controllable values
Static molecule                    :: static.parsed
.
.
.

Global grid cell span (angstroms)  ::      0.873

******

Residue level Pair Potential Scoring

Command line controllable values
Matrix                             :: /home/ftdock/data/best.matrix    (user defined)

Data
Type       ID    prvID    SCscore        RPscore          Coordinates              Angles

G_DATA      1      502        139         14.778        12   36   14        276  24 120
G_DATA      2     5839        114         14.011        37   -5    7         96  60 276
G_DATA      3       21        151         13.196        12   36   14        276 156 120
```

Output 2: Example output from `rpscore`

any previous runs of `rpscore` or `filter` are still at the top of the file, followed by the command line controllable matrix.

The `G_DATA` lines contain all the information corresponding to each putative complex. The fields are identical to those in the output from `ftdock` with the exception of `RPscore` which replaces `ESratio`. The complexes are now ordered by their residue level pair potential scores (`RPscore`), and the `prvID` field has values corresponding to the `ID` field in the input data file. The `prvID` field can be used to track the ranking of a complex as the successive programs are run.

B.5.5 `rpdock`

The residue level pair potentials scoring program for use with complexes generated by another docking program apart from `ftdock`.

Command line options

`-static` PDB style file of one side of the complex. Must have been parsed with `pre-process.perl` .
no default

`-mobile` PDB style file of the other side of the complex. Must have been parsed with `pre-process.perl`.
no default

`-matrix` matrix file name
default is *best.matrix*
this can be found in the data directory

Understanding the output

The program returns a line of the form

```
G_DATA     -3.646
```

to standard out. To screen a list of complexes it is advised to write a perl script wrapper.

B.5.6 `filter`

The biological filter program.

Command line options

`-in` input file name
default is *ftdock_rpscored.dat*

`-out` output file name
default is *ftdock_filtered.dat*

`-distance` the inter-atomic distance cut-off (in angstroms) for determining whether the residues, of which a given two atoms are members of, are in contact or not.
default is 4.5

`-constraints` a space separated list of the form
`chainID[residuenumber][Icode]:chainID[residuenumber][Icode]`
easiest explained by example

1. residue 45 of chain A to be in contact with chain B
   `A45:B`
2. residue 45 of chain A to be in contact with residue 3 of chain B
   `A45:B3`
3. residue 45 , insertion code A, of chain E to be in contact with chain I
   `E45A:I`

the list is treated as a set of logical OR statements, so if any are satisfied, the statement is satisfied.
there is a limit of 50 constraints
if no constraints are given, the program will simply not run

Understanding the output

Output 3 shows a typical output file to a run of `rpscore` (default *ftdock_filtered.dat* or defined by the `-out` option). All the information from the run of `ftdock` and any previous runs of `rpscore` or `filter` are still at the top of the file, followed by the command line controllable matrix.

The `G_DATA` lines contain all the information corresponding to each putative complex. The fields are identical to those in the output from `rpscore`. The complexes are ordered by their residue level pair potential scores (`RPscore`), and the `prvID` field has values corresponding to the `ID` field in the input data file.

B.5.7 `build`

The program to build a complex or a range of complexes.

Command line options

`-in` input file name
default is *ftdock_rpscored.dat*

`-b0` single complex number to build
no default

```
FTDOCK data file

Global Scan

Command line controllable values
Static molecule                   :: static.parsed
.
.
.

Matrix                            :: /home/ftdock/data/best.matrix    (user defined)

******

Filter

Command line controllable values
Constraints                       :: A57:I B102:I B195:I (3)
Distance                          ::      4.50   (default)

Data
Type       ID    prvID    SCscore       RPscore         Coordinates            Angles

