Honor Year Project Report

Rigid Body Protein Docking by Fast Fourier Transform

By

Huang Wenfan

Department of Computer Science

School of Computing

National University of Singapore

2004/2005

Honor Year Project Report

Rigid Body Protein Docking by Fast Fourier Transform

By

Huang Wenfan

Department of Computer Science

School of Computing

National University of Singapore

2004/2005

Project No: H040400

Advisor: Associate Professor Leow Wee Kheng

Deliverable:

Report: 1 Volume

Report, Program and Data: 1 Compact Disk

Abstract

Fast Fourier Transform based algorithms are the mainstream of rigid body protein docking procedures. In this report, a study on the original grid based FFT docking approach proposed by Kaltchalski-Katzir is carried out. During study, a two stage docking algorithm based on shape complementarity is developed with the intention to overcome the drawback of the original docking approach. A refinement stage performing coarse-to-fine search in the neighbouring rotational space of specific orientation is employed in this algorithm to improve the docking quality. However, applying two existing protein models designed for grid-based docking to this algorithm yields unsatisfactory results. Therefore, a new double layered protein model is proposed after studying the existing models. This new model is designed to allow close contact between van der Waal surfaces of proteins while it persist a relatively large angular tolerance which enable reasonably fast execution of the algorithm without missing the correct solutions. Experiments showed that this new model performs much better than the existing ones for this algorithm on sixteen bound docking cases. An experiment on six CAPRI unbound docking was also conducted with reasonable result obtained according to the CAPRI evaluation protocol.

Subject descriptor:

J.3 LIFE AND MEDICAL SCIENCES

Keywords:

protein-protein docking; molecular recognition; fast Fourier transform;

Implementation software:

Windows XP; Visual Studio .Net 2003; FFTW; JDK 1.5; JAMA;

Acknowledgement

I would like to thank Prof Leow Wee Kheng for his guidance throughout this project. I also like to thank him for giving me a chance to work as a part time research assistant. This experience is really valuable for me.

I would like to thank Lu Haiyun for her help in prototyping the evaluation program. Without her help, my life would have been more difficult.

I would like to also thank the developers of FFTW. Their powerful and free discrete Fast Fourier Transform C subroutine library enables me to implement the algorithm without extra concern about the efficiency of the FFT computation.

Table of Contents

Chapter 1

Introduction

Biomolecular interactions are the core of all regulatory and metabolic processes that constitute of the process of life. Intermolecular interactions, especially those between proteins, have become a central focus in the post-genomic biology (Mendez, Leplae, Maria and Wodak, 2003). In last few decades, a large number of possible protein interactions have been uncovered from genetic, biochemical and proteomics studies forming millions of putative protein-protein complexes. However, only a very small fraction of these complexes are available for analysis. In the meantime, although the development of structural biology has considerably accelerated the experimental determination of three dimensional protein structures, experimental determination of protein-protein complex structure remains difficult (Chen and Weng, 2002). For these two reasons, algorithms for computational prediction of protein-protein interactions are becoming increasingly important in recent years. These algorithms could not only serve as valuable tools for the industry, such as drug companies, but also potential utilities to give insight into the process of protein-protein reaction for scientific research.

1.1. Problem definition and background

Prediction of protein-protein interactions in a computational way is commonly addressed as the Protein docking problem. Docking is the term used for computational schemes that attempt to find the 'best' matching between two molecules. It essentially simulates the interaction of the protein surface. Therefore, docking usually involves the geometry of the molecular surfaces, as well as chemical and energy considerations. Protein docking problem can be formally defined as follows: Given the three-dimensional atomic coordinates of two protein molecules, find their 'correct' bound association (the relative orientation and position after interaction) between such two proteins. In the most general form, no additional data are provided (Halperin, Ma, Wolfson, and Nussinov, 2002).

Docking involves two separate proteins. By convention, the larger protein involved is referred to as the receptor, while the smaller one is known as the ligand. Depending on how the coordinates of a receptor and respective ligand are obtained, the docking problem can be divided into two categories, bounded docking and unbounded docking. Bounded docking is a simpler version of the problem. Both the receptor and the ligand are extracted from the structure of one protein complex, typically the product of interaction between the receptor and the ligand. And the goal is to reconstruct the complex. On the contrary, the unbound docking is designed for 'real' situations. The unbound structures of receptor and ligand are used as inputs for docking, and the goal is to predict how receptor and ligand could be bound to each other after interaction. As defined by Halperin et al (2002), an unbounded structure may be a native structure and a pseudo-native structure. A native structure is the structure when a protein is free in the solution, in its uncomplexed state. A pseudo-native structure is the structure of a protein complexed to a molecule which is different from the one involved in docking.

Investigation of the three-dimensional structures of most protein complexes deposited in the fast growing Protein Data Bank reveals a close geometric matching between the respective surfaces of the receptor and the ligand (Connolly, 1986). Physically, the van der Waals (VDM) surfaces of atoms cannot overlap in space and protein-protein interfaces between ligands and respective receptors generally do not contain large empty or water-filled holes (Hubbard and Argos, 1994) (See Figure 1.1, 1.2). Indeed, geometric complementarity between proteins plays a very important role in the process of docking. As Connolly stated, for docking, "Geometry is not everything, but it is the most fundamental thing." (Connolly, 1983). Although biologist may argue that physical and chemical properties play a more prominent role, shape complementarity between proteins surfaces was quickly used by people as a foundation for docking algorithms, which may also include chemical consideration as helpful complements.

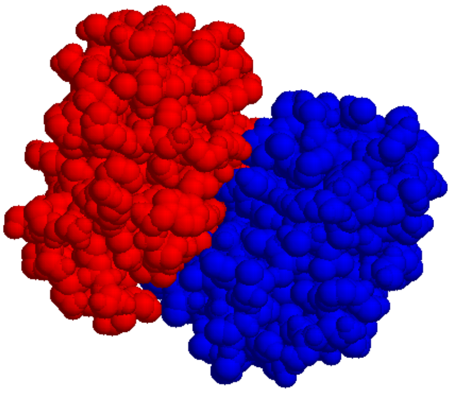


Figure 1.1: Shape Complementarity at alpha (red) -beta(blue) subunits interface of horse hemoglobin

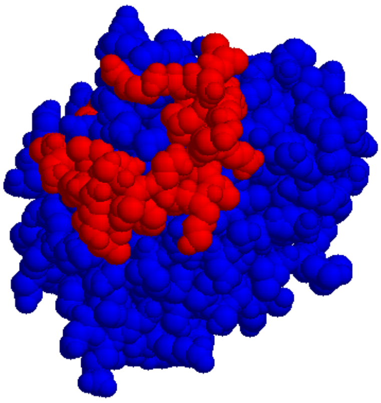


Figure 1.2: Shape Complementarity at alpha thrombin (blue) and hirulog 3 (red) interface of Hydrolase.

Docking is a difficult problem to address computationally especially for unbound cases. That is because unbound proteins could undergo conformational changes upon interaction. Such conformational changes will introduce additional difficulties to the problem. Algorithms that allow conformation changes by considering proteins as flexible shapes have already been proposed (Totrov and Abagyan, 1994; Halperin et al 2002). However, such algorithms are not computational affordable for large proteins at current stage, therefore they are not widely used. The more reasonable way is to consider proteins as rigid bodies. The conformation change is tolerated by allowing certain degree of penetration between input proteins. This so-called 'soft docking' approach reduces drastically the complexity of docking problem. Therefore, it is adopted for majority of the algorithms.

1.2. Objective and Contribution

Since 2002, nearly 20 docking teams, which represent the mass majority of docking community, have taken part in the CAPRI[1] experiment which aims at assessing the performance of docking procedures by blind trials. Among these docking procedures, a sizable fraction of them uses a cubic grid representation of the rigid body protein surface and Fast Fourier Transform (FFT) search algorithms, following the earlier work by Katchalsi-Katzir and his colleague (1992) (See Appendix A; Mendez et al 2003). It can also

[1] Critical Assessment of PRedicted Interactions. Hosted by European Bioinformatics Institute.

be noticed that this family of algorithms performed quite well in the first two rounds of CAPRI (See Appendix B). These facts triggered the original motivation of this project: to provide a foundation for further research on protein docking problem by studying Katchalsi-Katzir's grid-based FFT docking approach.

Katchalsi-Katzir's FFT docking approach is based on the shape complementarity which is measured using Fourier correlation. It's a 'soft docking' approach: the ligand and the receptor are considered as rigid bodies, and the conformation changes are accounted by allowing certain degree of inter protein penetration. This method is designed for docking problem in the most general form: the only input information is atomic coordinates of proteins.

During algorithm study, it has been found that the original Katchalsi-katzir's two-stage docking algorithm could not satisfy the current standard for docking due to the low accuracy of the results it produced. Therefore, a new two-stage algorithm based on Katchalsi-Katzir's FFT docking approach was proposed and implemented in C++. Same as Katchalsi-katzir's docking algorithm, this new algorithm is also based on shape complementarity only.

Through several experiments on the algorithm, a new double layered protein model was also proposed for this algorithm. Experiments on sixteen bound cases showed that this new model outperformed the model used by Katchalsi-Katzir et al (1992) and the model suggested by Gabb, Jackson, and Sternberg (1997) with much more accurate results produced. To further assess the new algorithm and the new model, an experiment with six CAPRI unbound testcases were also conducted with reasonable results obtained.

The rest of the thesis is structured as follows: Chapter 2 gives a literature review of rigid body docking algorithms. Chapter 3 presents the evaluation system used for performance study of the new algorithm and different protein models. Chapter 4 introduces the new algorithm and experimental results on existing protein models. Chapter 5 describes the new protein model, how it was derived and its performance on the docking problem. Chapter 6 concludes the whole project.

Chapter 2

Related Works

The first attempt toward the docking problem was made by Crick in early 1950s (Crick, 1953). However, due to the limitation of processing power of computer and lack of experimentally determined protein structures, the computational study of docking didn’t begin to flourish until middle of 1980s. The first practical method for docking was proposed by Connolly M.L. in 1983. His docking algorithm matches surface knobs with the surface depressions by describing protein surface using mathematical function. Docking is one of the most creative fields in computational biology. It is hard to enumerate all the algorithms that have been proposed. In this chapter, the focus will be mainly kept on shape complentarity-based docking procedures, while algorithms based on energy minimization will not be illustrated since the project scope is limited to geometry-based algorithms.

Docking consists of three key aspects: 1) Conformation space search, i.e. how conformation changes between bound and unbound structures of a protein are accounted 2) Representation of the proteins, i.e. how to represent the protein surface since docking simulates the interaction between protein surfaces; 3) Searching Algorithms and scoring schemes, i.e. how to find and rank the candidate solutions. Obviously, these aspects are closely related. The choice of representation will affect how search will be conducted. In the following subsections, an overview of Conformation space search and a justification for rigid body assumption will be given, followed a brief review on protein representations, and this chapter will be closed by a discussion on searching algorithms and scoring functions.

2.1. Conformation Space Search: Rigid vs. Flexible.

Docking is computationally difficult because there are many ways of putting two proteins in a complementary manner (six degrees of freedom for rigid transformation). This problem could become even more complicated when considering conformation changes for unbound docking. This additional difficulty of unbound docking derives from the conformational change that take place between the bound and unbound protein structures.

The conformation changes mainly result from protein ‘disorder’. A free protein can exist in a range of conformational substates, with low-energy barrier separating them. However, experimentally determined 3D structure for a protein is available, in most cases, for only one conformational substate. And the protein - protein interaction will stabilize both proteins, and force them into equilibrium, which will alter the structures of both participant of the interaction. Therefore, usually the experimentally detected structure of a protein in bound state will be different from the structure detected in unbound state. The complementarity, either shape or chemical, between structures of a bound protein pair, may not be easily observed in their unbound states.

According to how conformational changes are handled, docking algorithms are classified into three classes:

- *Flexible docking:* both receptor and ligand are considered as flexible. However, the extent of flexibility is either limited or simplified. Such flexibility is modeled by simulation methods (Halperin et al, 2002).
- *Rigid body docking:* a simplified model that regards two proteins as rigid bodies. The conformation change is tolerated by allowing certain degree of penetration between proteins. This assumption will limit the problem to a six-dimensional (three for translation and three for rotation) search space.
- *Semi-flexible docking:* Only one protein involved in docking is considered as flexible. Usually the smaller protein involved in docking, the ligand, is considered as flexible. The underlying reason for this is that small proteins are likely to have more conformational variations, and further more, compared to large proteins, simulating conformational changes of small proteins are computationally affordable.

Simulation of structural flexibility is a computational expensive process even for only one protein involved in docking. Due to the high computational complexity, flexible docking algorithms are not applicable to practical protein docking at present, while the semi-flexible docking is for docking that involves small molecules. On the contrary, the rigid body algorithms, which limit the search space to six dimensions by rigid body assumption, are extensively used for large protein docking.

A lot of literatures have been published on Semi-flexible docking. Generally speaking, Semi-flexible docking algorithms are designed for docking between a small molecule and a big molecule, such as protein drug docking. A general approach for semi-flexible docking is a two-stage process. The first stage is to produce sets of possible ligand conformations from conformational simulation. Several general algorithms have been applied for simulating such conformational flexibilities, for example: Monte Carlo (Totrov and Abagyan, 1994), simulated annealing (Goodsell and Olson, 1990) and genetic algorithm (Jones, Willet, Glen, Leach and Taylor, 1997). The second stage is to dock these generated ligand conformations with the receptor by certain rigid body docking algorithm. Rigid body docking actually serves as the foundation for semi-flexible docking.

For docking between large proteins, the number of degrees of freedom may be tremendous with conformational changes taken into consideration. Generally, rigid body docking algorithms will be used for such problem in order to reduce computational complexity. Although the ability to handle conformational changes is limited, those algorithms are remarkably successful for large protein docking (Halperin et al, 2002). That is because for large proteins, structural flexibility is mainly restricted to surface side chains (Betts and Sternberg, 1999), which could be tolerated if a rigid body docking algorithm is 'soft' enough.

