Conformational Changes In Protein Interations

Thesis Proposal

Lu Haiyun

HT040829Y

luhaiyun@comp.nus.edu.sg

Supervisor:

A/Prof. Leow Wee Kheng

School of Computing

National University of Singapore

July 14, 2006

Abstract

Protein interactions play important roles in various aspects of the molecular mechanisms of biological processes. Protein docking techniques are used in pharmaceutical applications such as drug design to predict protein interactions especially for flexible proteins. Conformational shape changes occur in flexible proteins during interactions. Studies of protein flexibility show that conformational changes have a crucial influence on protein docking. Most existing algorithms handle only side-chain flexibility, which accounts for a small amount of conformational changes. Treatment of backbone flexibility, which accounts for major conformational changes, remains a challenge. In this proposal, we propose to model the conformational changes in protein interactions. Proper modeling of conformational changes will allow flexible docking algorithm to produce more accurate predictions of protein interactions.

Contents

1 Introduction

1.1 Motivation

Research on proteomics has increased in recent years. Much effort is devoted to large-scale analysis of 3D structures and dynamics of proteins. The goal is to achieve scientific and commercial breakthroughs in drug discovery, especially new drug development.

Drugs are usually small molecules. In human body, they bind to the disease-causing proteins (Figure 1) and prevent the disease-causing activities to happen. The binding mode of both molecules shown in Figure 1 is determined experimentally using x-ray diffraction.

Often, data are available for a protein and a drug separately, but not for the two together. It is very costly to find the binding information by laboratory experiments. In drug design industry, the structure of a drug is modified constantly in order to search for the most effective binding with a protein, leading to very high development costs. Therefore, there is a need to predict possible binding using computational algorithm.

A protein interacts with another protein as well. When two proteins are bound to each other, they form a protein *complex*. For most of the proteins known to science, it is not completely understood which other proteins they interact with.

Genetic diseases are known to be caused by mutated proteins, and there is a desire to understand what anomalous protein-protein interactions a given mutation may cause. In the distant future, proteins may be designed to perform biological functions, and the determination of the potential interactions of such proteins will be essential.

Figure 1: A small molecule drug (green) binds to a disease-causing protein (red).

1.2 Protein Docking Problem

Protein docking problem is a computational problem that predicts the binding of a protein with a potential interacting partner. There are two kinds of protein docking problems [16]. The simpler problem is *bound docking*. It attempts to reconstruct a complex using the *bound structures* of the two molecules involved. A bound structure is extracted from a structure containing more than one molecule, typically a complex of the molecules.

The more difficult problem is *predictive docking*, also referred to as the *unbound docking*. It attempts to construct a complex using the *unbound structures* of the two molecules. A protein in its unbound structure usually changes its 3D shape to bind with the other molecule. Thus, the difficulty of the problem increases. An unbound structure may be a *native structure*, a *pseudo-native structure*, or a *modeled structure*. A native structure is the structure of a molecule when it is free in solution. A pseudo-native structure is the structure of a molecule when it is complexed with a molecule different from the one used in the docking problem. A modeled structure is the structure developed from a protein sequence based on the structures of homologous proteins. Homologous proteins have a common evolutionary origin, and there are similarities in their protein sequences and three-dimensional structures.

Many algorithms have been developed to address the protein docking problem. These algorithms typically generate a set of candidate solutions. A *scoring function* is used to distinguish the nearly correct solutions from the others. A rigorous docking algorithm would seek all possible bindings between the two molecules. However, this is impractical due to the large size of the search space. A balance should be reached between the computational expense and the amount of the search space examined.

1.3 Paper Overview

After introducing some background knowledge about protein docking in Chapter 2, we shall review existing protein docking algorithms in Section 3.1. Examples of scoring functions used for evaluating the goodness of the candidate solutions produced by the algorithms are discussed in Section 3.2. Based on these discussions, the research topic is proposed in Chapter 4. Furthermore, preliminary work is discussed in Chapter 5, followed by conclusions in Chapter 6.

2 Background

2.1 Structure of Protein

Protein is made from a long chain of amino acids, each links to its neighbor through a *covalent bond* [46]. Covalent bonds are formed as atoms seek to obtain a complete set of electrons in their outer electron shells. There are 20 types of amino acids in proteins, and each amino acid carries different chemical properties. The length of protein is in the range of 20 to more than 5000 amino acids. On the average, a protein contains around 350 amino acids. Therefore, protein is also known as a polypeptide.

Amino acid is the building block of proteins. Each amino acid consists of:

1. Amino Group (-NH2 group),
2. Carboxyl Group (-COOH group), and
3. R Group, which determines the type of the amino acid.

All the groups are attached to a single carbon atom called the α-carbon (Figure 2). The N, Cα, C, O atoms form the backbone of a protein molecule, while the R groups are side-chains (Figure 3).

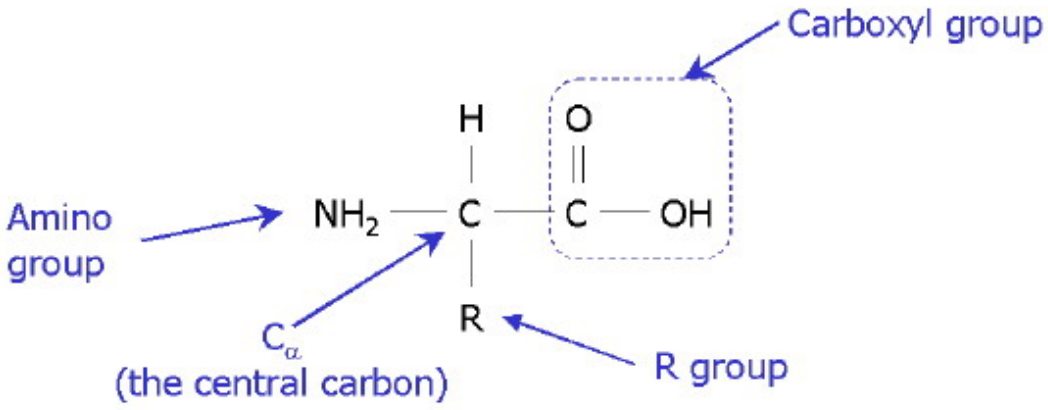

Figure 2: Structure of amino acid.

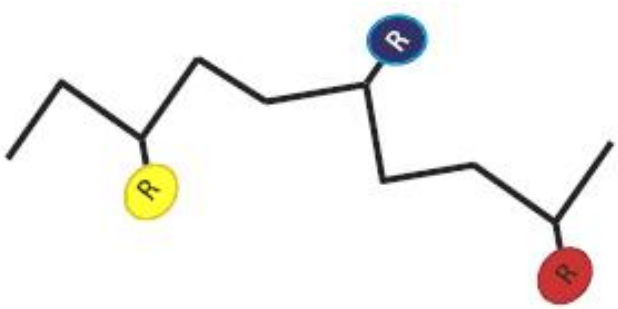

Figure 3: Backbone (black line) and side-chains (R) of protein structure.