G_DATA      1        1        139        14.778        12   36   14       276  24 120
G_DATA      2        3        151        13.196        12   36   14       276 156 120
```

Output 3: Example output from `filter`

`-b1` beginning of range of complex numbers to build
default is 1

`-b2` end of range of complex numbers to build
default is 10000

`-c_alpha` build only the Cα atoms

Understanding the output

The outputs from this program are the modelled complexes in PDB format. (There is extra information beyond column 80, but this should not cause problems to other programs such as visualisation tools.) The complexes are called *Complex_xg.pdb*, where x corresponds to the record ID number in the input file. If the `-c_alpha` option is used, this changes to *CA_Complex_xg.pdb*.

B.5.8 randomspin

A program that randomly spins a PDB file. This is useful for testing the stability of a docking algorithm with respect to the initial orientations of the molecules.

Command line options

`-in`	input file name default is *unspun.pdb*
`-out`	output file name default is *spun.pdb*

B.5.9 centres

A program to visualise the spread of positions of the mobile molecule with respect to the static molecule. The output PDB file *centres.pdb* contains the static molecule and water molecules, each of which represents the central position of the mobile molecule for each complex. Can show clustering of results when it occurs.

Command line options

`-in`	input file name default is *ftdock_global.dat*

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acid Res.*, 28:235–242, 2000.

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

[3] M.J. Betts and M.J.E. Sternberg. An analysis of conformational changes on protein-protein docking: implications for predictive docking. *Prot. Eng.*, 12:271–283, 1999.

[4] L.L. Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, 285:2177–2198, 1999.

[5] E.J. Sundberg and R.A. Mariuzza. Luxury accommodations: the expanding role of structural plasticity in protein-protein interactions. *Structure Fold Des*, 8:137–142, 2000.

[6] PM Colman. Structure-based drug design. *Current Opinion Structural Biol.*, 4:868–874, 1994.

[7] G. Moont, H.A. Gabb, and M.J.E. Sternberg. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, 35:364–373, 1999.

[8] P. Aloy, G. Moont, H.A. Gabb, E. Querol, F.X. Aviles, and M.J.E. Sternberg. Modeling repressor proteins binding to DNA. *Proteins*, 33:535–549, 1998.

[9] M. J. Sternberg, H. A. Gabb, R. M. Jackson, and G. Moont. Protein-protein docking. Generation and filtering of complexes. In D. Webster, editor, *Protein Structure Prediction: Methods and Protocols (Methods in Molecular Biology Series)*. Humana Press, 2000.

[10] M. J. E. Sternberg and G. Moont. Modelling protein-protein and protein-DNA docking. In T. Lengauer, editor, *Bioinformatics - From Genomes to Drugs*, pages 361–404. Wiley-VCH, 2002.

[11] G. R. Smith and M. J. E. Sternberg. Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial. *Proteins*, 52(1):74–79, 2003.

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

[13] J. Janin. Protein-protein recognition. *Prog. Biophys. Molec. Biol.*, 64(2/3):145–166, 1995.

[14] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and S. J. Wodak. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Nat. Acad. Sci.*, 89:2195–2199, 1992.

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

[16] E.E. Lattman. Optimal sampling of the rotation function. In M. G. Rossman, editor, *The Molelcular Replacement Method*, pages 179–185. Gordon and Breach, New York, 1972.

[17] B.E. Hingerty, R.H. Ritchie, T.L. Ferrell, and J.E. Turner. Dielectric effects in bio-polymers the theory of ionic saturation revisited. *Biopolymers*, 24:427–439, 1985.

[18] D.T. Edmonds, N.K. Rogers, and M.J.E. Sternberg. Regular representation of irregular charge distribution. applications to the electroststic potentials of globular proteins. *Molecular Physics*, 52:1487–1494, 1984.

[19] R. M. Jackson, H. A. Gabb, and M. J. E. Sternberg. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.*, 276(1):265–285, 1998.

[20] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side-chain conformations. *J. Biomol. Struct. Dynam.*, 8:1267–1289, 1991.

[21] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