As a conclusion for above paragraphs, rigid body assumption is reasonable for protein docking, especially for large protein docking. In addition, rigid body docking can also serve as the foundation for some flexible algorithms. Therefore, this assumption is widely adopted.

2.2. Representations of Proteins as Rigid bodies

The inputs of the docking problem are two sets of atomic coordinates of proteins. Such basic representation is usually not used for the docking algorithms. More often, the protein or the protein surface only will be reconstructed from the atomic coordinates and represented by certain mathematical models for ease of searching and ranking of possible solutions. There are two major branches of representations, namely, by geometric features and by grids.

- *Geometric features representation:* A bulk of algorithms chose to represent a protein by its geometric features of the protein surface. Connolly laid the foundation for this class of algorithms by introducing protein surface analysis. He proposed a protein surface model: the Connolly surface, which is also known as molecular surface. Based on the Connolly surface analysis, a surface is described by sparse critical points of the mathematical function describing the surface (Lin, Nussibov, Fischer and Wolfson, 1995). Those critical points could be cavities, local knob or holes on the actual surface. In the later stage, surface normals of those critical points are also included for surface representation (Norel, Lin, Wolfson and Nussinov, 1995). Besides Connolly surface, critical points can also be sampled on other kinds of rigid body protein surface models, such as solvent-accessible surface and molecular skin. For this class of representation, the sparseness of sampled points is critical for effective and accurate search for candidate rigid transformations associating the surfaces of two proteins in a complementary manner.
- *Grids representation:* Besides representing the protein by geometry features, another mainstream way is to represent the protein by grids. This approach was first applied to docking by Katchalsi-Katzir et al (1992). In this representation, the structure of a protein was discretized into three-dimensional Cartesian grid, and different numeric values are assigned to nodes of the grids. Similarly, grid representation could also be applied to different rigid body protein models. The original protein model by Katchalsi-Katzir is the most widely used one.

Comparing to the geometry feature representation of the protein, the grid-based one has several advantages. In addition to surface shapes, it can be easily applied to other protein surface properties such as electrostatic and hydrophobicity. Another advantage for this representation is that it can choose to represent the protein with either high resolution or low resolution depending on whether the accuracy is favored or the speed is favored. In conclusion, grid-based representation is more flexible than the geometry feature representation. Therefore, it has been widely adopted since it was proposed.

2.3. Searching algorithms and scoring schemes for Rigid body docking

For rigid body docking, a candidate solution is a rigid transformation which associates two proteins in a complementary manner. As stated in the beginning of this chapter, search for such transformations are closely related to protein representation. In the following section, two fundamental algorithms for rigid body docking will be illustrated.

Geometry feature representation represents a protein by critical points on the surface and the associated surface normals. To compute a rigid transformation to superimpose a receptor onto a ligand, three non-collinear points from each protein are needed. However, there may not always be three independent matching critical points pairs. In order to cope with this problem and reduce the complexity, Norel introduced geometry hashing-based docking algorithm (Norel et al, 1995; Norel, Petrey, Wolfson and Nussinov, 1999). Their algorithm picks two critical points from protein surface. For each pair of points from each protein, a 'signature' including certain geometry information about the two points and normals is computed. The computational complexity is reduced by breaking up the search into preprocess step and recognition step. In preprocess step, a look-up table with each entry consisting of a pair of points from ligand, a signature, and coordinate of a critical point using the two points as reference frame etc is built up for all possible pair of critical points on each protein. In recognition step, the best rigid transformation between the two pairs of points is computed by exploiting the pre-computed look up table only if the signatures are compatible according to some criteria. Those locally determined transformations will then undergo post-processing to remove spatially prohibited transformations, such as deep penetration between proteins.

Unlike the geometry hashing-based algorithm which explore only part of the solution space, algorithms using grid model of protein usually scan the entire solution space by systematically rotating and translating one protein about another. Matching of surfaces is accomplished by calculating correlation functions, which favors close contact and automatically penalizes surface overlap. The correlation calculation and successive translational increment can be performed efficiently using Fast Fourier Transform (FFT) algorithm. No post-processing step is needed for removing transformations causing deep penetrations since it is already integrated into the search stage.

As stated in the previous subsection, grid-based representation has several advantages over geometrical feature representation. These advantages make FFT based approaches more attractive than the geometry hashing based algorithms. First, FFT based algorithms could also correlate other protein surface properties such as electrostatic and hydrophobicity. And those properties can be integrated together when searching for the solutions instead of employing additional post-processing steps. In addition, FFT based algorithms can either perform fast low resolution scan for rough docking, or high resolution scan for accurate docking, and even a combination of low and high resolution docking to compromise between quality and speed.

The major disadvantage for FFT based algorithms is that it is quite slow compared to geometry hashing based ones due to its full solution space searching. However, restricting search space may bear the danger of missing correct solutions. In addition, at the current stage of development of docking algorithm, the docking community is much more concerned about the docking quality rather than speed of execution.

A docking algorithm may produce a large number of solutions during or after searching. To discriminate between ‘correct’ solutions and false positives, a reliable and fast scoring function is required. One of the most commonly used scoring functions is shape complementarity, which awards surface contact, penalizes overlap and rejects serious overlaps. However, geometry along is not powerful enough to filter out the undesired solutions. Some false positive solutions for some docking cases may appear to have a better shape complementarity than correct solutions. In recent years, biochemical properties such as electrostatic, hydrophobicity, and hydrogen bond have been extensively applied as scoring criteria in the development of docking algorithm. However, since geometry complementary calculation is highly efficient, they usually serve as a primary filter for the solutions (Halperin et al, 2002).

Chapter 3

Evaluation System

To evaluate the performance of an algorithm, a good evaluation protocol has to be developed and a pool of varied testcases must be collected. This chapter consists of two sections. The first section describes the parameters and methods used for evaluating the docking results whereas the second section lists all bound and unbound cases employed for development and verification of the new algorithm proposed in this report.

3.1. Evaluation protocol

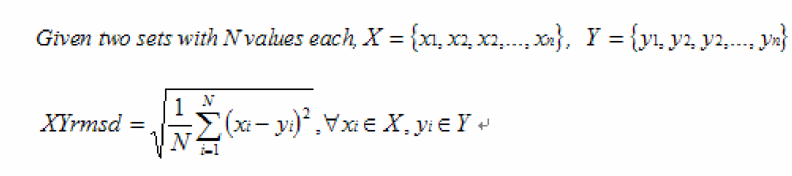
To assess the quality of a docked complex generated by a docking algorithm from a given pair of proteins, a natural way is to compare it with the known 'correct' structure. This is exactly what has been used for this project.

For bound docking, which aims at reconstruction, the 'correct' structure is obviously the structure of the protein complex from which both input receptor and input ligand are extracted. For unbound docking, which aims at prediction, the 'correct' structure is the experimentally determined structure of the complex formed by the receptor and the ligand in real biochemical interaction.

There are three parameters used for measuring the distance between a docked complex and the correct structure: namely, ligand RMSD (root mean standard deviation), interface RMSD and fraction of native residue-residue contact. These three parameters are adopted by CAPRI with the purpose of providing a reliable basis for performance evaluation and analysis.

3.1.1. Interface RMSD and Ligand RMSD.

These two quantities are used to evaluate the overall geometric fit between the 3D structures of the docked complex and the correct one. RMSD is a term measuring the distance between two sets of values, as formally defined:

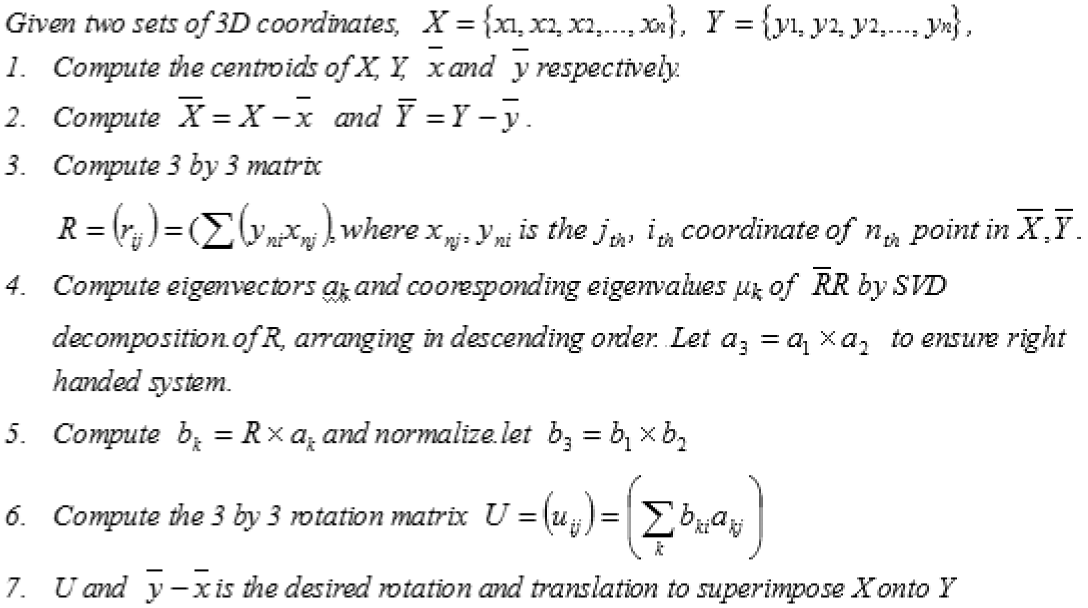


Given two sets with N values each, $X=\{x_1, x_2, x_2,\ldots, x_n\}$, $Y=\{y_1, y_2, y_2,\ldots, y_n\}$

$$XYrmsd=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2}, \forall x_i \in X, y_i \in Y$$

A lower RMSD indicates a better fit between the structures.

Because the docked complex and the correct complex may be at different orientations, before actual RMSD calculation, the receptor of both the docked complex and the correct one have to be superimposed/aligned. Based on Kabsch (1978)'s work. the procedure for superimposition has been developed as shown below:



Given two sets of 3D coordinates, $X=\{x_1, x_2, x_2,\ldots, x_n\}$, $Y=\{y_1, y_2, y_2,\ldots, y_n\}$,

1. *Compute the centroids of X, Y,* $\bar{x}$ *and* $\bar{y}$ *respectively.*
2. *Compute* $\overline{X}=X-\bar{x}$ *and* $\overline{Y}=Y-\bar{y}$.
3. *Compute 3 by 3 matrix*

   $R=(r_{ij})=(\sum(y_{ni}x_{nj}))$, *where* x_{nj}, y_{ni} *is the* j_{th}, i_{th} *coordinate of* n_{th} *point in* $\overline{X}, \overline{Y}$.
4. *Compute eigenvectors* a_k *and cooresponding eigenvalues* μ_k *of* $\overline{R}R$ *by SVD decomposition of R, arranging in descending order. Let* $a_3=a_1\times a_2$ *to ensure right handed system.*
5. *Compute* $b_k=R\times a_k$ *and normalize. let* $b_3=b_1\times b_2$
6. *Compute the 3 by 3 rotation matrix* $U=(u_{ij})=\left(\sum_k b_{ki}a_{kj}\right)$
7. *U and* $\bar{y}-\bar{x}$ *is the desired rotation and translation to superimpose X onto Y*

After superimposition of receptors, L_RMS, the RMSD between ligands in the docked complex and the correct complex, can be computed. Both the superimposition and L_rms are computed on coordinates of backbone atoms (C, C_{α}, N, O), for the reason that the structure of the backbone of a protein is usually stable upon protein conformational change.

L_RMS is a global measure. Therefore, it may not always portrait the real fit at protein-protein interface, especially when ligand is large. Hence, another local parameter, interface RMSD (I_RMS) is used for measuring the fit between the docked complex and the correct complex in the interface region. The interface region is defined in the correct complex.

It consists of interface residues[1]. A residue is said on the interface if any of its atoms is within 10 Å of an atom on the other protein in the correct complex. Once more, only the backbone atoms of those interface residues and their equivalents in the docked result will then be used to compute the I_RMS after superimposition.

3.1.2. Fraction of native residue-residue contact Fnat.

A pair of residues on different sides of protein-protein interface is considered to be in contact if any of their atoms were within 5Å (angstrom). Fnat is defined as the number of native (correct) residue-residue contacts in the docked complex divides by the number of contacts in the known correct complex. The number of residue-residue contacts is critical for protein interaction to take place. Therefore, Fnat is also used for evaluating docked complexes.

3.1.3. Evaluating a docking algorithm using these parameters

In most cases, a docking algorithm will produce several complexes, which are usually ranked according to their scores, for one docking case. Depending on whether the docking case is bound or unbound, the performance of the algorithm on this specific case is evaluated in different ways:

- For a bound docking case, only L_RMS will be computed, since L_RMS alone is good enough for measuring the quality of bound docking. The lowest L_RMS among all produced complexes and the rank of the complex with the lowest L_RMS will be used to evaluate how well the algorithm performs on this case. The lowest L_RMS is an indication for how good the best docked complex is, whereas the rank shows the ability of the algorithms to distinguish correctly docked complexes from false positives. A case is identified as Fail, if the lowest L_RMS is larger than 18 Å.
- For an unbound case, the evaluation protocol strictly follows CAPRI's procedure (Mendez et al 2003). All three parameters will be computed. The performance of the algorithm for s case is evaluated by the quality of the best docked complex. The quality is in terms of Fnat, L_RMS and I_RMS as defined in Table 3.1. A docked complex is considered as the best, if it has the lowest I_RMS.

[1] Residue is a term referring to those amino-acids which made up proteins.