A protein naturally folds into certain three-dimensional shape [46] (Figure 4). Folding is a spontaneous process mainly guided by van der Waals forces and entropic contributions to the Gibbs free energy: an increase in entropy is achieved by moving the hydrophobic parts of the protein inwards, and the hydrophilic ones outwards. Hydrophobic parts are electrically neutral and prefer other neutral solvent. Hydrophilic parts are electrically polarized and prefer polarized solvent, e.g., water molecules. Folding also depends on environmental conditions like temperature, solvent, concentration of salts, pH, etc. The duration of the folding process varies dramatically depending on the proteins. The slowest folding requires many minutes or hours. On the other hand, small proteins typically fold on a time scale of milliseconds.

2.2 Conformational Substates

A free protein can exist in a range of *conformational substates* [11]. Different conformational substates of a protein have the same coarse overall structure but differ in detail. In a conformational substate, a side chain may have rotated, some hydrogen bonds may have shifted, or a single helix may be displaced.

The conformational substates correspond to the local minima of the free energy landscape of a protein, where the minima are separated by energy barriers [12]. Under the influence of external factors such as temperature and pressure, one substate can transit to another. Transitions among substates are *protein motions* [22], which are essential for proteins to function.

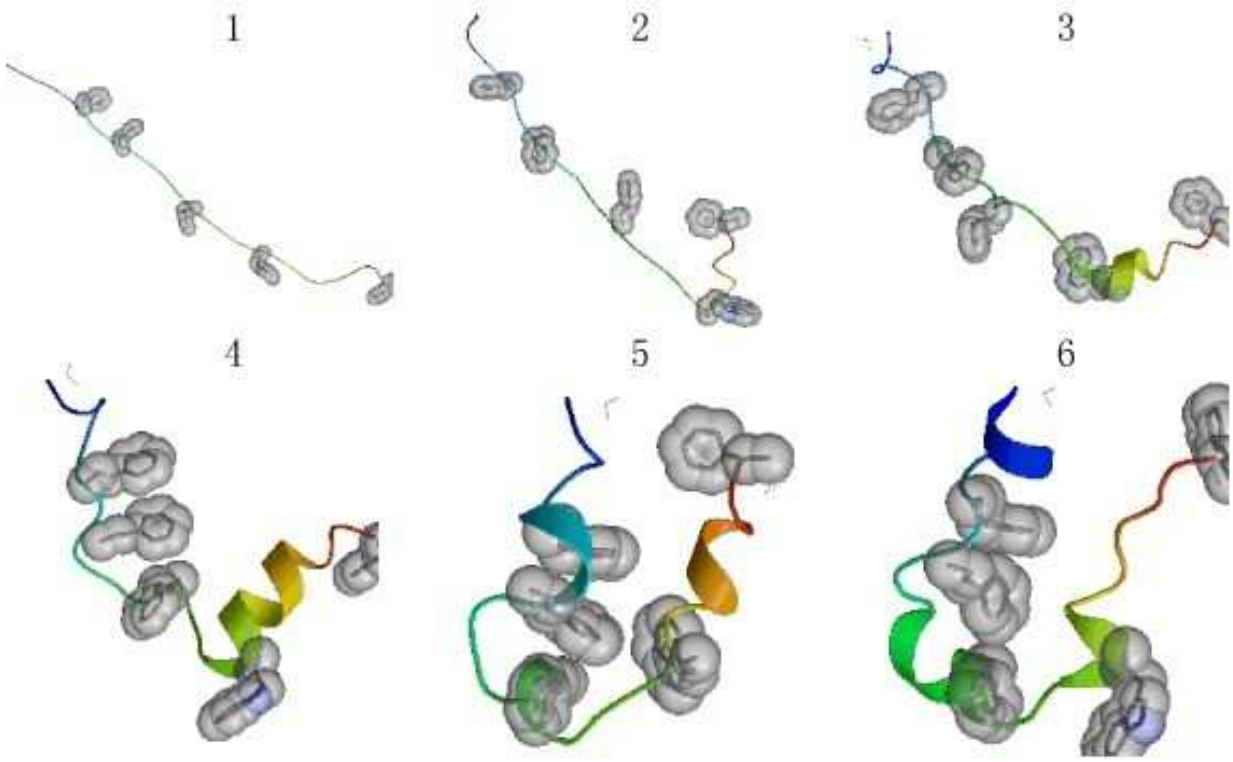

Figure 4: Protein folding. The process of protein folding is depicted from state 1 to 6 (from [32]).

2.3 Protein Binding

Proteins interact with each other and form complexes. There are thousands of complexes recorded in the Protein Data Bank (PDB). Of the two molecules that form a complex, the larger molecule is usually called the *receptor* and smaller one the *ligand*.

During protein binding, the structures of both participants of the interaction are usually altered and changed into other conformational substates (Figure 5). The binding interface regions typically have greater conformational changes. These changes occur for a variety of reasons such as to form chemical interactions, to avoid clashes, and to improve hydrogen bonding [19]. There are conformational changes in non-interface regions too, but they are usually due to flexibility and disorder. Some proteins have flexible and disordered regions existing as dynamic ensembles in which the atom positions and atomic bond angles vary significantly over time with no specific equilibrium values.