[22] D. Sitkoff, K.A. Sharp, and B. Honig. Accurate calculation of hydration free energies using macromolecular solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.

[23] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, 239:249–275, 1994.

[24] C. Lee. Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.*, 236:918–939, 1994.

[25] H.J. Hecht, M. Szardenings, J. Collins, and D. Schomburg. 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(3):711–722, 1991.

[26] M. Fujinaga, A.R. Sielecki, R.J. Read, W. Ardelt, M. Laskowski Jr., and M.N. James. Crystal and molecular structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8å resolution. *J. Mol. Biol.*, 195(2):397–418, 1987.

[27] W. Bode, Z. Chen, K. Bartels, C. Kutzbach, G. Schmidt-Kastner, and H. Bartunik. 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(2):237–282, 1983.

[28] R. Huber, D. Kukla, W. Bode, P. Schwager, K. Bartels, J. Deisenhofer, and W. Steigemann. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. II. Crystallographic refinement at 1.9å resolution. *J. Mol. Biol.*, 89(1):73–101, 1974.

[29] C.A. McPhalen and M.N. James. Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and Ci-2-subtilisin Novo. *Biochem.*, 27(17):6582–6598, 1988.

[30] T.O. Fischmann, G.A. Bentley, T.N. Bhat, G. Boulot, R.A. Mariuzza, S.E. Phillips, D. Tello, and R.J. Poljak. Crystallographic refinement of the three-dimensional structure of the FabD1.3-lysozyme complex at 2.5å resolution. *J. Biol. Chem.*, 266(20):12915–12920, 1991.

[31] B.C. Braden, H. Souchon, J.L. Eisele, G.A. Bentley, T.N. Bhat, J. Navaza, and R.J. Poljak. Three-dimensional structures of the free and the antigen-complexed Fab from monoclonal anti-lysozyme antibody D44.1. *J. Mol. Biol.*, 243(4):767–781, 1994.

[32] D.R. Davies, E.A. Padlan, and S. Sheriff. Antibody-antigen complexes. *Annu. Rev. Biochem.*, 59:439–473, 1990.

[33] E.W. Silverton, E.A. Padlan, D.R. Davies, S. Smith-Gill, and M. Potter. Crystalline monoclonal antibody Fabs complexed to hen egg white lysozyme. *J. Mol. Biol.*, 180(3):761–765, 1984.

[34] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical recipes in FORTRAN - The art of scientific computing.* Cambridge University Press, Cambridge, 2nd edition, 1992.

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

[36] I.A. Vakser, O.G. Matar, and C.F. Lam. A systematic study of low-resolution recognition in protein-protein complexes. *Proc. Nat. Acad. Sci.*, 96:8477–8482, 1999.

[37] P.N. Palma, L. Krippahl, J.E. Wampler, and J.J.G. Mora. BiGGER: A new (soft) docking algorithm for predicting protein interactions. *Proteins*, 39:372–384, 2000.

[38] D.W. Ritchie and G.J.L. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins*, 39:178–194, 2000.

[39] E. Ben-Zeev and M. Eisenstein. Weighted geometric docking: Incorporating external information in the rotation-translation scan. *Proteins*, 52(1):24–27, 2003.

[40] R. Chen, L. Li, and Z. Weng. ZDOCK: An initial-stage protein-docking algorithm. *Proteins*, 52(1):80–87, 2003.

[41] C. Zhang, G. Vasmatzis, J. L. Cornette, and C. DeLisi. Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, 267(4):707–726, 1997.

[42] L. Li, R. Chen, and Z. Weng. RDOCK: Refinement of rigid-body protein docking predictions. *Proteins*, 53:693–707, 2003.

[43] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

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

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

[46] D.K. Hendrix, T.E. Klien, and I.D. Kuntz. Macromolecular docking of a three-body system: The recognition of human growth hormone by its receptor. *Prot. Sci.*, 8:1010–1022, 1999.

[47] D. Fischer, R. Norel, H. Wolfson, and R. Nussinov. Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition. *Proteins*, 16:278–292, 1993.

[48] D. Fischer, S.L. Lin, H.J. Wolfson, and R. Nussinov. A suite of molecular docking processes. *J. Mol. Biol.*, 248:459–477, 1995.

[49] R. Norel, D. Petrey, H.J. Wolfson, and R. Nussinov. Examination of shape complementarity in docking of unbound proteins. *Proteins*, 36:307–317, 1999.

[50] H.P. Lenhof. New contact measures for the protein docking problem. *RECOM97 - Proceedings of the first annual international conference on computational molecular biology*, ACM, 1997.