Quality	F_{nat}	L_RMS	or I_RMS
High	≥ 0.5	≤ 1.0	or ≤ 1.0
Medium	≥ 0.3	≤ 5.0	or ≤ 5.0
Acceptable	≥ 0.1	≤ 10.0	or ≤ 10.0
Incorrect	< 0.1	> 10.0	or > 10.0

Table 3.1. Quality of a docked complex is determined according to Fnat AND (L_RMS OR I_RMS). These criterias are adopted in CAPRI (Norel et al 2003 Table II)

3.2. Docking cases.

3.2.1 Bound cases

A total of sixteen bound cases, consisting of proteins with a large variety of number of atoms, are used in this project (See Table 3.2). Structure files of all complexes are taken from Protein Data Bank. After ligands and receptors are extracted from the complexes, their orientations are randomized before docking. All bound cases were from Mendez et al (2002)

Complex name	Receptor name	No. Of Atom	Ligand name	No. Of Atoms
1CHO	Alpha-chymotrypsin Chain	1048	Alpha-chymotrypsin Chain 148-245	702
1ABI	Hydrolase alpha thrombin	2039	Hydrolase Chain L	265
1ACB	Hydrolase	1769	Eglin C (I)	522
1CSE	Subtilisin (E)	1920	Subtilisin Inhibitor (I)	522
1TGS	Trypsinogen (Z)	1646	Panecreatic Secreatic Inhibitor	454
2KAI	Kallikrein a	1799	Bovine Panecreatic trypsin Inhibitor	438
2MHB	Hemoglobin α	1069	Hemoglobin β	1134
2PTC	Beta-trypsin	1629	Panecreatic Secreatic Inhibitor	454
3HFM	IG * G1Fab fragment	3295	Lysozyme	1001
4HVB	HIV-1 protease Chain A	746	HIV-1 protease Chain B	746
4SGB	Serine proteinase	1310	Potato Inhibitor	300
4TPI	Trypsinogen (Z)	1629	Panecreatic Secreatic Inhibitor	456
9LDT	Lactate ddehydrogenase	2568	Lactate ddehydrogenase Chain B	2568
1FDL	IG * G1Fab fragment	3308	2-lysozyme	1001
2SIC	Subtlisin	1938	Subtilisin Inhibitor (I)	764

Table 3.2: Bound testcases used for performance analysis.

3.2.2 Unbound cases

Six unbound cases given by CAPRI team (Norel et al 2003) are used to evaluate the performance of the new algorithm (See Chapter 5). The table below gives a brief description of these unbound cases.

Complex name	Receptor name	No. Of Atom	Ligand name	No. Of Atoms
CAPRI02	bovine rotavirus VP6	9486	Fab	3237
CAPRI03	flu hemagglutinin	11679	Fab HC63	6677
CAPRI04	alpha-amylase	3898	Camelide antibody VH domain 1	882
CAPRI05	alpha-amylase	3908	Camelide antibody VH domain 2	905
CAPRI06	alpha-amylase	3908	Camelide antibody VH domain 3	899
CAPRI07	T cell receptor	1757	Toxin	1785

Table 3.3: Unbound testcases. Detail description of those cases can be found on CAPRI website[1].

[1] http://capri.ebi.ac.uk/capri.html

Chapter 4

The algorithm and experimental results

As a pioneer, Katchalsi-Katzir's algorithm inspired many researchers in docking filed. However, under the current standard, this algorithm is no longer applicable for docking due to the low accuracy of results it produced. A new algorithm based on Katchalsi-Katzir's work was developed for relatively high accuracy docking. Same as Katchalsi-Katzir's algorithm, the docking criteria used for this new algorithm is also shape complementarity only. However, with existing protein models, the performance of the algorithm is not as good as what is expected. In this chapter, the algorithm will be introduced in first section, and the experimental results with two protein model will be presented in second section.

4.1. The algorithm

4.1.1 The grid based FFT docking approach

Rigid body docking is essentially finding the best rigid transformations to associate two proteins. A rigid transformation consists of two components: translation and rotation. In the following paragraph, a detail explanation will be given on how these two components will be scanned.

<u>4.1.1.1 Measuring shape complementarity and scanning the translational space by FFT</u>

The grid-based FFT docking starts with representing proteins with grids. Both protein molecules are considered as rigid body and projected onto two three-dimensional grids of **N** × **N** × **N** nodes each by aligning the centroid of the protein with the center of the grid. Every node in the grids is assigned to a value according to the following function:

$$f_{A_{l,m,n}} = \begin{cases} 1 : \textit{on the surface of the protein molecule} \\ \rho : \textit{inside the protein molecule. a negative value} \\ 0 : \textit{outside the protein molecule} \end{cases} \qquad \textbf{Func. 1}$$

$$f_{B_{l,m,n}} = \begin{cases} 1 : \textit{on the surface of the protein molecule} \\ \delta : \textit{inside the protein molecule. a positive value} \\ 0 : \textit{outside the protein molecule} \end{cases} \qquad \textbf{Func. 2}$$

Where **A**, **B** are the grid representation for the receptor and the ligand respectively and l, m, n are the grid indices. A node is considered inside the protein molecule if there is at least one heavy atom (Carbon, Nitrogen or Oxygen) within ***r*** Å from it. The surface is defined as a boundary layer of finite width ***t*** Å between the inside and the outside of the protein molecule. A node is said to be the surface if the distance to the nearest heavy atom is between ***r*** and ***t*** + ***r*** (See Figure 4.1).

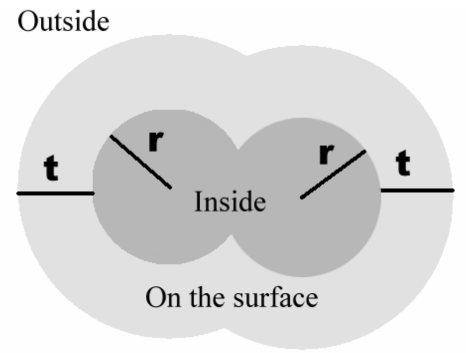


Figure 4.1: An illustration of 'inside', 'outside', 'on the surface', ***r*** and ***t***. There are two atoms.

The matching of surface is accomplished by calculating the correlation functions defined as:

$$f_{C_{\alpha,\beta,\gamma}} = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\left(f_{A_{l,m,n}} \cdot f_{B_{l+\alpha,m+\beta,n+\gamma}} \right) \quad \textbf{Eq. 1}$$

Where α, β, γ are the number of grid steps by which **B** is shifted with respect to protein **A** in each dimension. According to a translation vector {α, β, γ}, if there is no contact between the two proteins, the correlation value should be zero. If there is a good geometry match, the correlation value should be positive. If two proteins deeply penetrated each other, negative correlation values should be obtained. (See Figure 4.2) To formulate a clear distinction between the above three situations, a relatively large negative value should be assigned to ***ρ*** in f_A while a relatively small positive value should be assigned to ***δ*** in f_B. Therefore, if there is a penetration between two proteins after relative shifting, ***ρ*** times ***δ*** or 1 (the value assigned to grid in the surface layer) will contribute negatively to the overall correlation score. On the contrary, if there is a overlapping of surface, positive value will be contributed. In a conclusion, a correlation value is the score for surface contacts after being penalized by penetration. A positive correlation will be obtained if the contribution from contacts overweighs contribution from penetration. This scoring scheme is a ‘soft’ scoring scheme. Even for a translation vector such that there is certain degree of penetration, a large positive correlation score can still be obtained as long as there is a good surface overlapping.

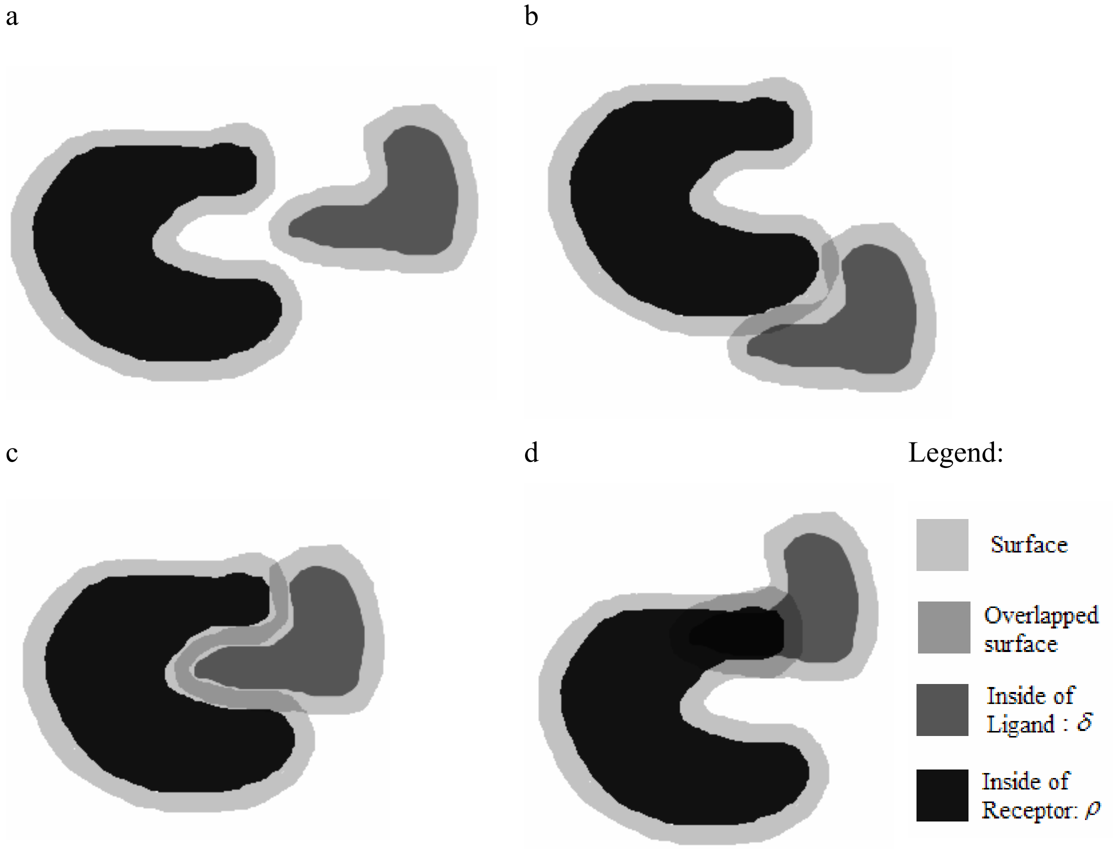


Figure 4.2: Different relative position of receptor and ligand, illustrated in 2D. a) No surface contact. b) Limited contact. c) Good Geometry match. d) Deep penetration.

If we plot a graph of f_C using correlation scores versus {α, β, γ} in the entire translation space (N × N × N), a good geometry match will be represented by a high peak in the graph (See Figure 4.3) while a poor match will be a low peak. Thus, the translation for the best shape match can be readily determined by the coordinates of the highest correlation peak.

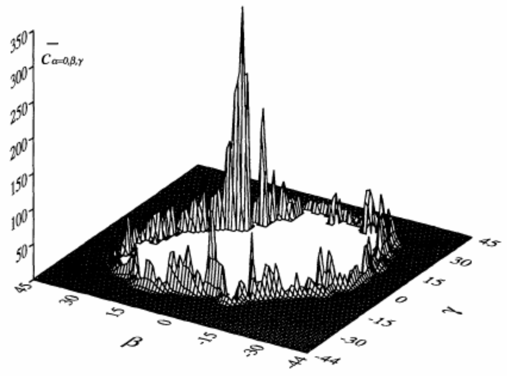


Figure 4.3: Cross Section at α = 0 through function $f_{C\alpha,\beta,\gamma}$ for a docking case. The height of the graph represents the correlation value at each shift vector {0, β, γ}. Negative values are omitted, and center area is left empty. Graph is taken from Katchalski-katzir et al (1992).

Direct calculation of $\boldsymbol{f_c}$ for each {α, β, γ} shift involves N^3 multiplications and additions, which means an $O(N^6)$ complexity for all N^3 possible shifts. However, since both $\boldsymbol{f_A}$ and $\boldsymbol{f_B}$ are discrete functions, such calculation can be much more rapidly done using Discrete Fast Fourier Transform. The discrete Fourier transformation (DFT) of a discrete function $f_{l,m,n}$ is defined as:

$$F_{o,p,q} = \sum^{N}\sum^{N}\sum^{N} \exp[-2\pi i(ol + pm + qn)/N] \times f_{l,m,n}\text{, } \textit{where o, p, q= \{1, 2,..., N\}}$$

And the inverse discrete Fourier transformation (IFT) is defined as:

$$f_{l,m,n} = \frac{1}{N^3}\sum_{o=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} \exp[-2\pi i(ol + pm + qn)/N] \times F_{o,p,q}$$

Apply DFT to both side of Equation. 1 yields:

$$F_C = F_A{}^* \times F_B\text{, } \textit{where } F_A{}^* = \text{Complex conjugation of DFT}(f_A)\text{,}$$
$$F_B = \text{DFT}(f_B)\text{, } F_C = \text{DFT}(f_C)\text{.} \qquad \textbf{Eq. 2}$$

The above equation reveals that f_C for all possible shifting vector can be obtained by an IFT operation on $\boldsymbol{F_C}$ which is computed by simplify multiplying complex functions $\boldsymbol{F_A}$* and $\boldsymbol{F_B}$ together. By performing fast Fourier transform algorithm (Eliot and Rao, 1982) for those DFT and IFT, calculation for correlation scores for all possible shifts requires $O(N^3\ln(N))$ steps, which is significantly faster.

4.1.1.2 Scanning the rotational space.

The correlation calculation and scan for the highest correlation peak must be performed for all relative orientations of two input proteins in order to find the transformations that produce good geometry matches. An orientation is defined by three Eulerian angles: ϕ (0° ~ 360°), θ (0° ~ 180°) and ψ (0° ~ 360°) (See Goldstein, (1980) for definition of three angles). It can be represented below by a rotation matrix R.

$$R_{\phi,\theta,\psi} = \begin{pmatrix} -\sin\phi\cos\theta\sin\psi + \cos\phi\cos\psi & \cos\phi\cos\theta\sin\psi + \sin\phi\cos\psi & \sin\theta\sin\psi \\ -\sin\phi\cos\theta\cos\psi - \cos\phi\sin\psi & \cos\phi\cos\theta\cos\psi - \sin\phi\sin\psi & \sin\theta\cos\psi \\ \sin\phi\sin\theta & -\cos\phi\sin\theta & \cos\theta \end{pmatrix}$$

In practice, the receptor is fixed, while the ligand is rotated with respect to its centroid according to the three Eulerian angles which are varied at fixed angular step **Δ**. However, such sampling of the rotational space is biased. The authors of Katchalsi-katzir's algorithm didn't handle this drawback. Gabb et al (1997) suggested computing the pair-wise distance between two orientations, which is measured by the following formula:

$$dist((\phi_1,\theta_1,\psi_1),(\phi_2,\theta_2,\psi_2)) = \arccos\frac{tr\left(R_{\phi_1,\theta_1,\psi_1} \times R^T_{\phi_2,\theta_2,\psi_2}\right)-1}{2}, \textit{ where tr() is matrix trace}$$

In geometry sense, this distance is the magnitude of the angle by which one orientation is rotated to another with respect to certain axis (Lattman, 1971). Therefore, orientations within 1° distance from any already scanned orientation are defined as degenerate and will not be scanned again. For example, when **Δ** = 15°, preventing scanning degenerate orientations reduces total number of orientations from 360 × 180 × 360/$\mathbf{\Delta}^3$ = 6912 to 6360. For scanned orientation, the highest peaks found will be saved.