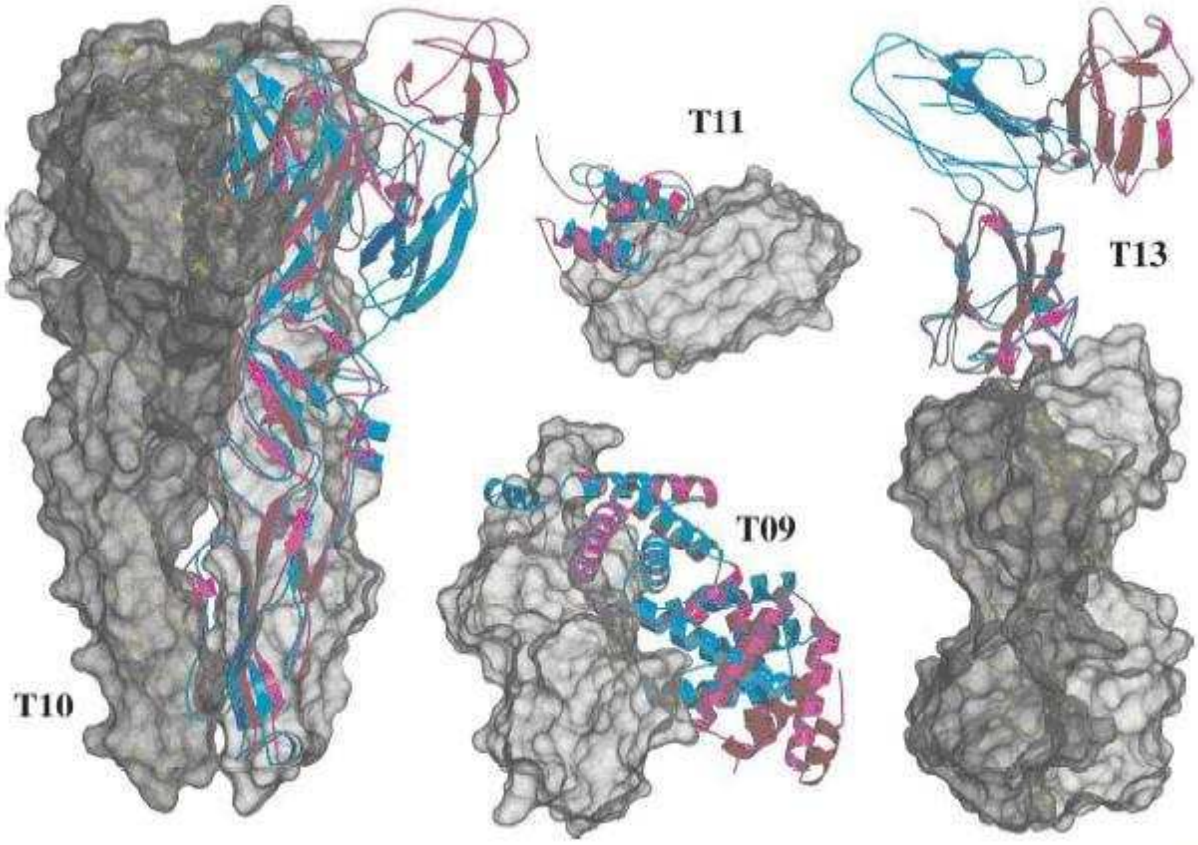

Figure 5: Conformational changes in protein binding. Bound ligands are shown in blue, and unbound ligands in red. The receptors are drawn as a solid surface (from [28]).

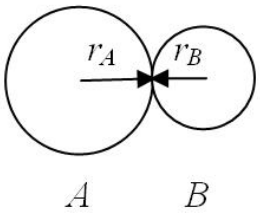

Figure 6: Two unbonded atoms that just touch each other. r_A and r_B are the van der Waals radii of atoms A and B.

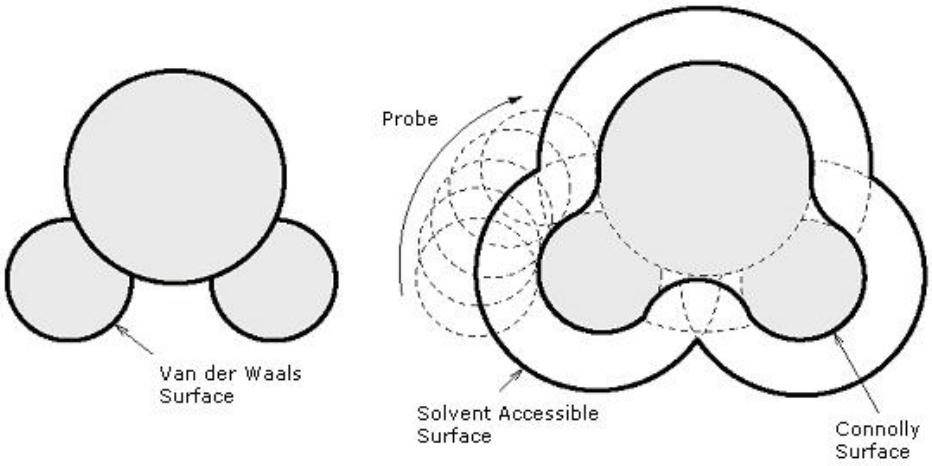

Figure 7: Molecular surface models: Van der Waals surface model, solvent accessible surface model and the Connolly surface model.

2.4 Models of Molecular Surfaces

During protein binding, two molecules interact at their surfaces in contact. There are several ways of representing and modeling molecular surfaces, namely van der Waals surface, solvent accessible surface, and Connolly surface.

When two atoms come in contact, there exists a minimum distance between them due to electrical repulsion. This suggests that atoms must occupy a well-defined molecular volume. The simplest model of an atom is a sphere. The radius of the sphere called van der Waals radius depends on the complexity of the atom, i.e., the number of electrons it contains. When the spheres of two atoms just touch, the measured inter-atomic distance equals the sum of their van der Waals radii (Figure 6). The van der Waals surface of the molecule is the boundary of the union of spheres of the atoms in the molecule [41].

The models of solvent accessible surface and the Connolly surface use a probe to define a protein's surface. The probe is a solvent molecule, usually a water molecule, that rolls over the protein's van der Waals surface. The solvent accessible surface [26] is the trace of the probe's centroid (Figure 7). It models the surface of the molecule which can come into contact with the solvent molecules. The Connolly surface [7] is the boundary of the volume which the probe cannot penetrate (Figure 7). One of the advantages of the Connolly model is the smoothness of the surface.

3 Related Work

3.1 Protein Docking

Protein docking is a computational problem that predicts the binding of two proteins or one protein with another molecule. It can be defined as follows: Given the atomic coordinates of two molecules, predict their correct *bound association* [16], which is given by the relative orientation and position between the molecules after interaction. Protein docking algorithms usually produce a set of candidate solutions instead of a single solution.

Depending on the extent of molecular flexibility taken into account, docking algorithms can be classified to two categories [10]: (1) Rigid-body docking: Both molecules are regarded as rigid solid bodies. No conformational change is performed. (2) Flexible docking: At least one of the molecules, possibly the smaller ligand, is considered flexible. The molecules that are considered flexible may undergo conformational changes.

Protein docking algorithms use search algorithms to find stable configurations between the ligand and the receptor. There are two different searching approaches: (1) full solution space search and (2) gradual guided progression through solution space. The first approach scans the entire solution space in a predefined systematic manner. The second approach searches part of the solution space in a random or criteria-guided manner, or generates the solutions incrementally.

3.1.1 Rigid-body Docking

Most rigid-body docking algorithms perform exhaustive search in a six dimensional space: 3D rotation and 3D translation. The molecule is represented by mapping its surfaces onto a three-dimensional grid. The molecule is represented by a discrete function where the value 1 denotes grid voxels on the molecule's surface, p denotes grid voxels inside the molecule, and the value 0 denotes grid points outside the molecule (Figure 8). The p value for the interior of the molecule is positive for one molecule and negative for the other.

The matching of the surfaces of two interacting molecules is computed by the correlation of their discrete representations. When two molecules have no contact, the correlation value is 0 (Figure 9a). When there is contact, the correlation value is positive (Figure 9b). When there is penetration (Figure 9c), the correlation value is negative. When the geometric match

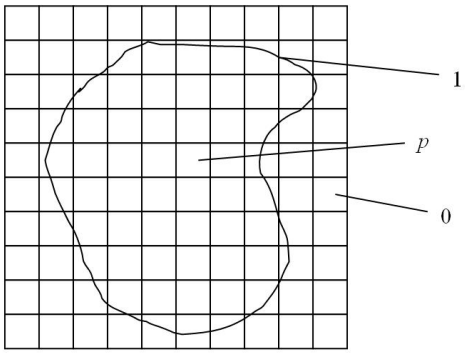

Figure 8: Mapping of the surface of a molecule onto grid.