[51] E. Althaus, O. Kohlbacher, H. P. Lenhof, and P. Muller. A combinatorial approach to protein docking with flexible side-chains. *RECOM2000 - Proceedings of the fourth international conference on computational molecular biology*, ACM, 2000.

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

[53] M. Totrov and R. Abagyan. Detailed *ab initio* prediction of lysozyme-antibody complex with 1.6å accuracy. *Nature Struct. Biol.*, 1:259–263, 1994.

[54] J. Fernández-Recio, M. Totrov, and R. Abagyan. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*, 52(1):113–117, 2003.

[55] J. Fernández-Recio, M. Totrov, and R. Abagyan. Soft protein-protein docking in internal coordinates. *Prot. Sci.*, 11:280–291, 2002.

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

[57] M. K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins*, 4(7):7–18, 1988.

[58] F. M. Richards. Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

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

[60] R.M. Jackson and M.J.E. Sternberg. Application of scaled particle theory to model the hydrophobic effect: implications for molecular association and protein stability. *Prot. Eng.*, 7:371–383, 1994.

[61] C.H. Robert and J. Janin. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J. Mol. Biol.*, 283:1037–1047, 1998.

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

[63] R.E. Bruccoleri and J. Novotny. Antibody modeling using the conformational search program CONGEN. *Immunomethods*, 1:96–106, 1992.

[64] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

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

[66] S. Vajda, Z. Weng, R. Rosenfeld, and C. DeLisi. Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochem.*, 33:13977–13988, 1994.

[67] J. J. Gray, S. Moughon, T. Kortemme, O. Schueler-Furman, K. M. S. Misura, A. V. Morozov, and D. Baker. Protein-protein docking predictions for the CAPRI experiment. *Proteins*, 52(1):118–122, 2003.

[68] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, 331:281–299, 2003.

[69] N.C. Strynadka, M. Eisensten, B. Shoichet, Kunts T., Duncan R., A. Olson, R. Abagyan, M. Totrov, R. Jackson, M. Sternberg, J. Cherfils, J. Janin, and James M.N. Current molecular docking programs successfully predict a large protein-protein complex. *Nature Struct. Biol.*, 3:233–239, 1996.

[70] B.S. Duncan and A.J. Olson. Approximation and characterization of molecular surfaces. *Biopolymers*, 33:219–229, 1993.

[71] S.J. Dixon. Evaluation of the CASP2 docking section. *Proteins*, Supplement 1:198–204, 1997.

[72] I.A. Vakser. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins*, Supplement 1:226–230, 1997.

[73] D.M. Webster and A.R. Rees. Macromolecular recognition: antibody-antigen complexes. *Prot. Eng.*, 65:94, 1993.

[74] R. Méndez, R. Leplae, L. De Maria, and S. J. Wodak. Assesment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins*, 52(1):51–67, 2003.

[75] C. J. Camacho and D. W. Gatchell. Successful discrimination of protein interactions. *Proteins*, 52(1):92–97, 2003.

[76] J. Janin. The kinetics of protein-protein recognition. *Proteins*, 28:153–161, 1997.

[77] C. Zhang, J. Chen, and C. DeLisi. Protein-protein recognition: Exploring the energy funnels near the binding sites. *Proteins*, 34:255–267, 1999.

[78] C.J. Tsai, S. Kumar, B.Y. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Prot. Sci.*, 8:1181–1190, 1999.

[79] M.D. Cummings, T.N. Hart, and R.J. Read. Atomic solvation parameters in the analysis of protein-protein docking results. *Prot. Sci.*, 4(10):2087–2099, 1995.

[80] A. Wallqvist and D.G. Covell. Docking enzyme-inhibitor complexes using a preference-based free-energy surface. *Proteins*, 25(4):403–419, 1996.

[81] S. Tanaka and H.A. Scheraga. Medium and long range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):45–950, 1976.

[82] S. Miyazawa and R.L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.*, 256:623–644, 1996.

[83] M.J. Sippl. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213(4):85–883, 1990.

[84] A. Godzik, A. Kolinski, and J. Skolnick. Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, 227(1):227–238, 1992.

[85] S.H. Bryant and C.E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins*, 16(1):92–112, 1993.

[86] A.E. Torda. Perspectives in protein-fold recognition. *Current Opinion Structural Biol.*, 7(2):200–205, 1997.