After the entire rotational space has been traversed, all the peaks saved will be sorted according their correlation score. Each of these peaks indicates a geometric match and represents a potential docked complex. The higher a peak is ranked, the more likely it could represent the correct complex. The relative transformation between two input proteins to produce such complex can be easily derived from the coordinates of the peak and the three Eulerian angles at which the peak was found.

It is also noteworthy that orientation sampling is discrete. It's not reasonable to assume that the correct orientation will be sampled during rotational space scanning. However, by assigning appropriate values to these parameters, such as **Δ,** ***t*** etc., an orientation that slightly deviates from the correct orientation would still produce a distinct correlation peak. The maximal deviation from the correct orientation that would still result a correlation peak is defined as angular tolerance. This quantity is crucial for the FFT docking approach. Docking with a **Δ** larger than the angular tolerance will result in missing correct orientations for some docking cases.

4.1.2. Katchalsi-katzir’s docking algorithm

Katchalsi-katzir and his colleagues did not stop when they had developed the above docking approach. Instead, they proposed a two-stage docking algorithm which compromises between computation load and docking quality. The first stage is a coarse global search with a larger **η** (grid step size) and a smaller **N** (the number of nodes in each dimension of grid). The second stage is a discrimination stage using a finer grid with smaller **η** and larger **N**. The global search stage will scan the entire rotational space and translational space for the highest peak of each sampled orientation. These peaks will be sorted and the top **k** peaks will be passed to the discrimination stage. In the discrimination stage, the surface correlation scores for orientations that yields the **k** peaks will be recalculated using the smaller **η** and the bigger **N** and the highest peaks will be scanned one more time. During this stage, correct correlation will be enhanced while spurious peaks will be suppressed. After this stage, recalculated peaks will be sorted according to their correlation scores, and then used to generate potential docked complexes.

4.1.3. A new variation based on the FFT docking approach

Katchalsi-katzir’s algorithm can only produce roughly correct docked results because of its discrete sampling of the rotational space during the global search stage. A correct orientation could not be accurately detected unless it is fortunately sampled during global search. Hence, a smaller **Δ** is always desired to produce high quality docked complexes. However, the angular step **Δ** in global search stage can not be too small otherwise the computation load will be too heavy for practical use. This contradiction between speed and quality was not handled in their algorithm. If we want to get some accurate results after running their algorithm, we have to assign a small value to **Δ**, but such small value may result in days of computation. If we want the program to finish in reasonable time, the results produced will not be accurate enough under the current standard of docking.

Therefore, a variation aiming to resolve this contradiction has been proposed and implemented in this project (See Figure 4.4).

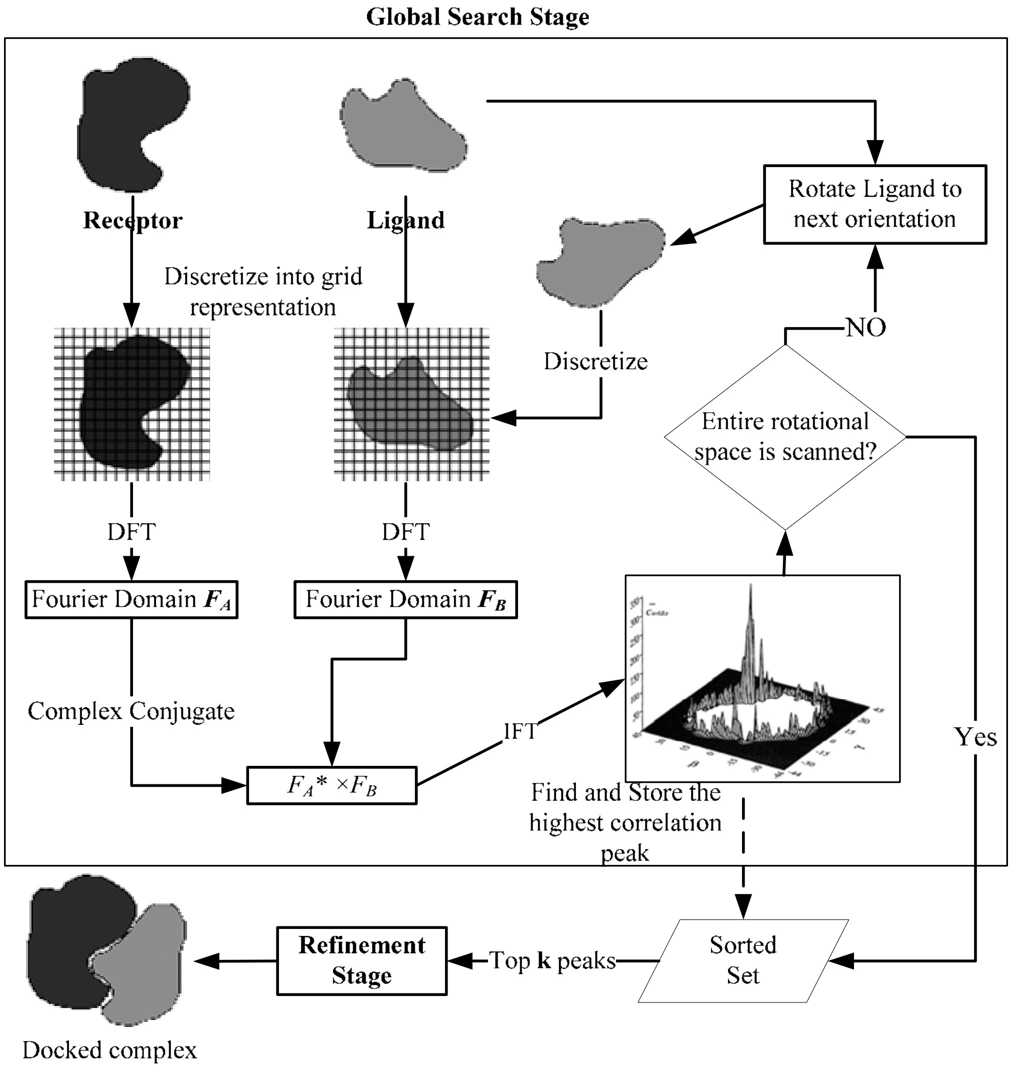


Figure 4.4: The flowchart of the docking algorithm based on the approach of Katchalsi-katzir et al.

As shown in Figure 4.4, this new algorithm also consists of two stages. The first stage is a global search which can be either coarse or fine depending on the choice of grid step **η** and number of nodes **N**. The top **k** peaks found in the first stage will passed to the next stage. In the refinement stage, each of the **k** orientations at which the top **k** peaks are found will be refined by performing an iterative coarse-to-fine search in its neighbouring rotational space for the locally best orientation to dock the receptor and the ligand in a shape-complementary manner. The translational space scanning by FFT will be performed for every newly sampled orientations using a grid specified by **N'** and **η'**, which could be either finer or the same as the one used in previous stage. The detail procedure of refining one orientation is given below:

Δ_0 *= angular step in global search stage.*
(Φ, θ, ψ) *= an orientation that needs to be refined.*

```
For $\Delta = \Delta_0/2, \Delta_0/4, \ldots\ldots, \Delta_0/2^m \geq 1^o$
    For each of the $3 \times 3 \times 3$ orientations ( $\boldsymbol{\Phi}$', $\boldsymbol{\theta}$', $\boldsymbol{\psi}$') within a small neighborhood of size
    $\Delta$ centered at ( $\boldsymbol{\Phi}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$),
        Find the highest correlation peak **P** at ( $\boldsymbol{\Phi}$', $\boldsymbol{\theta}$', $\boldsymbol{\psi}$') by FFT correlation approach
        with a grid of grid step $\boldsymbol{\eta}$' and number of nodes **N'**.
    Replace ( $\boldsymbol{\Phi}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$) by the angles ( $\boldsymbol{\Phi}$', $\boldsymbol{\theta}$', $\boldsymbol{\psi}$') at which the largest peak **P** is found.
Return ( $\boldsymbol{\Phi}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$) as refined and its corresponding peak.
```

After refinement is performed for all **k** orientations, all the peaks corresponding to the refined orientations will be sorted again according to their surface correlation scores. And then they will be used to generate docked complexes for the input docking case.

The time complexity of this algorithm is determined by **N**, **N'**, **k** and **Δ**. The time complexity for the global search stage is $O(\frac{N^3 \log(N)}{\Delta^3})$ while the time complexity for the refinement stage is $O(kN'^3 \log(N')\log(\Delta))$. In actual running, for **N, N'**= 128, **Δ** = 15° and **k** = 60, the running time for the global search stage and refinement stage are roughly 65 minutes and 50 minutes using a Pentium 4 2.0G CPU.

Compared to the algorithm proposed by Katchalsi-katzir et al, this new algorithm can still exploit the coarse global search and fine discrimination technique to reduce computational load depending on the choice of related parameters. Furthermore, the use of the refinement stage will remarkably increase the quality of docked complexes (See Figure 4.5).

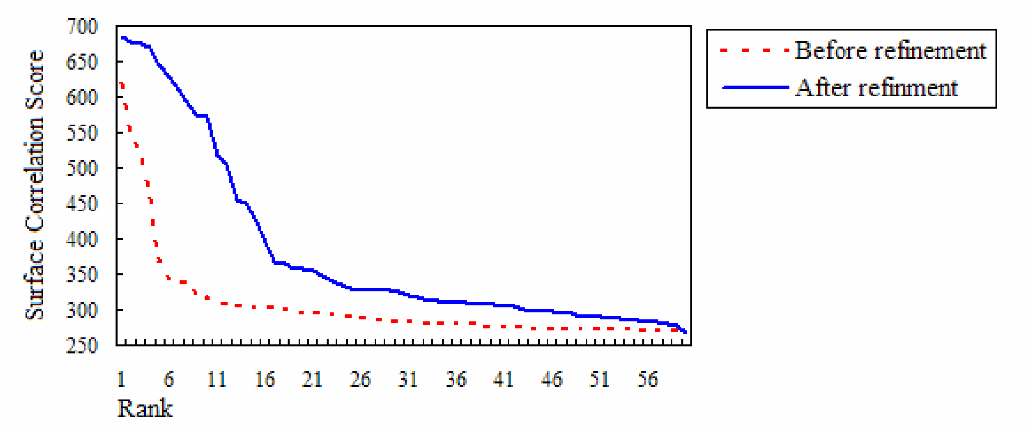


Figure 4.5: The effect of refinement. Plotting 60 peaks from 1ABI case in table 5.5 using rank as X-axis and surface correlation score as Y-axis. As can be observed from the graph, after refinement, the surface correlation scores become larger for each peaks.

4.2. Experimental Results on different configuration of parameters.

To implement the new algorithm, a number of parameters have to be specified. Those parameters can be classified into two groups:

- Parameters that specify the protein model: ***r:*** the radius of an atom; ***t:*** the surface thickness; ***ρ***: the value assigned to the nodes inside the molecule in the receptor grid; ***δ***: the value assigned to the nodes inside the molecule in the ligand grid;
- Parameters that control the algorithm running: **Δ**: the angular step; **N**: the number of nodes in each dimension of the grid in global search stage; **η**: the grid step size used in global search stage; **N'**: the number of nodes of the grids used in refinement stage; **η'**: the grid step size used in refinement; **k**: the number of peaks passed to the refinement stage.

An additional constraint should be noticed that the products **Nη** and **N'η'** have to be larger than any potential complex otherwise the algorithm might function improperly due to the periodicity of Fourier space.

4.2.1. Two existing protein models for grid-based docking

4.2.1.1. Katchalsi-katzir's model and corresponding parameter values

In Katchalsi-katzir's implementation, atom radius ***r*** is 1.8 Å which is 0.2 Å larger than the average VDW radius of carbon, nitrogen and oxygen. The additional 0.2 compensated for fact that the hydrogen atoms are not projected on the grids. Thickness ***t*** is chosen to be 2.0 Å which is used to tolerate the penetration due to conformation change. ***ρ*** and ***δ***, the value for interior nodes, are assign to -15 and 1 respectively (See a visualization in Figure 4.6).

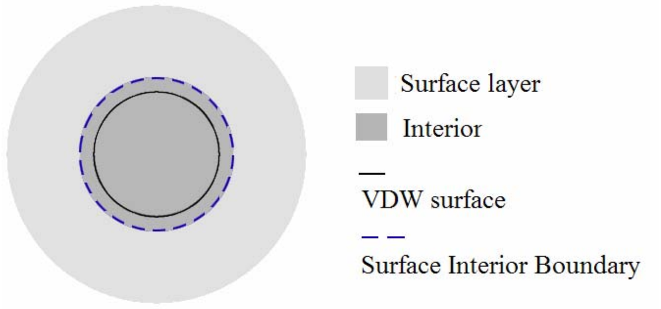


Figure 4.6: One atom in Katchalsi-katzir's protein model. This model is for both receptor and ligand. The VDW surface is contained in the interior region.

4.2.1.2. Gabb's model and corresponding parameter values

Gabb et al (1997) proposed to use two different models for receptor and ligand respectively. They also assigned 1.8 Å to ***r***, but 1.5 Å for ***t***. The major difference is that they chose to model the ligand with no surface layer. In other words, they chose to assign 0 to nodes within the surface layer of ligand molecule whereas Katchalsi-katzir and his colleagues assigned 1 (See Figure 4.7). ρ and δ in Gabb's model are also set to be -15 and 1.