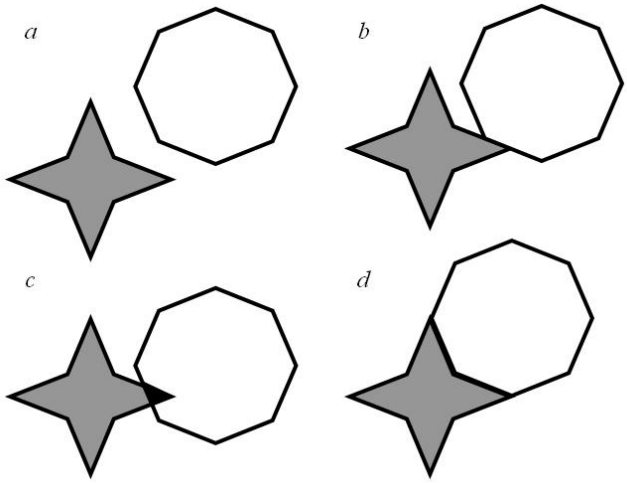

Figure 9: Docking of two molecules. (a) No contact. (b) Limited contact. (c) Penetration. The penetrated part is represented in black. (d) Good geometric match.

is good (Figure 9d), the correlation has a large positive value.

Translational correlation in the spatial domain corresponds to multiplication in the Fourier domain. Therefore, 3D fast Fourier transform (FFT) method is frequently used to compute translational correlation efficiently [5, 21, 42]. On the other hand, 3D rotational match has to be searched exhaustively and thus slowly. This is an exhaustive shape-based algorithm, which works well when both molecules are rigid bodies and shape complementarity is essential. However, the computation cost is high due to the exhaustive rotational search.

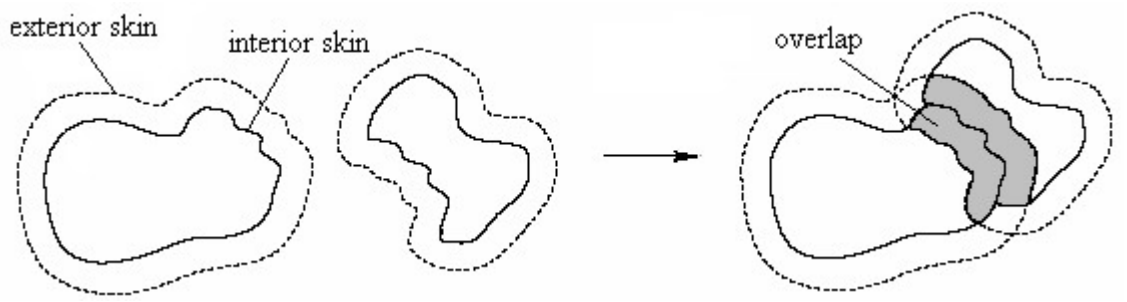

Figure 10: A "double skin" model used in spherical polar Fourier correlation algorithm. The overlap between opposing interior and exterior regions are used to compute the shape complementarity.

Spherical polar Fourier correlation is another approach [30, 35]. The protein surface is represented by a "double skin" model (Figure 10) that describes thin regions of space exterior and interior to the molecular surface. Each skin is represented by a Fourier series expansion of real orthogonal radial and spherical harmonic basis functions.

Shape complementarity is computed by correlating the coefficients of the spherical harmonic functions of opposing pairs of interior and exterior skins. The six-dimentional search space can be represented by 5 rotational angles and an intermolecular separation. Thus, for each separation, the two molecules remaining at fixed positions are rotated about their own centroids and the ligand is twisted about the intermolecular axis. The coefficients of each rotational increment can be computed just once and stored. So, shape complementarity can be computed efficiently using the stored coefficients over each twist angle and intermolecular separation. Similar to the FFT method, spherical polar Fourier correlation algorithm is an exhaustive shape-based search method. The computational cost is high for high resolution search.

Another approach of rigid-body docking is called geometric docking [1, 47]. The surfaces of the molecules are first divided into patches according to rough surface shape, such as concave, convex, or flat surface patches. Complementary patches are identified and superimposed by shape-matching techniques. Concave patches are matched with convex ones and flat patches with any type of patches. The goal is to compute a rigid-body transformation that would align as many complementary patches as possible. One possible technique that can be applied is geometric hashing. Other geometric docking methods use points or spheres

Figure 11: Surface layers of two molecules are allowed to be penetrated to achieve small-scale flexibility.

to represent the geometric elements [23, 24]. Without performing an exhaustive search in the six-dimensional space, geometric docking focuses on aligning complementary features and consumes small computational cost. One drawback is that the algorithm depends greatly on the geometric features used and such features are usually approximated and simplified. So, the results are not accurate.

3.1.2 Flexible Docking

Rigid-body docking algorithms can account for small-scale flexibility by incorporating a surface layer on the molecules and allow for surface penetration (Figure 11). In this case, the penetrations are considered to be equivalent to the conformational changes of the surface.

Monte Carlo algorithm is one of the major algorithms used for flexible docking. Several methods are developed based on Monte Carlo algorithm [4, 9, 15, 43]. The receptor is regarded as a rigid body, and ligand is considered as flexible. The ligand is represented by a set of variables consisting of rotation angles of the whole molecule, translation of the whole molecule, and torsion angles of each atomic bond. In each Monte Carlo cycle, these variables are assigned randomly selected values according to a uniform distribution of allowed values. Then, a cost is computed according to a scoring function (Section 3.2). Those solutions with costs smaller than the lowest cost found up to the current Monte Carlo cycle are saved as candidate solutions.

The main advantage of Monte Carlo algorithm is that its representation of the molecular flexibility is very detailed and can model explicitly all degrees of freedom of the system if necessary. Unfortunately, the high level of details in the modeling comes with a high computational cost.

Genetic algorithm can also be used for flexible protein docking problem [14, 29, 31]. The global translation, global rotation, and torsion angles of internal atomic bonds are encoded as genes for the molecules. A candidate docking result consists of a collection of genes, and it is assigned a fitness value accoring to a scoring function. The number of genes depends on the number of internal bonds. A new population of candidate results is generated from the old one by the use of three genetic operations, namely mutation, crossover and migration. The quality of the solutions usually depends on the starting genes, the number of evolutionary events, and the scoring function that picks the more favorable individuals to generate new population.

In genetic algorithm, a large number of degrees of freedom can be modeled to account for the molecular flexibility. The major drawback of genetic algorithm is that it is too slow for extensive flexible docking of two large molecules.

Another algorithm for flexible docking is an incremental construction algorithm [25, 34]. The ligand is divided into fragments separated by rotatable bonds. The algorithm works by first placing a base fragment into the pre-defined active site of the receptor, followed by a greedy search to incrementally add more fragments and grow the base fragment to the final optimal conformation. The fragments are added to maximize interactions and optimize the scoring function. This process is repeated by placing the base fragment in different orientations at the pre-defined active site.