[87] S. Vajda, M. Sippl, and J. Novotny. Empirical potentials and functions for protein folding and binding. *Current Opinion Structural Biol.*, 7:222–228, 1997.

[88] B.H. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.*, 249(2):493–507, 1995.

[89] E.S. Huang, S. Subbiah, and M. Levitt. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.*, 252(5):709–720, 1995.

[90] D.T. Jones, W.R. Taylor, and J.M. Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–89, 1992.

[91] D.A. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243(4):668–682, 1994.

[92] A. Godzik, A. Kolinski, and J. Skolnick. Lattice representations os globular proteins: How good are they? *J. Comp. Chem.*, 14:1194–1202, 1993.

[93] P.D. Thomas and K.A. Dill. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.*, 257:457–469, 1996.

[94] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Prot. Sci.*, 6:676–688, 1997.

[95] M.N. James and A.R. Sielecki. Structure and refinement of penicillopepsin at 1.8å resolution. *J. Mol. Biol.*, 163(2):299–361, 1983.

[96] S.T. Freer, J. Kraut, J.D. Robertus, H.T. Wright, and N.H. Xuong. Chymotrypsinogen: 2.5å crystal structure, comparison with alpha-chymotrypsin, and implications for zymogen activation. *Biochem.*, 9(9):1997–2009, 1970.

[97] M.W. Empie and M. Laskowski Jr. Thermodynamics and kinetics of single residue replacements in avian ovomucoid third domains: effect on inhibitor interactions with serine proteinases. *Biochem.*, 21(10):2274–2284, 1982.

[98] R.A. Blevins and A. Tulinsky. Comparison of the independent solvent structures of dimeric alpha-chymotrypsin with themselves and with gamma-chymotrypsin. *J. Biol. Chem.*, 260(15):8865–8872, 1985.

[99] S. Parkin, B. Rupp, and H. Hope. Structure of bovine pancreatic trypsin inhibitor at 125 k: Definition of carboxyl-terminal residues Gly57 and Ala58. *Acta Crystallogr D*, 52(Part 1):18–29, 1996.

[100] R. Huber, D. Kukla, A. Ruhlmann, and W. Steigemann. Pancreatic trypsin inhibitor (Kunitz). I. Structure and function. *Cold Spring Harb Symp Quant Biol*, 36:141–148, 1972.

[101] W. Bode and P. Schwager. The refined crystal structure of bovine beta-trypsin at 1.8å resolution. II. Crystallographic refinement, calcium binding site, benzamidine binding site and active site at pH 7.0. *J. Mol. Biol.*, 98(4):693–717, 1975.

[102] G.M. Clore, A.M. Gronenborn, M.N. James, M. Kjaer, C.A. McPhalen, and F.M. Poulsen. Comparison of the solution and X-ray structures of barley serine proteinase inhibitor 2. *Prot. Eng.*, 1(4):313–318, 1987.

[103] D.T. Gallagher, J.D. Oliver, R. Bott, C. Betzel, and G.L. Gilliland. Subtilisin Bpn' at 1.6 ångstroms resolution: Analysis of discrete disorder and comparison of crystal forms. *Acta Crystallogr D*, 52(Part 6):1125–1135, 1996.

[104] K. Maenaka, M. Matsushima, H. Song, F. Sunada, K. Watanabe, and I. Kumagai. Dissection of protein-carbohydrate interactions in mutant hen egg-white lysozyme complexes and their hydrolytic activity. *J. Mol. Biol.*, 247(2):281–293, 1995.

[105] T.N. Bhat, G.A. Bentley, T.O. Fischmann, G. Boulot, and R.J. Poljak. Small rearrangements in structures of Fv and Fab fragments of antibody D1.3 on antigen binding. *Nature*, 347(6292):483–485, 1990.

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

[107] A. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a Structural Classification of Proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[108] F. Melo and E. Feytmans. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, 267:207–222, 1997.

[109] M.J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des.*, 7(4):473–501, 1993.

[110] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accesability. *J. Mol. Biol.*, 55:379–400, 1971.

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

[112] D. Rhodes, J.W. Schwabe, L. Chapman, and L. Fairall. Towards an understanding of protein-DNA recognition. *Philos Trans R Soc Lond B Biol Sci.*, 351(1339):501–509, 1996.

[113] M.H. Werner, A.M. Gronenborn, and G.M. Clore. Intercalation, DNA kinking, and the control of transcription. *Science*, 271(5250):778–784, 1996 (Erratum in *Science*, 276(5321):1957 1997).

[114] B.E. Raumann, M.A. Rould, C.O. Pabo, and R.T. Sauer. DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. *Nature*, 367(6465):754–757, 1994.