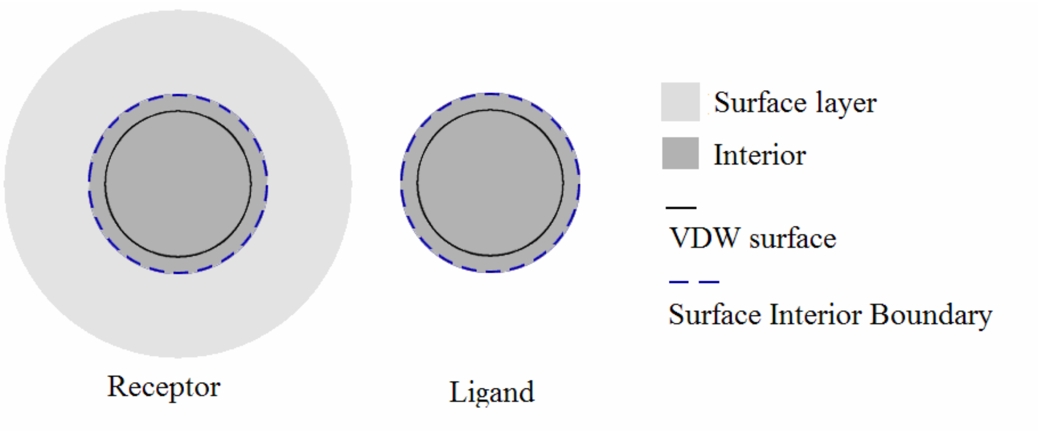


Figure 4.7: Two atoms in Gabb's protein models for receptor and ligand respectively. As can be observed from the above Figure, the ligand has no surface layer.

4.2.2. Performance of Katchalsi-katzir's model

To find a set of suitable parameters for Katchalsi-katzir's model, several experiments using fifteen bound cases have been conducted. It can be concluded that the Katchalsi-katzir's protein model is not suitable for the new docking algorithm according to the experimental presented in following subsections.

4.2.2.1. Experiment 1

The first experiment was conducted on the values used by Katchalsi-katzir (See Table 4.1 A.). The angular step **Δ** was set to 20º because they believed the 2.0 Å surface thicknesses yield an angular tolerance of about ±10º. **N**, **η**, **N'** and **η'** were set to 90, 1.1Å, 128, 0.8Å to compromise between computation load and docking quality. The choice of 0.8Å for **η'** is because it is the half of carbon-carbon bond length.

Parameter	**Δ**	**N**	**η**	**N'**	**η'**	ρ	δ	**k**
Value	20º	90	1.1Å	128	0.8Å	-15	1	60

Table 4.1 A: Parameters used in Experiment 1. (Katchalsi-katzir et al 1992).

Cases		1CHO	1ABI	1ACB	1CSE	1TGS	2KAI	2MHB	2PTC	3HFM
	Rank	Fail	52	40	37	25	48	17	5	34
	L_RMS (Å)	Fail	13.46	15.2	14.5	8.22	13.85	14.97	11.2	17.3

Cases		4HVB	4SGB	4TPI	9LDT	9RSA	1FDL	2SIC
	Rank	Fail	14	19	4	55	Fail	—
	L_RMS (Å)	Fail	13.55	12.71	12.28	14.47	Fail	—

Table 4.1 B: Results from Experiment 1: Rank and L_rms rows list the rank and L_rms of the docked complex with the lowest L_rms. A case is considered as **fail** if the lowest L_RMS is bigger than 18 Å. See Chapter 3 for reasons of using these quantities to assess the performance. —: not tried.

As can be seen in above table, only 1TGS case is acceptable if we strictly follow CAPRI evaluation criteria (See Chapter 3). Nevertheless, some cases can still be considered as roughly correct because the correct protein-protein interaction interface can be roughly observed in the docked complex, for example, 2MHB. Although its lowest L_RMS is 14.97Å, certain degree of similarity between the best docked complex and the correct complex can still be observed (See Figure 4.7). However, such accuracy is far from good enough for docking applications.

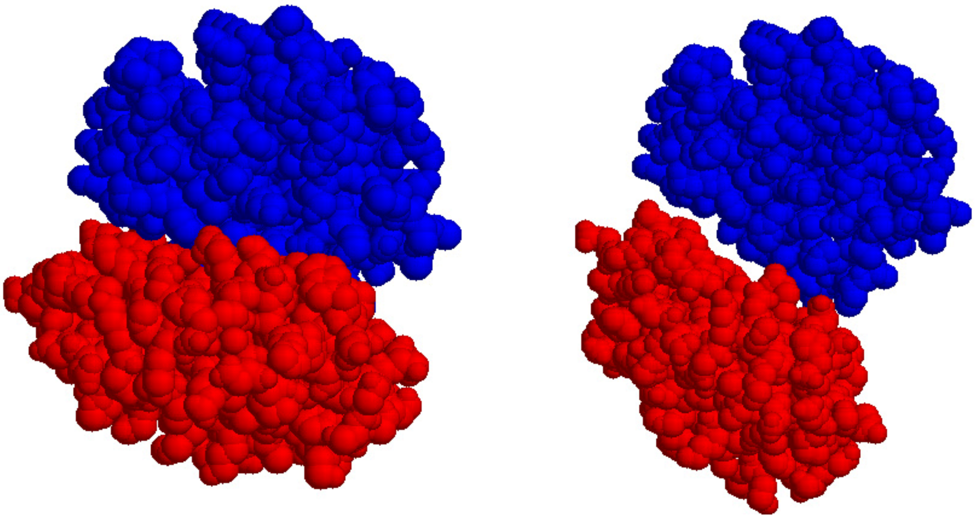


Figure 4.7: 2MHB case: Although the ligand (red) in the docked complex (right) is skewed and deviated compared to the ligand in the correct complex (left), the correct interaction interface can still be roughly observed from the docked complex.

4.2.2.2. Experiment 2 , 3, 4.

Three other experiments were conducted in order to determine whether the poor performance of the algorithm using Katchalsi-katzir's protein model is due to the choice of parameters (other than those parameters for modeling protein).

Parameter	Δ	N	η	N'	η'	ρ	δ	r	t	k
Value	15°	90	1.1Å	128	0.8Å	-15	1	1.8 Å	2.0 Å	60

Table 4.2 A: Parameters used in Experiment 2.

Parameter	Δ	N	η	N'	η'	ρ	δ	r	t	k
Value	20°	128	0.8 Å	128	0.8Å	-15	1	1.8 Å	2.0 Å	60

Table 4.2 B: Parameters used in Experiment 3.

Parameter	Δ	N	η	N'	η'	ρ	δ	r	t	k
Value	20°	128	0.8 Å	128	0.8Å	-5	1	1.8 Å	2.0 Å	60

Table 4.2 C: Parameters used in Experiment 4.

Cases	Expriment 2		Experiment 3		Experiment 4	
	Rank	L_RMS	Rank	L_RMS	Rank	L_RMS
1CHO	Fail	Fail	Fail	Fail	Fail	Fail
1ABI	Fail	Fail	55	15.42	29	15.64
1ACB	22	13.11	51	14.31	49	13.13
1CSE	18	6.80	8	14.02	53	12.57
1TGS	51	7.77	23	7.69	37	6.80
2KAI	6	13.41	33	9.70	59	9.43
2MHB	15	4.70	45	13.59	4	2.03
2PTC	31	9.48	39	10.23	6	12.33
3HFM	Fail	Fail	Fail	Fail	Fail	Fail
4HVB	28	17.32	33	15.11	51	14.35
4SGB	27	14.37	33	9.73	32	5.55
4TPI	19	11.16	30	12.28	60	10.51
9LDT	54	17.06	26	12.91	46	12.30
9RSA	9	17.92	10	6.01	35	14.70
1FDL	51	14.99	Fail	Fail	Fail	Fail
2SIC	—	—	—	—	—	—

Table 4.2 D: Resullts for experiment 2 — 4

The poor performance remained, although for some cases the performance was slightly better.

4.2.2. Performance of Gabb's model

Only one experiment using all parameters proposed by Gabb et al (1997) was conducted for this model. The result (See Table 4.3) is still not satisfactory, although there are five correct cases (2MHB, 2PTC, 4SGB, 4TPI and 9RSA) according to CAPRI criteria.

Parameter	Δ	**N**	η	**N'**	η'	ρ	δ	R	t	**k**
Value	15º	128	0.8 Å	128	0.8Å	-15	1	1.8 Å	1.5 Å	60

Table 4.3 A: Parameters used in Experiment 5 on Gabb's model: Ligand has no surface layer. (Gabb et al 1997)

Cases		1CHO	1ABI	1ACB	1CSE	1TGS	2KAI	2MHB	2PTC	3HFM
	Rank	Fail	Fail	16	1	21	1	1	10	Fail
	L_RMS (Å)	Fail	Fail	10.70	16.84	14.65	12.72	7.96	1.03	Fail

Cases		4HVB	4SGB	4TPI	9LDT	9RSA	1FDL	2SIC
	Rank	30	1	7	39	2	58	30
	L_RMS (Å)	16.01	0.56	4.89	12.81	4.05	18.03	16.01

Table 4.3 B: Results from Experiment 5.

Chapter 5

A new protein model and its performance

The experimental results in chapter 4 show the proposed docking algorithm doesn’t perform well using Katchalsi-Katzir’s and Gabb’s protein models. A new protein model must be proposed to achieve high quality docking. This chapter will introduce the new model in the first section, whereas its performance on both bound and unbound cases will be presented in the second section.

5.1. A double-layer protein model

By investigation of the failed cases of previous experiments, it could be found that there is an always-fail case, 1CHO. This interesting case is the starting point of the new protein model.

5.1.1 The Drawback of both protein models

The 3D structure of 1CHO shows very close contact between nearly half of entire VDW surfaces of the receptor and the ligand. Furthermore, the receptor and the ligand are actually combined in a ‘locked’ manner: three protrusions on the ligand were tightly held by the same number of holes on the receptor (See Figure 5.1).

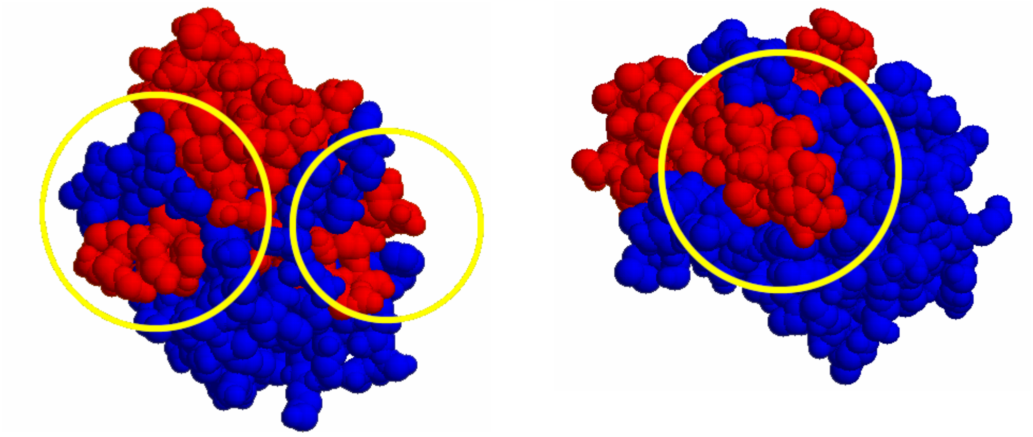


Figure 5.1: 3D structure of 1CHO. The protrusion-hole pairs are marked by yellow circles. A close contact can also be observed between the receptor (blue) and the ligand (red)

However, close contact between VDW surfaces of two proteins is prohibited if we use Katchalsi-Katzir's model for docking. This is because of 2.0Å thick surface layers of both receptor and ligand. Once the actual VDW surfaces of two atoms of different proteins are in contact, the surface layer of one atom will deeply penetrated into the interior of the other one (See Figure 5.2). Such penetration will result in a deduction of correlation score because $\boldsymbol{\rho}$ is set to be << -1 and $\boldsymbol{\delta}$ falls in [0, 1]. Therefore, a small gap between the receptor and ligand can always be observed in any docked complex produced using this model (See Figure 5.3). That is why 1CHO case always fails, no matter how we change the other parameters. One may argue that it's possible to assign a small negative value to $\boldsymbol{\rho}$, but this might cause the docking procedure to fail to reject deep penetrations for some cases. Besides, the thicker the surface layer is; the more faulty matches could be.

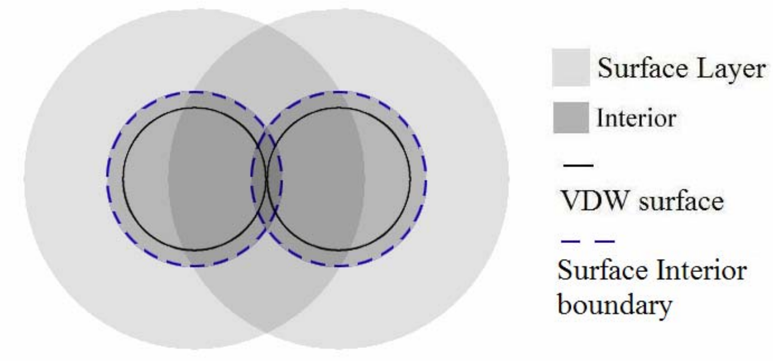


Figure 5.2: Surface layer of one atom penetrates into the interior of other atom when their VDW surfaces are in contact. Negative contribution will be made to the correlation score If these two atoms residing on different proteins.

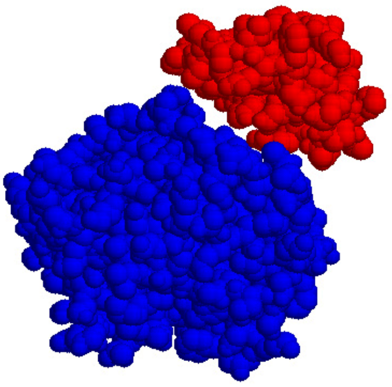


Figure 5.3: A gap between the receptor and ligand can be easily observed in the docked complex produced by the proposed algorithm using Katchalsi-Katzir's model.