Using incremental construction algorithm, flexibility is achieved since fragments of ligand are added separately and the speed is satisfactory. However, the result is highly dependent on the selection of an appropriate base fragment and prior knowledge of the binding site. The quality of each construction step and thus the final result is strongly dependent on the previous step.

Domain movement algorithm is another algorithm that simulates protein flexibility [36, 37]. Either the ligand or the receptor is modeled as a hinge-articulated object. Molecules are divided into domains connected by hinges, and the molecules can bend by rotating about the hinge points (Figure 12(a, c)). Movements are allowed either in the ligand or in the

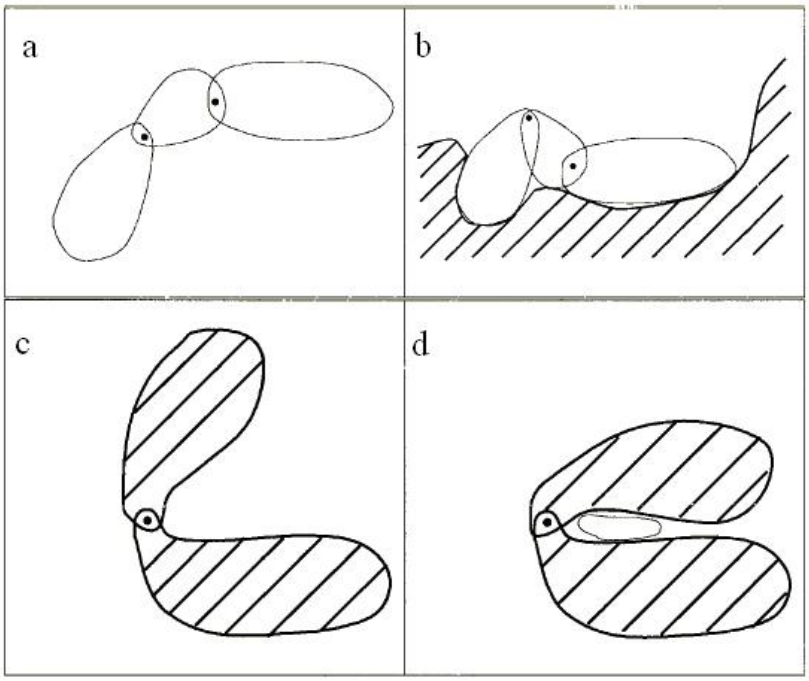

Figure 12: Hinge-bending movements of domains. (a) Hinge-articulated ligand. (b) Ligand bends about the hinge to fit the shape of the receptor (shaded). (c) Hinge-articulated receptor. (d) Receptor closes on the ligand.

receptor. Like pliers closing on a screw, the receptor closes on its ligand and vice versa (Figure 12(b, d)), hence achieving the molecular fit. More than one hinge can be allowed in the docking. By allowing several hinge motions to occur at the same time, the method simulates the cumulative effect of flexibility.

The performance of the domain movement algorithm depends largely on the choice of hinge points. By considering only domains of molecules, backbone flexibility is modeled but the flexibility inside the domain is ignored, which limits the level of flexibility achievable.

A motion planning approach to flexible docking also avoids a full solution space search [39]. By modeling the flexible ligand as an articulated robot, the docking problem is transformed into the robot motion planning problem. Each atomic bond of the ligand molecule maps to a joint of the robot with one degree of freedom: torsion angle. Bond lengths and angles in the plane made by two neighboring bonds are considered constant. Bonds involved in a ring are modeled as a rigid structure. The root atom, which represents the free base of the robot (with 6 degrees of freedom for 3D rotation and translation), is an arbitrarily

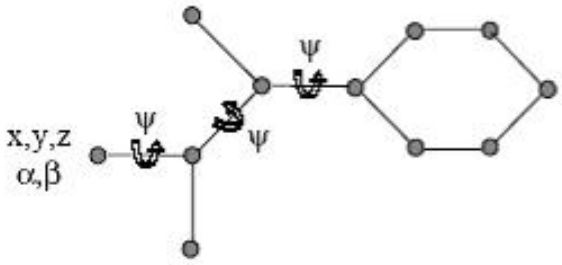

Figure 13: A ligand with 8 degrees of freedom: 3 position coordinates (x, y, z) and 2 rotation angles (α, β) for the root atom, and 3 torsion angles (ψ) for the atomic bonds.

chosen terminal atom of the ligand (Figure 13).

Traditional robot motion planning is based on manipulating a robot through a workspace while avoiding collisions with obstacles. This algorithm applied on protein-ligand docking is to determine potential paths that a robot (ligand) may naturally take based on energy distribution of the workspace. Hence, it examines the possible motions of the robot induced by the energy landscape of its immediate environment. The more energetically favorable paths between the initial and the goal positions of the ligand are computed. The knowledge of binding site is needed and used as the goal position of the ligand. This algorithm is energy-based only and has less concern about the shape of the molecules.

3.1.3 Levels of Flexibility

Studies on protein flexibility show that conformational changes have a crucial influence on protein docking [3, 8]. The flexibility of a molecule can be modeled at two levels, namely backbone level and side-chain level. Treatment of side-chain flexibility alone is not sufficient to achieve satisfactory results for unbound docking. Minor change in the backbone level causes a larger change in the molecule's overall shape compared to side-chain conformational change. However, the majority of the existing docking methods assume that proteins can be consider rigid at the backbone level, while allowing for limited flexibility at the side-chain level. Most of them include a step for optimizing side-chain conformations in the refinement stage.

The conformational space accessible to all side-chains of a protein is very large. A key

approximation which alleviates this problem is the discretization of the side-chain conformational space, whereby a side-chain is only allowed to adopt a discrete set of conformations [20, 38]. This approximation is based on the observation that, in high-resolution experimental protein structures, side-chains tend to cluster around a discrete set of favored conformations, known as *rotamers*. In most cases, these rotamers correspond to local minima of potential energy on the side-chain. Many rotamer libraries are presently available. A rotamer library can be added into some algorithms mentioned above, such as Monte Carlo algorithm [43] and FFT algorithm [18]. It reduces the search space on a large scale and thus allows for fast searching.

Principal component analysis (PCA) can also be applied to handle side-chain flexibility [30]. First, a large number of 3D protein conformations are generated randomly with a set of upper and lower interatomic distance limits. Each conformational structure is considered as a sample point within the multidimensional conformational space of the protein. Then the eigenvectors and eigenvalues of the set of samples are computed. The eigenvectors are orthonormal and span the conformational coordinate space of the protein. Thus any 3D conformation may, in principle, be constructed from an appropriate linear combination of the eigenvectors. Up to 10^5 conformations can be generated but many of them have infeasible covalent bond lengths and angles. A simple bond length filter can be used to prune the unrealistic conformations.

Treatment of backbone flexibility remains a major challenge. Some algorithms [9, 27] include a global refinement step that performs energy minimization before applying the scoring function. The global refinement enables small adjustment of the backbone. Another way of handling backbone flexibility is to use PCA technique [30] in the similar manner as handling side-chain flexibility. Algorithms such as incremental construction algorithm and domain movement algorithm account to backbone flexibility as discussed in the previous section.