[115] J.N. Breg, J.H. van Opheusden, M.J. Burgering, R. Boelens, and R. Kaptein. Structure of arc repressor in solution: evidence for a family of beta-sheet DNA-binding proteins. *Nature*, 364(6284):586–589, 1990.

[116] A. Mondragon, C. Wolberger, and S.C. Harrison. Structure of phage 434 Cro protein at 2.35å resolution. *J. Mol. Biol.*, 205(1):179–188, 1989.

[117] R. Marmorstein, M. Carey, M. Ptashne, and S.C. Harrison. DNA recognition by GAL4: structure of a protein-DNA complex. *Nature*, 356(6368):408–414, 1992.

[118] M. Lewis, G. Chang, N.C. Horton, M.A. Kercher, H.C. Pace, M.A. Schumacher, R.G. Brennan, and P. Lu. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, 271(5253):1247–1254, 1996.

[119] M. Slijper, A.M. Bonvin, R. Boelens, and R. Kaptein. Refined structure of lac repressor headpiece (1-56) determined by relaxation matrix calculations from 2D and 3D NOE data: change of tertiary structure upon binding to the lac operator. *J. Mol. Biol.*, 259(4):761–773, 1996.

[120] L.J. Beamer and Pabo C.O. Refined 1.8å crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.*, 227(1):177–196, 1992.

[121] W.S. Somers and S.E. Phillips. Crystal structure of the met repressor-operator complex at 2.8å resolution reveals DNA recognition by beta-strands. *Nature*, 359(6394):387–393, 1992.

[122] M.A. Schumacher, K.Y. Choi, H. Zalkin, and R.G. Brennan. Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science*, 266(5186):763–770, 1994.

[123] Z. Otwinowski, R.W. Schevitz, R.G. Zhang, C.L. Lawson, A. Joachimiak, R.Q. Marmorstein, B.F. Luisi, and P.B. Sigler. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, 335(6188):321–329, 1988 (Erratum in *Nature*, 335(6193):837 1988).

[124] R.M.A. Knegtel, R. Boelens, and R. Kaptein. Monte Carlo docking of protein-DNA complexes: incorporation of DNA flexibility and experimental data. *Prot. Eng.*, 7:761–767, 1994.

[125] R.M.A. Knegtel, J. Antoon, C. Rullmann, R. Boelens, and R. Kaptein. MONTY: a Monte Carlo approach to protein-DNA recognition. *J. Mol. Biol.*, 235:318–324, 1994.

[126] G. Campbell, Y. Deng, Glimm J., Eisenberg M., and Grollman A. Analysis and prediction of hydrogen bonding of protein-DNA complexes on parallel processors. *J. Comp. Chem.*, 17(15):1712–1725, 1996.

[127] S. Arnott, R. Chandrasekharan, D.L. Birdsall, A.G.W. Leslie, and R.L. Ratliff. Left-handed DNA helices. *Nature*, 283:743–745, 1980.

[128] R. Lavery. Junctions and bends in nucleic acids: a new theoretical and modelling approach. In W.K Olson, R.H. Sarman, M.H. Sarma, and M. Sundralingham, editors, *Structure and Expression*, volume 3 DNA Bending and Curvature, pages 1–211. Adenine Press, Schenectady, New York, 1988.

[129] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T Demeny, S.H. Hsieh, A.R. Srinivasan, and B. Schneider. The nucleic acid database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysics J.*, 63:751–759, 1992.

[130] W.J. Cook, L.C. Jeffrey, M. Carson, Z. Chen, and C.M. Pickart. Structure of a diubiquitin conjugate and a model for interaction with ubiquitin conjugating enzyme (E2). *J. Biol. Chem.*, 267(23):16467–16471, 1992.

[131] J.E. Anderson, M. Ptashne, and S.C. Harrison. Structure of the repressor-operator complex of bacteriophage 434. *Nature*, 326(6116):846–852, 1987.

[132] J.D. Baleja, R. Marmorstein, S.C. Harrison, and G. Wagner. Solution structure of the DNA-binding domain of Cd2-GAL4 from S. Cerevisiae. *Nature*, 356(6368):450–453, 1992.

[133] E.R. Zuiderweg, R. Kaptein, and K. Wuthrich. Sequence-specific resonance assignments in the 1h nuclear-magnetic-resonance spectrum of the lac repressor DNA-binding domain 1-51 from Escherichia coli by two-dimensional spectroscopy. *European J. Biochem.*, 137(1-2):279–292, 1983.

[134] D.H. Ohlendorf, W.F. Anderson, M. Lewis, C.O. Pabo, and B.W. Matthews. Comparison of the structures of cro and lambda repressor proteins from bacteriophage lambda. *J. Mol. Biol.*, 169(3):757–769, 1983.

[135] J.B. Rafferty, W.S. Somers, I. Saint-Girons, and S.E.V. Phillips. Three-dimensional crystal structures of *escherichia coli* MET repressor with and without corepressor. *Nature*, 341:705–710, 1989.

[136] A. Joachimiak, R.Q. Marmorstein, R.W. Schevitz, W. Mandecki, J.L. Fox, and P.B. Sigler. Crystals of the trp repressor-operator complex suitable for X-ray diffraction analysis. *J. Biol. Chem.*, 262(10):4917–4921, 1987.