Gabb's model allows such close contact between the VDW surfaces by eliminating the surface layer of ligand. If two atoms' VDW surface are in contact, only positive contribution will be made to total correlation score. However, using this model for the algorithm still failed to generate a complex that was close to the original 1CHO. The possible reasons for that might be: 1) Gabb's choice of -15 for $\boldsymbol{\rho}$ is too large, which may result in penetrations to be too heavily penalized. Then, even though an orientation close to the correct orientation is sampled, its peak's correlation score could not be higher enough for refinement stage. 2) This model requires an angular step smaller that 15º because of the single layered model of ligand, which may result in missing orientations in global scan stage. Both these two could be the reasons for the absence of close to correct one among the docked complexes.

5.1.2 The new protein model

To overcome the drawbacks, a novel double-layer protein model has been proposed for both the receptor and ligand are modeled in the same way. The fundamental difference between this new model and other two is that it models the protein to have an inner core layer and an surface layer such that the VDW surface is approximately in the middle of the surface layer (See Figure 5.4). This model allows contact between VDW surfaces (See Figure 5.5), while its angular tolerance is about 15º (See next section).

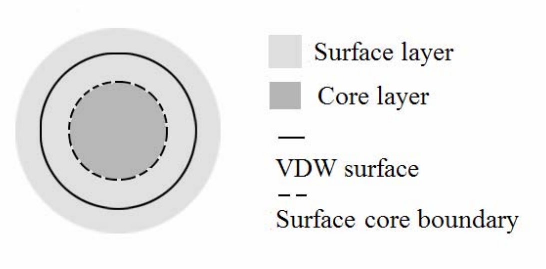


Figure 5.4: An atom in the new protein model with a core layer and surface layer. This model is applied to both receptor and ligand.

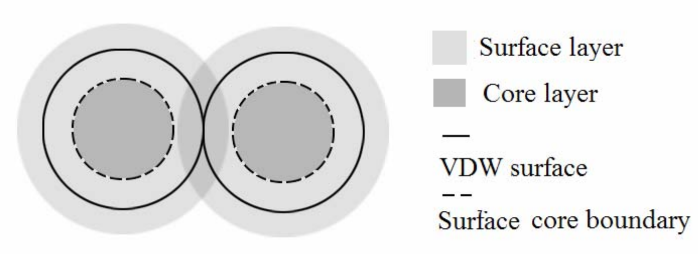


Figure 5.5: Surface layers of both atoms overlap each other when their VDW surfaces are in contact. Positive contribution will be made to correlation score if both atoms residing on different proteins.

This model can be easily used by the docking algorithm by assigning 1.1Å to both $\boldsymbol{r}$ and $\boldsymbol{t}$. After several experiments (See next section), -5 and 1 is found to be suitable values for $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ while 1 is still assigned to the surface layer.

5.1.3. Configurations of other parameters

The new model alone is not enough for the proposed algorithm to run; several other parameters have to be specified. As stated by Katchalsi-Katzir et al (1992), "optimal results were obtained when the grid step size was 0.7-0.8Å". In refinement stage, $\boldsymbol{\eta'}$ is fixed to be 0.8Å. The corresponding numbers of nodes **N'** have a default value 128, but it could be larger if 128 ×0.8Å is smaller than any potential docked complex (Refer to pg. 24). -5 is assigned to $\boldsymbol{\rho}$ for two reasons: 1) Docking using Katchalsi-Katzir's model will produce a docked complex with a small gap between the receptor and ligand. The original intention for this gap to exist is to tolerate certain degree of conformation changes. However, docking using the new model won't produce such gaps; hence this degree of conformation changes has to be tolerated by assigning a smaller negative value to $\boldsymbol{\rho}$ to core layer. 2) A thinner surface layer leads to a smaller angular tolerance. A smaller angular tolerance requires a smaller angular step $\boldsymbol{\Delta}$ which will increase the computation load. Therefore, a smaller $\boldsymbol{\rho}$ should be used to enhance the angular tolerance. In order to figure out the appropriate values for the rest of the parameters, several experiments using the bound cases have been conducted.

Experiment 6 is designed to determine the angular step $\boldsymbol{\Delta}$. In this experiment, $\boldsymbol{\Delta}$ is set to 20° while $\boldsymbol{\eta}$ is assigned to be the optimal value 0.8Å. Among the results (See Table 5.1), only one case is incorrect according to CAPRI's 10Å L_RMS upper bound for correct docked complexes. It can also be observed that the lowest L_RMS of the correct cases are all quite small. This suggests that the correct orientations were actually missed during the global search stage for the incorrect case. Therefore, it could be concluded that 20° as angular step is too large for the new model.

Parameter	$\boldsymbol{\Delta}$	**N**	$\boldsymbol{\eta}$	**N'**	$\boldsymbol{\eta'}$	$\boldsymbol{\rho}$	$\boldsymbol{\delta}$	$\boldsymbol{r}$	$\boldsymbol{t}$	$\boldsymbol{k}$
Value	20°	128*[1]	0.8 Å	128*	0.8Å	-5	1	1.1 Å	1.1 Å	60

Table 5.1 A: Values used in Experiment 6. Shadowed column is the focus of this experiment.

Cases		1CHO	1ABI	1ACB	1CSE	1TGS	2KAI	2MHB	2PTC	3HFM
	Rank	1	1	5	11	1	3	1	1	1
	L_RMS (Å)	0.96	0.85	1.68	1.00	0.97	1.61	1.26	1.28	0.65

Table 5.1 B: Results from Experiment 6.

[1] '*' mark indicates that the marked value could be altered. For example, if 0.8 × 128 is smaller than any possible docked complex (Refer to pg. 24), a larger value has to be assigned to **N** and **N'** whereas $\boldsymbol{\eta}$ and $\boldsymbol{\eta'}$ are still fixed at 0.8Å.

Cases		4HVB	4SGB	4TPI	9LDT	9RSA	1FDL	2SIC
	Rank	1	9	54	2	—	44	4
	L_RMS (Å)	1.41	1.04	14.50	1.30	—	1.96	1.54

Table 5.1 B cont'd: Results from Experiment 6.

Experiment 7 is designed for testing whether 1.1Å is an appropriate value for grid step size $\boldsymbol{\eta}$ in global search stage. In this experiment, $\boldsymbol{\Delta}$ is set to 15 º in order to minimize the influence of angular step on the docked results. Among the result (See Table 5.2), three cases are failed and other two cases are incorrect according to CAPRI. These results clearly show that 1.1Å is not an appropriate choice for $\boldsymbol{\eta}$ although it could reduce the computation load.

Parameter	$\boldsymbol{\Delta}$	**N**	$\boldsymbol{\eta}$	**N'**	$\boldsymbol{\eta}'$	$\boldsymbol{\rho}$	$\boldsymbol{\delta}$	***r***	***t***	***k***
Value	15º	90*	1.1 Å	128*	0.8Å	-5	1	1.1 Å	1.1 Å	60

Table 5.2 A: Values used in Experiment 7. Shadowed column is the focus of this experiment

Cases		1CHO	1ABI	1ACB	1CSE	1TGS	2KAI	2MHB	2PTC	3HFM
	Rank	1	1	17	2	1	6	1	2	Fail
	L_RMS (Å)	0.89	1.03	14.06	2.32	1.95	3.38	1.25	1.27	Fail

Cases		4HVB	4SGB	4TPI	9LDT	9RSA	1FDL	2SIC
	Rank	1	7	1	Fail	2	Fail	—
	L_RMS (Å)	1.36	1.34	1.52	Fail	17.51	Fail	—

Table 5.2 B: Results from Experiment 7.

Experiment 8 aims to verify -5 is a good choice for $\boldsymbol{\rho}$ by investigation of the results produced by $\boldsymbol{\rho}$ = -10. $\boldsymbol{\Delta}$ and $\boldsymbol{\eta}$ are set to 15º and 0.8 to minimize their influence on the results. Four failed cases (See Table 5.3) could be found in this experiment. This poor performance clearly demonstrates that, the magnitude of $\boldsymbol{\rho}$ could not be too large in order for the docking algorithm to function properly.

Parameter	$\boldsymbol{\Delta}$	**N**	$\boldsymbol{\eta}$	**N'**	$\boldsymbol{\eta}'$	$\boldsymbol{\rho}$	$\boldsymbol{\delta}$	***r***	***t***	***k***
Value	15º	128*	0.8 Å	128*	0.8Å	-10	1	1.1 Å	1.1 Å	60

Table 5.3 A: Values used in Experiment 8. Shadowed column is the focus of this experiment

Cases		1CHO	1ABI	1ACB	1CSE	1TGS	2KAI	2MHB	2PTC	3HFM	4HVB	4SGB	4TPI	9LDT	9RSA	1FDL	2SIC
	Rank	Fail	1	1	1	1	3	23	1	Fail	2	22	1	Fail	5	Fail	2
	L_RMS (Å)	Fail	1.13	1.92	0.61	0.40	0.75	14.20	1.59	Fail	0.85	14.22	1.35	Fail	17.76	Fail	1.19

Table 5.3 B: Results from Experiment 8.

From these three experiments, an optimal configuration for the parameters of the proposed algorithm could be deduced (See Table 5.4) as listed in the following table.

Parameter	**Δ**	**N**	**η**	**N’**	**η'**	***ρ***	***δ***	***r***	***t***
Value	15°	128*	0.8 Å	128*	0.8Å	-5	1	1.1 Å	1.1 Å

Table 5.4: The best configuration for the parameters of the proposed algorithm.

5.2. Performance of the algorithm with the new model.

To evaluate the performance of the proposed algorithm with the best configuration, sixteen bound cases and six CAPRI unbound cases have been tried. The results from these trials are quite encouraging. The execution time for each cases were also recorded, a typical running time using the best configuration is about two hours on a P4 2.0 GHz machine. **k** is fixed as 60 in the following experiments.

5.2.1. On the bound cases.

Cases		1CHO	1ABI	1ACB	1CSE	1TGS	2KAI	2MHB	2PTC	3HFM	4HVB	4SGB	4TPI	9LDT	9RSA	1FDL	2SIC
	Rank	1	1	6	10	5	3	2	3	15	1	8	6	3	1	50	5
	L_RMS (Å)	0.66	0.65	1.23	0.60	0.78	0.87	1.34	1.08	0.93	1.00	0.98	0.87	1.73	1.37	1.34	0.81

Table 5.5: Results for sixteen bound cases produced by the proposed algorithm with the best configuration.

Results listed in Table 5.5 are quite good in terms of the lowest L_RMS. The lowest

L_RMS of ten cases fall in the interval from 0 to 1, which means that ten high quality docked complexes have been produced by the algorithm according to the CAPRI criteria. The other six cases' lowest L_RMS fall in the interval from 1 to 5, which could be considered as with medium accuracy. The results are much better compared to the results produced by the algorithm using the other two models presented in previous chapter. This demonstrates the ability of the proposed algorithm to match the protein surfaces as long as appropriate configuration is specified.

In terms of the ranking, ten cases have ranked the lowest L_RMS docked complexes in top five, while four cases ranked such complexes in top ten. There are two exceptions, 3HFM case and 1FDL case. The algorithm with the best configuration ranked the best docked complex as 15th for 3HFM case and 50th for 1FDL case. These two cases have the same receptor that has a hole in the center. The algorithm always tends to place the ligand into that hole, because such placement appears to be more 'complementary' in term of shape. As shown in Figure 5.6 A and B, it is obvious that the shape complementarity between receptors and ligands in the two top ranked docked complexes is better than the complementarity in the real complex. This fact reveals that shape complementarity alone is not powerful enough to discriminate the false positives. A conclusion could be made that a high surface correlation score doesn't necessarily indicate a correctly docked complex (See Figure 5.7).

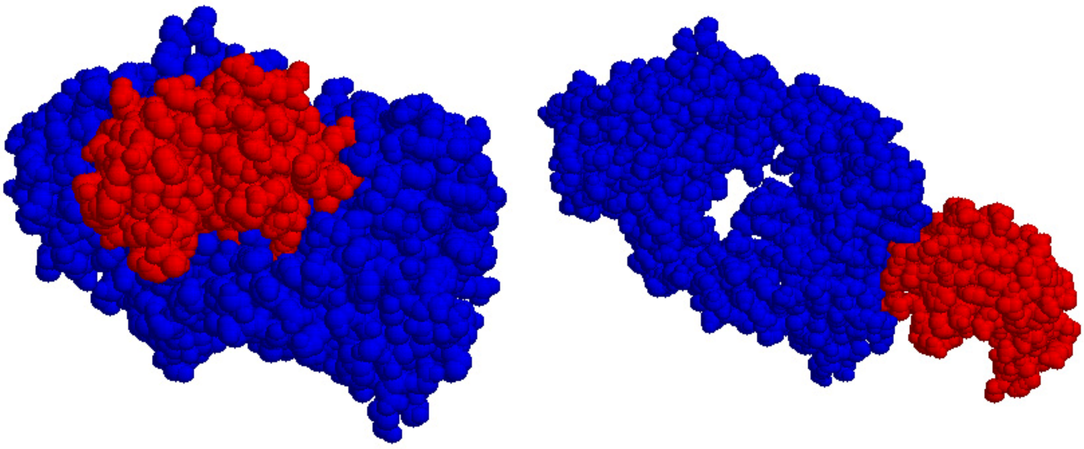


Figure 5.6 A: the top ranked complex (left) for 1FDL case vs. the structure of the real 1FDL. Receptors are colored in blue whereas ligands are colored in red.