3.2 Scoring Functions

Protein docking algorithms typically produce a set of candidate solutions. A scoring function is used to assess the goodness of the candidate solutions. It may contain multiple aspects, such as shape, intermolecular overlap, intramolecular overlap, hydrogen bonds, electrostatic

potential, van der Waals potential and other energy models. Scoring functions may be applied after the searching stage, or used during searching procedure to prune the candidate solutions. The latter approach is required for methods such as Monte Carlo algorithm and genetic algorithm.

3.2.1 Shape Measure

- Geometric Complementarity

  Geometric complementarity is the measure of how the 3D shapes of two molecules match each other at the contacting interface. It plays an important role in protein docking since most protein-protein interactions actually shows a good geometric complementarity [21, 40].

  There are several definitions available for geometric complementarity [16]. One definition is based on surface normals of contact area between two molecules. If the surface normals are in opposite directions, the corresponding surfaces are complementary. The other definition is based on the contacting area of two molecules. A large contacting area suggests better complementarity. Algorithms such as FFT method can directly compute the amount of contacting area.

- Intermolecular Overlap

  Intermolecular overlap is the overlap between two interacting molecules. By allowing some intermolecular overlap, some amount of conformational flexibility is taken into account.

  The general approach to intermolecular overlaps is to allow for a small amount of interface penetration and to penalise large interior clashes. The tolerance is usually implemented by a surface layer of non-penalized penetration area. As shown in Figure 11, a surface layer is used at the surface of the molecules and overlaps of surface layers are allowed.

- Intramolecular Overlap

  When ligand or receptor flexibility is taken into account, the overlaps inside a molecule may occur. For example, when a ligand is divided into fragments in incremetal algorithm for flexible docking, fragments may clash with each other during docking.

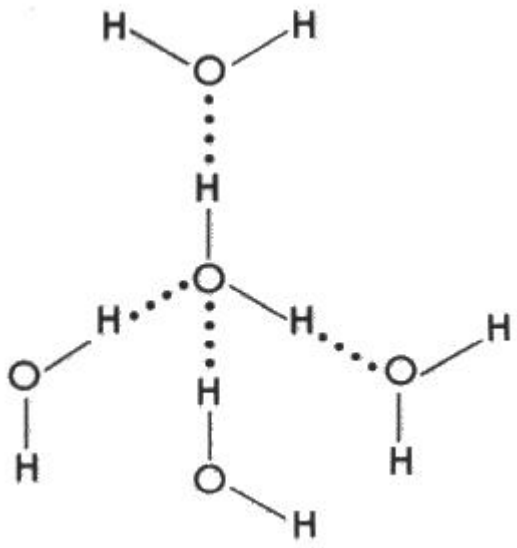

Figure 14: Hydrogen bonds (dotted lines) among five water molecules.

Though slight overlap may be considered as flexibility, large-scale self-collision is not present in real protein interactions. Usually, self penetration is penalised.

3.2.2 Hydrogenn Bonds

Polar molecules, such as water molecules, have a weak, partial negative charge at one region of the molecule (the oxygen atom in a water molecule) and a partial positive charge elsewhere (the hydrogen atoms in a water molecule). Thus, when water molecules are close together, their positive and negative regions are attracted to the oppositely-charged regions of nearby molecules. The force of attraction, shown as a dotted line in Figure 14, is called a hydrogen bond.

There tends to be uniformity in the static features of the complex interface despite a variety of shapes. The interface between two molecules of a complex has 1.13±0.47 hydrogen bonds per 100 Å^2 buried accessible surface area [3]. Å stands for Angstrom and 1 Å $= 10^{-10}$ meter. The buried accessible surface area is the surface buried and not accessible by solvent when two proteins form a complex. Thus the number of hydrogen bonds on the interface of two interacting molecules is another important measurement of interaction.

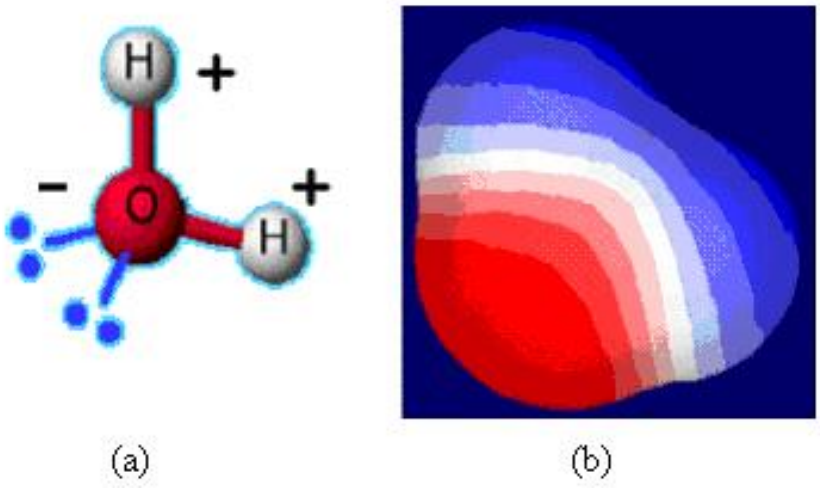

Figure 15: Electrostatic potential on the surface of a water molecule. (a) The partial charges of the atoms of a water molecule. (b) The electrostatic potentials on the surface of a water molecule. Negative potentials are colored as red and positive potentials are colored as blue.

3.2.3 Electrostatic Potential

The common definition of electrostatic potential is the potential energy of a proton at a particular location near a molecule. Negative potential corresponds to attraction of the proton by the concentrated electron density, and positive potential corresponds to repulsion of the proton by the atomic nuclei in regions where the positive nuclear charge is not completely shielded by low electron density.

Water molecule is a good example for understanding electrostatic potential (Figure 15). The bond between H and O are formed by two electrons, one from H and the other from O. However, the two electrons are closer to the oxygen's atomic nucleus than those of the hydrogen. As a result, O and H have, respectively, negative and positive partial charges, and thus there are corresponding negative (red) and positive (blue) electrostatic potentials on the molecular surface.

Electrostatic interactions play an important role in energy evaluation for scoring candidate solutions. When two molecules are interacting with each other, the existence of complementary charged surfaces is a good indicator of a good docking interface.

The classical treatment of electrostatic interactions in solution is based on the Poisson-

Boltzmann equation (PBE) [17]:

$$\nabla\left[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})\right] - \epsilon(\mathbf{r})\kappa^2(\mathbf{r})\sinh\left[\phi(\mathbf{r})\right] + \frac{4\pi\rho(\mathbf{r})}{k_B T} = 0 \tag{1}$$

where $\mathbf{r}$ is a 3D position, $\phi(\mathbf{r})$ is the electric potential at $\mathbf{r}$, $\epsilon(\mathbf{r})$ is the dielectric constant at $\mathbf{r}$, and $\rho(\mathbf{r})$ is the charge density. The term $\kappa^2(\mathbf{r}) = 8\pi q^2(\mathbf{r})I/\epsilon(\mathbf{r})k_B T$, where $q(\mathbf{r})$ is the charge of a proton at $\mathbf{r}$, T is the absolute temperature, k_B is the Boltzmann constant, and I is the ionic strength of the bulk solution. Analytical solutions to the PBE are only available for a limited number of cases involving idealized geometries such as spheres and cylinders. Numerical methods for computing electrostatic potentials can be categorized into two approaches [17]: finite difference method (FDM) and boundary element method (BEM).