[137] A. Joachimiak, R.W. Schevitz, R.L. Kelley, C. Yanofsky, and P.B. Sigler. Functional inferences from crystals of Escherichia coli trp repressor. *J. Biol. Chem.*, 258(20):12641–12643, 1983.

[138] B. Lustig and R.L. Jernigan. Consistencies of individual DNA base-amino acid interactions in structures and sequences. *Nucleic Acids Research*, 23:4707–4711, 1995.

[139] M. Frigo and S. G. Johnson. FFTW: An adaptive software architecture for the FFT. *ICASSP conference proceedings*, 3:1381, 1998.

[140] F. Glaser, D.M. Steinberg, I.A. Vakser, and N. Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43:89–102, 2001.

[141] T.R. Hynes, M. Randal, L.A. Kennedy, C. Eigenbrot, and A.A. Kossiakoff. X-ray crystal structure of the protease inhibitor domain of Alzheimer's amyloid beta-protein precursor. *Biochem.*, 29(43):10018–10022, 1990.

[142] J.J. Perona, C.A. Tsu, M.E. McGrath, C.S. Craik, and R.J. Fletterick. Relocating a negative charge in the binding pocket of trypsin. *J. Mol. Biol.*, 230(3):934–949, 1993.

[143] W. Bode. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. II. The binding of the pancreatic trypsin inhibitor and of isoleucine-valine and of sequentially related peptides to trypsinogen and to p-guanidinobenzoate-trypsinogen. *J. Mol. Biol.*, 127(4):357–374, 1979.

[144] H.J. Hecht, M. Szardenings, J. Collins, and D. Schomburg. Three-dimensional structure of a recombinant variant of human pancreatic secretory trypsin inhibitor (Kazal type). *J. Mol. Biol.*, 225(4):1095–1103, 1992.

[145] J. Walter and R. Huber. Pancreatic trypsin inhibitor. A new crystal form and its analysis. *J. Mol. Biol.*, 167(4):911–917, 1983.

[146] S. Hirono, H. Akagawa, Y. Mitsui, and Y. Iitaka. Crystal structure at 2.6å resolution of the complex of subtilisin bpn' with streptomyces subtilisin inhibitor. *J. Mol. Biol.*, 178(2):389–414, 1984.

[147] M. Huang, R. Syed, E.A. Stura, M.J. Stone, R.S. Stefanko, W. Ruf, T.S. Edgington, and I.A. Wilson. The mechanism of an inhibitory antibody on tf-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF.G9 complex. *J. Mol. Biol.*, 275(5):873–894, 1998.

[148] K. Harlos, D.M. Martin, D.P. O'Brien, E.Y. Jones, D.I. Stuart, I. Polikarpov, A. Miller, E.G. Tuddenham, and C.W. Boys. Crystal structure of the extracellular region of human tissue factor. *Nature*, 370(6491):662–666, 1994.

[149] J. Foote and G. Winter. Antibody framework residues affecting the conformation of the hypervariable loops. *J. Mol. Biol.*, 224(2):487–499, 1992.

[150] M. Ramanadham, L.C. Sieker, and L.H. Jensen. Refinement of triclinic lysozyme: II. The method of stereochemically restrained least squares. *Acta Crystallogr B*, 46 (Pt 1):63–69, 1990.

[151] Y. Li, H. Li, S.J. Smith-Gill, and R.A. Mariuzza. Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63(,). *Biochem.*, 39(21):6296–6309, 2000.

[152] S.E. Mylvaganam, Y. Paterson, K. Kaiser, K. Bowdish, and E.D. Getzoff. Biochemical implications from the variable gene sequences of an anti-cytochrome C antibody and crystallographic characterization of its antigen-binding fragment in free and antigen-complexed forms. *J. Mol. Biol.*, 221(2):455–462, 1991.

[153] G.W. Bushnell, G.V. Louie, and G.D. Brayer. High-resolution three-dimensional structure of horse heart cytochrome C. *J. Mol. Biol.*, 214(2):585–595, 1990.

[154] G.V. Louie and G.D. Brayer. High-resolution refinement of yeast iso-1-cytochrome c and comparisons with other eukaryotic cytochromes c. *J. Mol. Biol.*, 214(2):527–555, 1990.

[155] D.B. Goodin and D.E. McRee. The Asp-His-Fe triad of cytochrome c peroxidase controls the reduction potential, electronic structure, and coupling of the tryptophan free radical to the heme. *Biochem.*, 32(13):3313–3324, 1993.

[156] Y. Mauguen, R.W. Hartley, E.J. Dodson, G.G. Dodson, G. Bricogne, C. Chothia, and A. Jack. Molecular structure of a new family of ribonucleases. *Nature*, 297(5862):162–164, 1982.

[157] G.S. Ratnaparkhi, S. Ramachandran, J.B. Udgaonkar, and R. Varadarajan. Discrepancies between the NMR and X-ray structures of uncomplexed Barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochem.*, 37(19):6958–6966, 1998.