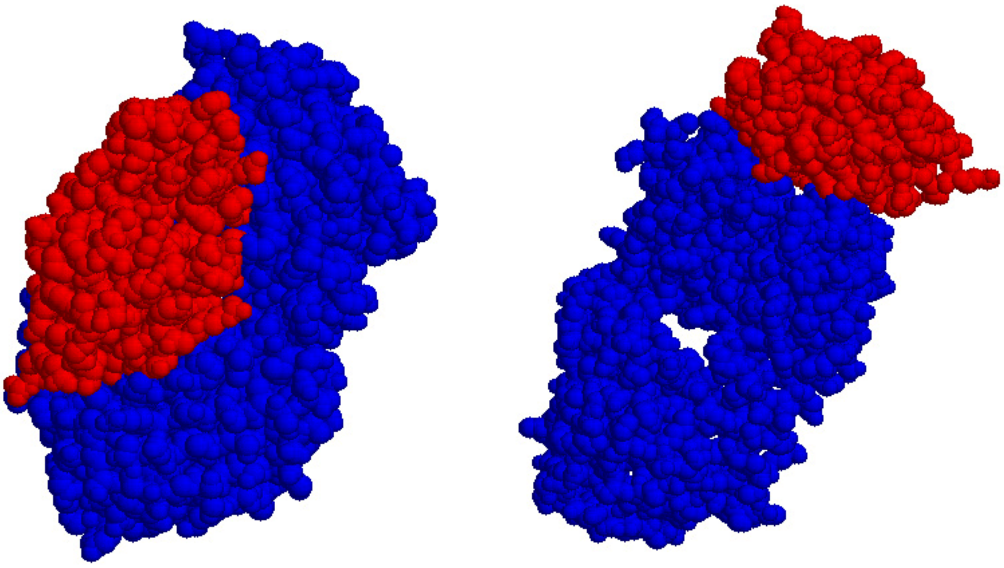


Figure 5.6 B: the top ranked complex (left) of 3HFM case vs. the structure of the real 3HFM. Receptors are colored in blue whereas ligands are colored in red.

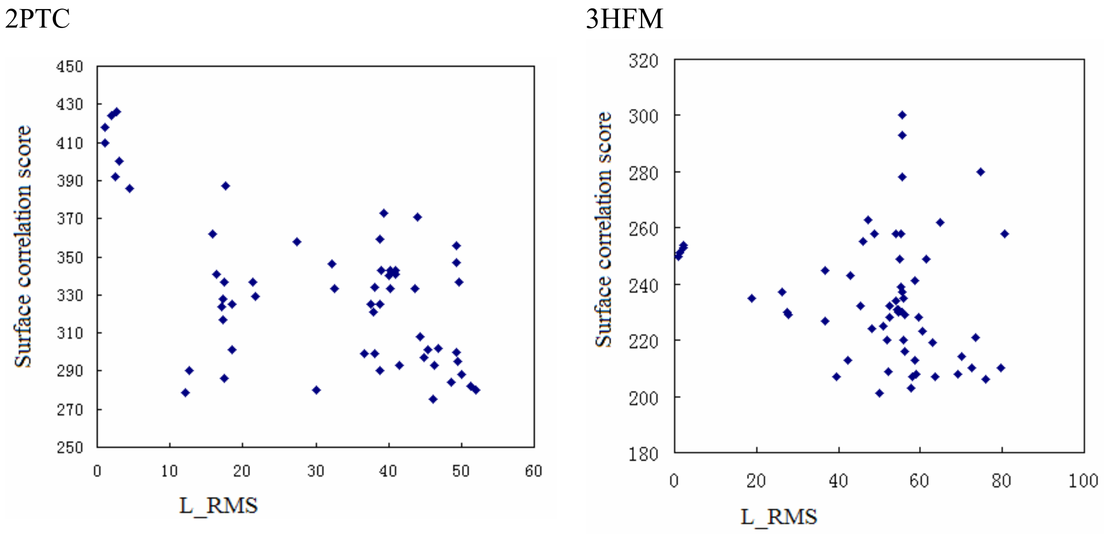


Figure 5.7: Plotting the 60 docked results of 2PTC case and 3HFM case using surface correlation score as Y axis and L_RMS as X axis, A high surface score does correspond to low L_RMS in 2PTC plot. However, in 3HFM plot, such relation can not be observed. A high surface correlation score may not necessarily indicate a correctly docked complex.

5.2.2. On the unbound cases.

Only the top ten docked complexes for each CAPRI case produced by the proposed algorithm with the best configuration were evaluated. This strictly followed the CAPRI procedure in which each participant can only submit 10 predictions. As shown in table 5.6; the docking algorithm failed in four cases while succeeded in two cases. It is not surprising that there are four failed cases, because the docking algorithm uses only shape complementarity for docking while the participants of CAPRI experiments employs various kinds of chemical and physical information to guide their docking. But it is quite encouraging that there are one high quality case and one medium quality case. This fact shows that shape complementarity could be the dominant factor for some protein-protein interactions.

Capri Cases		2	3	4	5	6	7
	L_RMS (Å)	36.32	44.95	26.68	2.055	1.023	36.18
	I_RMS (Å)	22.87	24.93	15.59	1.100	0.65	27.46
	Fnat	0.0	0.0	0.0107	0.95	0.908	0.0
Quality		incorrect	incorrect	incorrect	medium	high	incorrect

Table 5.6: Results on CAPRI cases produced by the proposed algorithm with the best configuration. Refer to Chapter 3 for details on how to determine the quality of a docked complex. I_RMS, L_RMS and Fnat rows lists the corresponding value of the complex with lowest I_RMS.

Another interesting fact is that every participant of CAPRI failed the fifth case (See Appendix B) while it has been correctly predicted by my docking procedure based on only shape complementarity. Mendez et al (2003) has given the reason for why all participants failed the fifth case: they used prior knowledge to restrict the search space, but unfortunately, by doing so they also excluded the correct complex. On the contrary, a full search space scan is always performed by the proposed algorithm, and hence the correct docked complex could be generated. This fact demonstrates that applying biochemical and physical information is not always beneficial for docking, especially when such information would prune the search space.

Chapter 6

Conclusion

6.1. Conclusion

In this project, a study on the Katchalsi-katzir's grid-based FFT docking approach was presented. Although this approach is proposed 10 years ago, it could still be used as a foundation for protein docking algorithms. A new docking algorithm based on this approach was developed in this project. This new algorithm overcame the drawbacks of the original two-stage algorithm proposed by Katchalsi-katzir et al (1992) by removing degenerated orientations during rotational space scanning in the global scan stage and incorporating a refinement procedure into the discrimination stage. The refinement procedure carries out a coarse-to-fine adaptive search in the neighbouring rotational space for each of the given orientations from the global search stage. The quality of docking at this specific orientation is improved by iteratively replacing it with the neighbouring orientation which yields a better docking result in terms of shape complementarity.

Several Experiments were conducted in order to find the best configuration for the new algorithm. Two existing protein models designed for grid based protein representation were tried for this algorithm. However, both of these two models were not suitable for this new algorithm according to experimental results on different configurations for other parameters. By analyzing the drawbacks of two models, a new protein model was proposed. This protein model has a core layer and surface layer such that the VDW surface of the protein molecule is in the middle of the surface layer. In such a way, the contact between VDW surfaces of receptor and ligand is allowed while the angular tolerance remains relatively large. An optimal configuration for other parameters was found for this model and the proposed algorithm through several experiments. Pretty good results were obtained from the combination of the new model, the optimal configuration and the proposed algorithm on the sixteen bounded cases. Two unbound CAPRI cases were also successfully predicted in spite of only shape complementarity is used in this algorithm. These facts demonstrate the ability of the new algorithm on matching proteins when an appropriate configuration is specified.

Several facts have also been drawn from the experiments.

- Shape Complementarity alone is not powerful enough for discriminate false positives. Chemical or physical properties should also be employed.
- For some unbound cases, shape complentarity solely can successfully predicte the docked complex. However, it is not possible in general.
- Pruning of the search space by prior knowledge may not always be an advantage.

6.2. Future work

The work has been done so far is only a start in protein docking research. There are several problems that need to be addressed in future work:

- Biochemical and physical properties of the protein surface should be incorporated in the docking algorithm, but how to balance the weights of shape complementarity and other properties is still a problem. Intensive experiments should be employed in order to find an optimal solution.
- The rotational space scan sampled a lot of undesired orientations, how to detect these undesired orientation so that translational scan does not have to be performed for those orientations.
- Current rotational space sampling in global search stage is still biased even though degenerate orientations are removed. An optimal sampling strategy might be employed to achieve efficient sampling.

Reference

Betts, M.J., Sternberg, M.J. (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. Protein Engineering, Vol.19, 1999, pp: 271 – 283.

Chen, R., Weng, Z.P. (2002) Docking Unbound Proteins Using Shape Complementarity, Desolvation and Electrostatics. Proteins: Structure, Function, and Genetics, Vol.47, 2002, pp: 281 – 294.

Connolly, M. (1983). Analytical molecular surface calculation, Journal of Applied Crystallography, Vol.16, 1983, pp: 548 – 558

Connolly, M. (1986). Shape Complementarity at the hemoglobin α β subunit interface. Biopolyners, Vol.25, 1986, pp: 1229 – 1247

Crick, H.F.C. (1953). The packing of α-helices: simple coiled-coils. Acta Crystallographica, Vol.6, 1953, pp: 689 – 697

Eliot, D.F., Rao, K.R. (1982). Fast Fourier Transforms: Algorithms, Analyses, Applications, Academic Press, New York, 1982.

Gabb, H.A., Jackson, R.M., Sternberg, M.J.E. (1997). Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information. Journal of Molecular Biology, Vol.272, 1997, pp: 106 – 120.

Goldstein, H. (1980). Classical Mechanics, Addison–Wesley, Reading, Massachusetts, 1980. pp: 108.

Goodsell, D.S., Olson, A.J. (1990). Automated docking of substrates to proteins by simulated annealing. Proteins: Structure, Function & Genetics, Vol.8, 1990, pp: 195 – 202.

Halperin, I., Ma, B.Y., Wolfson, H., and Nussinov, R. (2002) Principle of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. Proteins: Structure, Function, and Genetics, Vol.47, 2002, pp: 409 – 443.

Hubbard, S. J., Argos, P. (1994). Cavities and packing at protein interfaces. Protein Science Vol.3, 1994, pp: 2194 – 2206.

Jones, G., Willet, P., Glen, R., Leach, A., Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. Journal of Molecular

Biology, Vol.267, 1997, pp: 727 – 748.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section D, Vol.34, 1978, pp: 827 – 828

Katchalski-katzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C., & Vakser, I.A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands hy correlation techniques, Proceedings of the National Academy of Sciences USA, Vol.89, 1992, pp: 2195 – 2199.

Lattman, E.E. (1971). Optimal Sampling of the Rotation Function, Acta Crystallographica Section B, Vol.28, 1972, pp: 1065 – 1066.

Lin, S.L., Nussibov R, Fischer, D., Wolfson. H.J. (1995). Molecylar surface representation by sparse critical points. Proteins: Structure, Function & Genetics, Vol.18, 1995, pp:94 – 101.

Mendez, R., Leplae, R., Maria, L.D., Wodak, S.J. (2003) Assessment of Blind Predictions of Protein-Protein Interactions: Current Status of Docking Method. Proteins: Structure, Function, and Genetics, Vol.52, 2003, pp: 51 – 67.

Norel, R., Lin, S.L., Wolfson, H.J., Nussinov, R. (1995). Molecular Surface Complementarity at Protein-Protein Interfaces: The Critical Role played by Surface Normals at Well Placed, Sparse, Points in Docking, Journal of Molecular Biology, Vol.252, 1995, pp: 263 – 273;

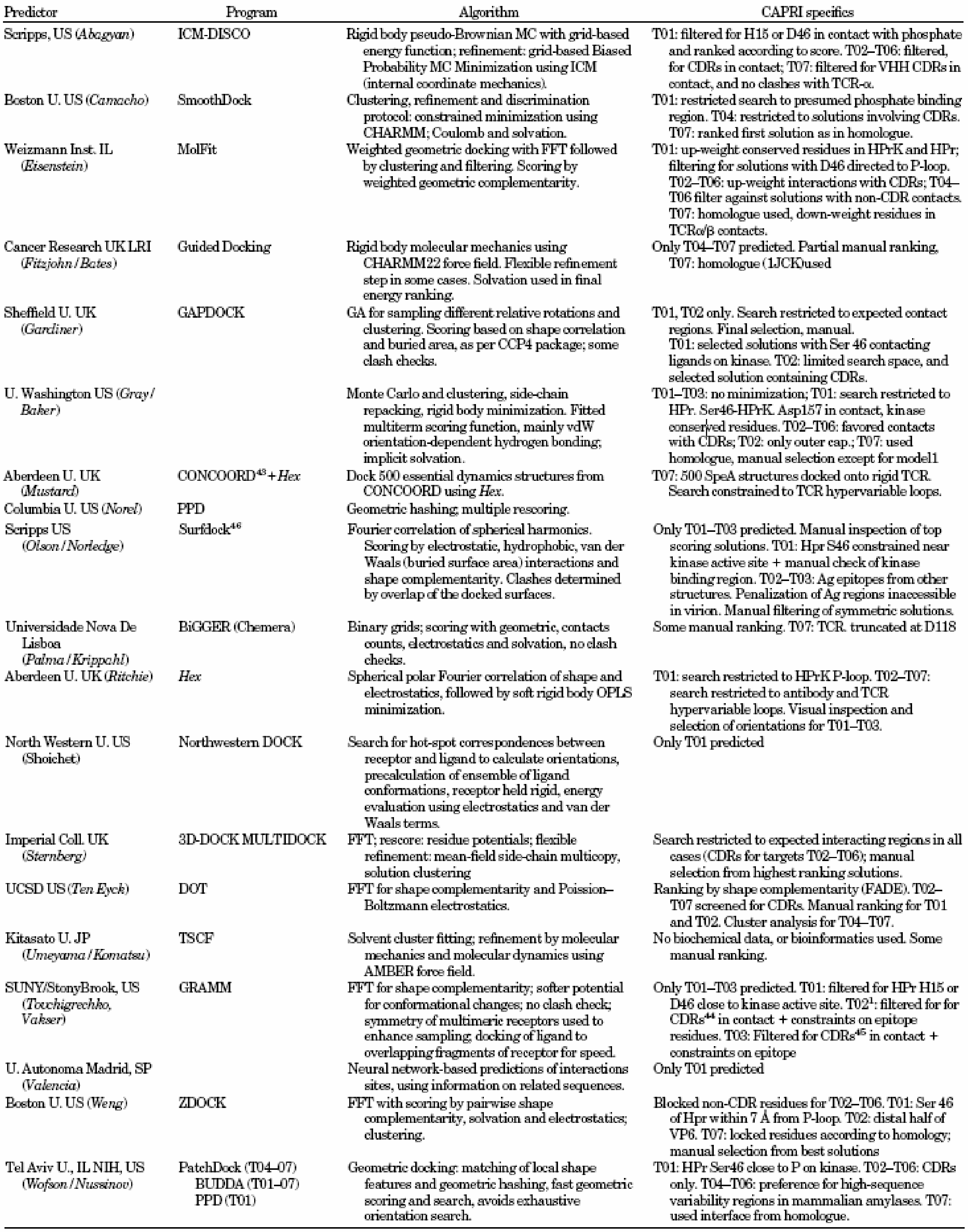
Norel, R., Petrey, D., Wolfson, H., Nussinov, R. (1999). Examination of shape complementarity in docking of unbound proteins. Proteins: Structure, Function & Genetics, Vol.35, 1999, pp: 403 – 419.