3.2.4 Van de Waals Potential

When two non-bonded atoms are at close proximity, van der Waals attraction occurs. When their distance is less than the sum of their van der Waals radii, van der Waals repulsion occurs. Theoretically, van der Waals interaction should be minimum when two molecules are at the equilibrium separation (Figure 16).

The van der Waals potential $V(r)$ between two non-bonded atoms can be expressed as a function of their seperation r as follows:

$$V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2}$$

where r is the distance between two atoms, ϵ is the well depth at the equilibrium separation r_{eqm}, and σ is the collision constant such that $2^{1/6}\sigma = r_{eqm}$. At $r = r_{eqm}$, $V(r) = -\epsilon$.

3.2.5 Other Energy Terms

In addition to the measures described in the precedubg sections, other energy terms have also been used, such as bond potential, bond angle potential, torsion angle potential, hydrophobicity, etc. Please refer to [6] and [33] for more details.

3.3 Summary & Comparison

Protein docking is a difficult computational problem. In the past decade, many methods have been proposed, and significant progress has been made. However, this problem is far from being solved.

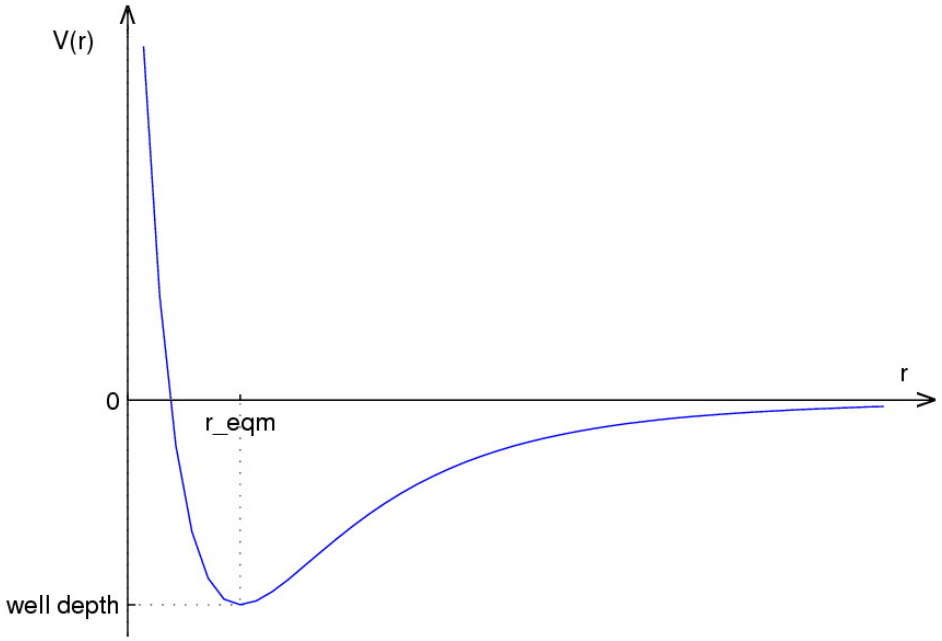

Figure 16: Van der Waals potential $V(r)$ between two non-bonded atoms. r is the distance between two atoms, r_{eqm} is the equilibrium separation, well depth is the potential at r_{eqm}.

Bound docking is a simple version of the protein docking problems. A rigid-body docking algorithm is able to find good solutions for bound docking. FFT algorithm, spherical polar Fourier correlation algorithm and geometric docking algorithm are examples of rigid-body docking methods (Table 1). FFT and spherical polar Fourier correlation algorithm are exhaustive methods which will find the best shape match given enough time. They can also incorporate a surface layer to allow penetration and overlap, and thus account for a small amount of flexibility. Geometric docking algorithm uses molecular surface features to find the best shape match. This non-exhaustive algorithm can be fast and efficient, but it may not be accurate enough due to the approximation of features. All three algorithms take shape complementarity as their primary objective, which is sufficient in the case of bound docking. However, rigid-body docking algorithms do not always give good solutions to unbound docking problem because proteins usually undergo conformational changes when they are bound to their partners.

Unbound docking problems are more difficult as the search space increases dramatically. One approach to unbound docking is to add some amount of flexibility into rigid-body docking algorithms, such as allowing penetration on the molecular surfaces in FFT method.

The results may be acceptable for a small set of test cases. However, this approach is not reliable for general flexible docking in practice [21].

Flexible docking algorithms are developed to address unbound docking problems. No existing flexible docking algorithm attempts an exhaustive search (Table 1) due to the large search space. In order to reduce the search space, random sampling (Monte Carlo, genetic algorithm) or criteria guided construction (incremental construction algorithm) are used. Algorithms that perform molecular articulation are feasible only for small number of hinges (domain movement algorithm) or small molecules (motion planning algorithm).

Different levels of flexibility are modeled in existing flexible docking algorithms (Table 1). Monte Carlo and gentic algorithm can account for both side-chain and backbone flexibility because they can potentially generate possible conformations at both levels. However, backbones are usually held rigid or limited to a small number of degrees of freedom in the existing applications to reduce search space. Incremental construction and motion planning algorithm can account for both side-chain and backbone flexibility only for small molecules, but knowledge of possible binding site is required. Domain movement algorithm accounts for coarse backbone flexibilty with a small number of hinges. Generally speaking, treatment of backbone flexibility is still a major challenge.

Another important aspect of protein docking problem is scoring function. Usually, flexible docking algorithms use scoring functions during the search to prune away poor candidate results or to guide the search, and rigid-body docking algorithms use them after the search to select good solutions. Shape complementarity is a primary criterion used in rigid-body docking, but it is only a part of the scoring function for most flexible docking methods (Table 1). Though many existing scoring functions can rank the correct solutions among the top hundred or even top ten possible solutions, many solutions with good scores are false positives. A selective and effective scoring function should give good scores to correct and nearly correct solutions while give poor scores to others.

Table 1: Comparison of existing docking methods.