[158] L. Maveyraud, I. Massova, C. Birck, K. Miyashita, J.P. Samama, and S. Mobashery. Crystal structure of 6 alpha-(hydroxymethyl)penicillanate complexed to the TEM-1 beta-lactamase from Escherichia coli: Evidence on the mechanism of action of a novel inhibitor designed by a computer-aided process. *J. Am. Chem. Soc.*, 118(32):7435–7440, 1996.

[159] B. Kobe, Z. Ma, and J. Deisenhofer. Complex between bovine ribonuclease A and porcine ribonuclease inhibitor crystallizes in a similar unit cell as free ribonuclease inhibitor. *J. Mol. Biol.*, 241(2):288–291, 1994.

[160] B. Kobe and J. Deisenhofer. Crystallization and preliminary X-ray analysis of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *J. Mol. Biol.*, 231(1):137–140, 1993.

[161] A. Wlodawer, R. Bott, and L. Sjolin. The refined crystal structure of ribonuclease A at 2.0å resolution. *J. Biol. Chem.*, 257(3):1325–1332, 1982.

[162] M.H. le Du, P. Marchot, P.E. Bougis, and J.C. Fontecilla-Camps. 1.9å resolution structure of fasciculin 1, an anti-acetylcholinesterase toxin from green mamba snake venom. *J. Biol. Chem.*, 267(31):22122–22130, 1992.

[163] C.B. Millard, G. Koellner, A. Ordentlich, A. Shafferman, I. Silman, and J.L. Sussman. Reaction products of acetylcholinesterase and VX reveal a

mobile histidine in the catalytic triad. *J. Am. Chem. Soc.*, 121(42):9883–9884, 1999.

[164] C.H. Lee, K. Saksela, U.A. Mirza, B.T. Chait, and J. Kuriyan. Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. *Cell*, 85(6):931–942, 1996.

[165] M.E. Noble, A. Musacchio, M. Saraste, S.A. Courtneidge, and R.K. Wierenga. Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J*, 12(7):2617–2624, 1993.

[166] C.D. Mol, A.S. Arvai, R.J. Sanderson, G. Slupphaug, B. Kavli, H.E. Krokan, D.W. Mosbaugh, and J.A. Tainer. Crystal structure of human Uracil-DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA. *Cell*, 82(5):701–708, 1995.

[167] C.D. Mol, A.S. Arvai, G. Slupphaug, B. Kavli, I. Alseth, H.E. Krokan, and J.A. Tainer. Crystal structure and mutational analysis of human Uracil-DNA glycosylase: structural basis for specificity and catalysis. *Cell*, 80(6):869–878, 1995.

[168] C.D. Putnam, M.J. Shroyer, A.J. Lundquist, C.D. Mol, A.S. Arvai, D.W. Mosbaugh, and J.A. Tainer. Protein mimicry of DNA from crystal structures of the Uracil-DNA glycosylase inhibitor protein and its complex with Escherichia coli Uracil-DNA glycosylase. *J. Mol. Biol.*, 287(2):331–346, 1999.

[169] K. Scheffzek, M.R. Ahmadian, W. Kabsch, L. Wiesmuller, A. Lautwein, F. Schmitz, and A. Wittinghofer. The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science*, 277(5324):333–338, 1997.

[170] K. Scheffzek, A. Lautwein, A. Scherer, S. Franken, and A. Wittinghofer. Crystallization and preliminary X-ray crystallographic study of the Ras-GTPase-activating domain of human p120GAP. *Proteins*, 27(2):315–318, 1997.

[171] A. Scherer, J. John, R. Linke, R.S. Goody, A. Wittinghofer, E.F. Pai, and K.C. Homes. Crystallization and preliminary X-ray analysis of the human c-H-ras-oncogene product p21 complexed with GTP analogues. *J. Mol. Biol.*, 206(1):257–259, 1989.

[172] M. Kato, T. Shimizu, T. Mizuno, and T. Hakoshima. Structure of the histidine-containing phosphotransfer (HPt) domain of the anaerobic

sensor protein ArcB complexed with the chemotaxis response regulator CheY. *Acta Crystallogr D*, 55(Part 7):1257–1263, 1999.

[173] K. Volz and P. Matsumura. Crystal structure of Escherichia coli CheY refined at 1.7å resolution. *J. Biol. Chem.*, 266(23):15511–15519, 1991.

[174] M. Kato, T. Mizuno, T. Shimizu, and T. Hakoshima. Insights into multistep phosphorelay from the crystal structure of the C-terminal Hpt domain of ArcB. *Cell*, 88(5):717–723, 1997.

[175] B. Kobe and J. Deisenhofer. Mechanism of ribonuclease inhibition by Ribonuclease Inhibitor Protein based on the crystal structure of its complex with Ribonuclease A. *J. Mol. Biol.*, 264(5):1028–1043, 1996.

[176] H. Zhou and Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Prot. Sci.*, 11(11):2714–2726, 2002.

[177] C. Zhang, S. Liu, H. Zhou, and Y. Zhou. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Prot. Sci.*, 13(2):400–411, 2004.

[178] C. J. Camacho, D. W. Gatchell, S. R. Kimura, and S. Vajda. Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins*, 40(3):525–537, 2000.

[179] R. Chen, J. Mintseris, J. Janin, and Z. Weng. A protein protein docking benchmark. *Proteins*, 52(1):88–91, 2003.