Strynadka, N.C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, BK., Kuntz, I.D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M., James, M.N. (1996). Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase, National Structural Biology USA, Vol.3, 1996, pp:233–239.

Totrov, M., Abagyan, R. (1994). Detailed ab initio prediction of lysozymeantibody complex with 1.6 Å accuracy, National Structural Biology USA, Vol.1, 1994, pp: 259 – 263.

Appendix A

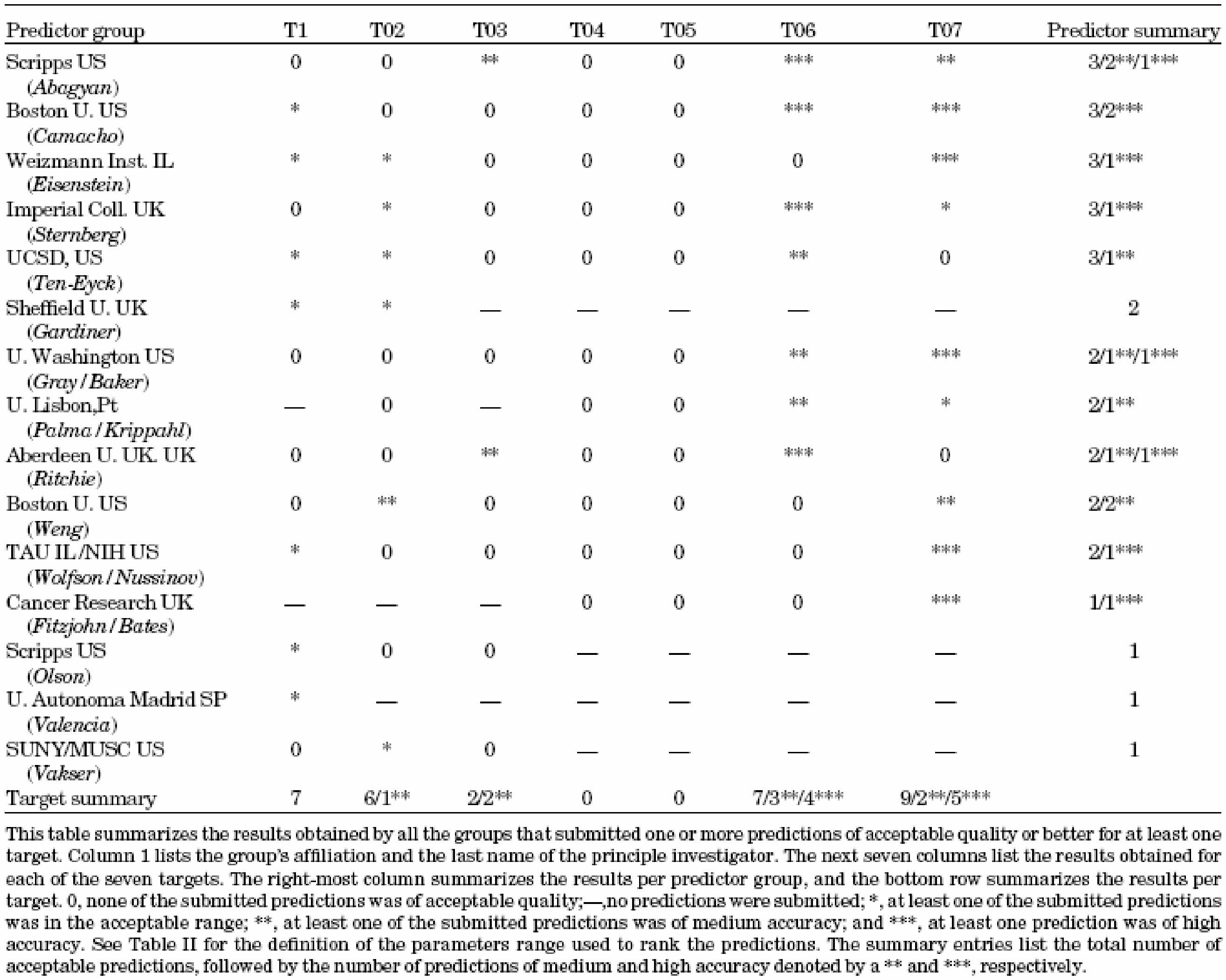
The participants of CAPRI experiment.



Predictor	Program	Algorithm	CAPRI specifics
Scripps, US (*Abagyan*)	ICM-DISCO	Rigid body pseudo-Brownian MC with grid-based energy function; refinement: grid-based Biased Probability MC Minimization using ICM (internal coordinate mechanics).	T01: filtered for H15 or D46 in contact with phosphate and ranked according to score. T02–T06: filtered, for CDRs in contact; T07: filtered for VHH CDRs in contact, and no clashes with TCR-α.
Boston U. US (*Camacho*)	SmoothDock	Clustering, refinement and discrimination protocol: constrained minimization using CHARMM; Coulomb and solvation.	T01: restricted search to presumed phosphate binding region. T04: restricted to solutions involving CDRs. T07: ranked first solution as in homologue.
Weizmann Inst. IL (*Eisenstein*)	MolFit	Weighted geometric docking with FFT followed by clustering and filtering. Scoring by weighted geometric complementarity.	T01: up-weight conserved residues in HPrK and HPr; filtering for solutions with D46 directed to P-loop. T02–T06: up-weight interactions with CDRs; T04–T06 filter against solutions with non-CDR contacts. T07: homologue used, down-weight residues in TCRα/β contacts.
Cancer Research UK LRI (*Fitzjohn / Bates*)	Guided Docking	Rigid body molecular mechanics using CHARMM22 force field. Flexible refinement step in some cases. Solvation used in final energy ranking.	Only T04–T07 predicted. Partial manual ranking, T07: homologue (1JCKused
Sheffield U. UK (*Gardiner*)	GAPDOCK	GA for sampling different relative rotations and clustering. Scoring based on shape correlation and buried area, as per CCP4 package; some clash checks.	T01, T02 only. Search restricted to expected contact regions. Final selection, manual. T01: selected solutions with Ser 46 contacting ligands on kinase. T02: limited search space, and selected solution containing CDRs.
U. Washington US (*Gray / Baker*)		Monte Carlo and clustering, side-chain repacking, rigid body minimization. Fitted multiterm scoring function, mainly vdW orientation-dependent hydrogen bonding; implicit solvation.	T01–T03: no minimization; T01: search restricted to HPr. Ser46-HPrK. Asp157 in contact, kinase conserved residues. T02–T06: favored contacts with CDRs; T02: only outer cap.; T07: used homologue, manual selection except for model1
Aberdeen U. UK (*Mustard*)	CONCOORD[43] + *Hex*	Dock 500 essential dynamics structures from CONCOORD using *Hex*.	T07: 500 SpeA structures docked onto rigid TCR. Search constrained to TCR hypervariable loops.
Columbia U. US (*Norel*)	PPD	Geometric hashing; multiple rescoring.	
Scripps US (*Olson / Norledge*)	Surfdock[46]	Fourier correlation of spherical harmonics. Scoring by electrostatic, hydrophobic, van der Waals (buried surface area) interactions and shape complementarity. Clashes determined by overlap of the docked surfaces.	Only T01–T03 predicted. Manual inspection of top scoring solutions. T01: Hpr S46 constrained near kinase active site + manual check of kinase binding region. T02–T03: Ag epitopes from other structures. Penalization of Ag regions inaccessible in virion. Manual filtering of symmetric solutions.
Universidade Nova De Lisboa (*Palma / Krippahl*)	BiGGER (Chemera)	Binary grids; scoring with geometric, contacts counts, electrostatics and solvation, no clash checks.	Some manual ranking. T07: TCR. truncated at D118
Aberdeen U. UK (*Ritchie*)	*Hex*	Spherical polar Fourier correlation of shape and electrostatics, followed by soft rigid body OPLS minimization.	T01: search restricted to HPrK P-loop. T02–T07: search restricted to antibody and TCR hypervariable loops. Visual inspection and selection of orientations for T01–T03.
North Western U. US (Shoichet)	Northwestern DOCK	Search for hot-spot correspondences between receptor and ligand to calculate orientations, precalculation of ensemble of ligand conformations, receptor held rigid, energy evaluation using electrostatics and van der Waals terms.	Only T01 predicted
Imperial Coll. UK (*Sternberg*)	3D-DOCK MULTIDOCK	FFT; rescore: residue potentials; flexible refinement: mean-field side-chain multicopy, solution clustering	Search restricted to expected interacting regions in all cases (CDRs for targets T02–T06); manual selection from highest ranking solutions.
UCSD US (*Ten Eyck*)	DOT	FFT for shape complementarity and Poission–Boltzmann electrostatics.	Ranking by shape complementarity (FADE). T02–T07 screened for CDRs. Manual ranking for T01 and T02. Cluster analysis for T04–T07.
Kitasato U. JP (*Umeyama / Komatsu*)	TSCF	Solvent cluster fitting; refinement by molecular mechanics and molecular dynamics using AMBER force field.	No biochemical data, or bioinformatics used. Some manual ranking.
SUNY/StonyBrook, US (*Touchigrechko, Vakser*)	GRAMM	FFT for shape complementarity; softer potential for conformational changes; no clash check; symmetry of multimeric receptors used to enhance sampling; docking of ligand to overlapping fragments of receptor for speed.	Only T01–T03 predicted. T01: filtered for HPr H15 or D46 close to kinase active site. T02[1]: filtered for for CDRs[44] in contact + constraints on epitope residues. T03: Filtered for CDRs[45] in contact + constraints on epitope
U. Autonoma Madrid, SP (*Valencia*)		Neural network-based predictions of interactions sites, using information on related sequences.	Only T01 predicted
Boston U. US (*Weng*)	ZDOCK	FFT with scoring by pairwise shape complementarity, solvation and electrostatics; clustering.	Blocked non-CDR residues for T02–T06. T01: Ser 46 of Hpr within 7 Å from P-loop. T02: distal half of VP6. T07: locked residues according to homology; manual selection from best solutions
Tel Aviv U., IL NIH, US (*Wofson / Nussinov*)	PatchDock (T04–07) BUDDA (T01–07) PPD (T01)	Geometric docking: matching of local shape features and geometric hashing, fast geometric scoring and search, avoids exhaustive orientation search.	T01: HPr Ser46 close to P on kinase. T02–T06: CDRs only. T04–T06: preference for high-sequence variability regions in mammalian amylases. T07: used interface from homologue.

Appendix B

Summary of Docking Results for each CAPRI participant, T02 is referred as CAPRI 02 in main text.



Predictor group	T1	T02	T03	T04	T05	T06	T07	Predictor summary
Scripps US (*Abagyan*)	0	0	**	0	0	***	**	3/2**/1***
Boston U. US (*Camacho*)	*	0	0	0	0	***	***	3/2***
Weizmann Inst. IL (*Eisenstein*)	*	*	0	0	0	0	***	3/1***
Imperial Coll. UK (*Sternberg*)	0	*	0	0	0	***	*	3/1***
UCSD, US (*Ten-Eyck*)	*	*	0	0	0	**	0	3/1**
Sheffield U. UK (*Gardiner*)	*	*	—	—	—	—	—	2
U. Washington US (*Gray / Baker*)	0	0	0	0	0	**	***	2/1**/1***
U. Lisbon,Pt (*Palma / Krippahl*)	—	0	—	0	0	**	*	2/1**
Aberdeen U. UK. UK (*Ritchie*)	0	0	**	0	0	***	0	2/1**/1***
Boston U. US (*Weng*)	0	**	0	0	0	0	**	2/2**
TAU IL /NIH US (*Wolfson / Nussinov*)	*	0	0	0	0	0	***	2/1***
Cancer Research UK (*Fitzjohn / Bates*)	—	—	—	0	0	0	***	1/1***
Scripps US (*Olson*)	*	0	0	—	—	—	—	1
U. Autonoma Madrid SP (*Valencia*)	*	—	—	—	—	—	—	1
SUNY/MUSC US (*Vakser*)	0	*	0	—	—	—	—	1
Target summary	7	6/1**	2/2**	0	0	7/3**/4***	9/2**/5***	

This table summarizes the results obtained by all the groups that submitted one or more predictions of acceptable quality or better for at least one target. Column 1 lists the group's affiliation and the last name of the principle investigator. The next seven columns list the results obtained for each of the seven targets. The right-most column summarizes the results per predictor group, and the bottom row summarizes the results per target. 0, none of the submitted predictions was of acceptable quality;—,no predictions were submitted; *, at least one of the submitted predictions was in the acceptable range; **, at least one of the submitted predictions was of medium accuracy; and ***, at least one prediction was of high accuracy. See Table II for the definition of the parameters range used to rank the predictions. The summary entries list the total number of acceptable predictions, followed by the number of predictions of medium and high accuracy denoted by a ** and ***, respectively.