Docking method	Docking type	Search strategy	Knowledge of binding site	Flexibility treatment	Flexibility level	Use of scoring function	Scoring function
FFT	Rigid-body	Exhaustive	Not required	Surface layer	Side-chain	After search	Shape complementarity
Spherical polar Fourier correlation	Rigid-body	Exhaustive	Not required	Surface layer	Side-chain	After search	Shape complementarity
Geometric docking	Rigid-body	Geometric hashing	Not required	N.A.	N.A.	After search	Shape complementarity
Monte Carlo	Flexible	Local optimization	Not required	Generate possible conformations	Side-chain, backbone	During search	Various
Genetic algorithm	Flexible	Evolution	Not required	Generate possible conformations	Side-chain, backbone	During search	Various
Incremental construction	Flexible	Greedy search	Required	Ligand fragments	Side-chain, backbone	During search	Various
Domain movement	Flexible	Geometric hashing	Not required	Articulation	Coarse backbone	After search	Shape complementarity
Motion planning	Flexible	Motion planning	Required	Articulation	Full (small molecule)	Before search (generate configurations)	Energy

4 Research Proposal

4.1 Problem Statement

Protein interactions play important roles in various aspects of the molecular mechanisms of biological processes. Protein docking techniques are used in pharmaceutical applications such as the modulation of protein interactions with small molecules in drug design. In practice, finding the binding of two flexible molecules is more important than reconstructing the complex of two bound molecules. Thus, unbound docking is the major problem in protein docking. The difficulty of this problem is due to the conformational changes that occur during the binding process, which increase the complexity of the problem by a large scale.

Existing flexible docking algorithms are developed to solve the unbound docking problem (Chapter 3). Most of them model molecular flexibility by perturbing a set of parameters coresponding to the degrees of freedom of the molecules. The idea is to generate a large set of possible conformations and then select the good ones. In spite of these efforts, treatment of backbone flexibility is still a major challenge (Section 3.3).

We propose to model the conformational changes that occur during the protein binding process. Instead of seeking the correct solution from a large set of possible conformations, we model how and why the conformational changes occur, guided by forces acting on the atoms. The model can be used as part of a flexible docking algorithm to predict and refine the conformational changes given candidate docking results.

We know that in physics, conformational changes occur because of electrical attraction and repulsion, changes in atomic bonding, solvent, temperature, etc. It is computationally expensive to model all the relevant factors. So, our goal is to find a guided way to model conformational changes that is computationally efficient.

There are two research problems involved. Problem 1 is to determine an effective fitness function that gives high fitness value to good candidate docking results and low fitness value to bad ones. Problem 2 is to determine the conformational changes of the ligand given a candidate docking result. The fitness function obtained in Problem 1 is used in Problem 2. To describe these problems more precisely, we formulate the problems as follows.

Consider two proteins, a rigid receptor and a flexible ligand, that interact during binding.

Let $R = \{\mathbf{r}_j\}$ represent the set of 3D coordinates of the atoms in the receptor. Let $U = \{\mathbf{u}_i\}$ represent the set of 3D coordinates of the atoms in the unbound ligand before binding, and $B = \{\mathbf{b}_i\}$ the set of 3D coordinates of the atoms in the bound ligand after binding to the receptor. Note that R and B are given in the same 3D coordinate system. So, information about the binding site can be derived from the atoms that are in contact. Let T denote the global rotation and translation of the ligand, and C denote the conformational changes of the ligand. Then,

$$\mathbf{b}_i = C(T(\mathbf{u}_i)),\ \forall\ i. \tag{3}$$

Let $F(R, B)$ denote a fitness function that computes the fitness of a possible complex of receptor and ligand.

- Problem 1: Determine an effective fitness function.

  Given R_k, U_k, B_k, find $F(R_k, B_k)$ such that $F(R_k, B_k) - F(R_k, C(T(U_k)))$ is a monotonic function of the alignment error $\|B_k - C(T(U_k))\|$ between B_k and $C(T(U_k))$, where

  $$\|B_k - C(T(U_k))\| = \frac{1}{|B_k|} \sum_i \|\mathbf{b}_{ki} - C(T(\mathbf{u}_{ki}))\|^2 . \tag{4}$$

  For Problem 1, an effective fitness function can be found by using a set of training samples including known receptors with bound and unbound ligands. Since there is a infinite number of C and T, it is impossible to verify all the possibilities. Thus, the evaluation of the fitness function can be performed by sampling C and T and measuring the correlation between $F(R_k, B_k) - F(R_k, C(T(U_k)))$ and $\|B_k - C(T(U_k))\|$ and checking whether F is monotonic.

- Problem 2: Determine the conformational changes.

  Let T_0 denote the initial guess of the global rotation and translation of the ligand. Given R, U, F, T_0, find C, T that maximize $F(R, C(T(U)))$.

  For Problem 2, a set of testing samples with known receptor R, bound ligand B and unbound ligand U can be used to evaluate C and T found by the algorithm by measuring the error E_C given by

  $$E_C = \|B - C(T(U))\| . \tag{5}$$

4.2 Research Plan

Given the research problems, we present a research plan as follows:

1. Implement an FFT docking program.

   The program can be used to generate initial guesses T_0 of the ligand's position and orientation, which are the candidate docking results.

2. Determine an effective fitness function F.

   The fitness function may consist of many factors including shape complementarity, van der Waals potential, electrostatic complementarity, etc. The various factors can be combined using a weighted sum, but other combinations are possible.

3. Develop an algorithm to refine T given an initial guess T_0.

   Regard the ligand as rigid, and refine the global rotation and translation to maximize the fitness value given by function F. The refined T may be obtained by identifying the forces acting on the ligand as a whole.

4. Develop an algorithm to find C, T given an initial guess T_0.

   Regard the ligand as flexible, and determine the conformational changes and the best global rotation and translation of the ligand that maximize the fitness value given by F. The conformational changes may be modeled by identifying flexible bonds and the forces acting on each atom.

5 Preliminary Work

Three sets of preliminary work have been accomplished:

1. implementation of FFT docking algorithm,
2. study of van der Waals potential and electrostatic complementarity, and
3. development of robust registration algorithm for the study of conformational changes.

5.1 FFT Docking

An FFT docking program is implemented which can produce candidate docking results for bound docking problem. It can also provide initial guesses of the ligand's rotation and translation for unbound docking problem.

FFT rigid-body docking is an effective docking method for the bound docking problem. Our project team member implemented an FFT docking program based on Katchalski-Katzir and Gabb's work [13, 21], but used a different atom model [45] (Figure 17). Katchalski-Katzir and Gabb's atom models both use thick surface layers and they allow large overlap of the surface layer and the atom core. The new model has a thinner surface layer and the surface core boundary is inside the van der Waals surface. Using the new model effectively avoids deep penetration among atoms and improves the performance of the FFT docking program. For details of the algorithms, please refer to Section 3.1.1 and [45].

The program takes bound cases as inputs and try to reconstruct the complexes. 60 candidate solutions are produced for each test case and ranked according to their FFT score. To evaluate the performance of FFT docking algorithms, the root-mean-square deviation (RMSD) between the docking result and the actual ligand in the complex (the ground truth) is computed.

$$RMSD(B, T(U)) = \left[\frac{1}{|B|} \sum_i \|\mathbf{b}_i - T(\mathbf{u}_i)\|^2 \right]^{1/2} \tag{6}$$

Table 2 lists the ranking and RMSD of the best candidate solution produced by the program using various atom models. The best candidate solution is the one with the smallest RMSD with respect to the ground truth. Using the new atom model, the program produces better results compared to using the existing atom models. For the new model, 15 of the

Table 2: Comparison of docking results using three different atom models. The results listed are the RMSD and FFT rank of the best candidate solution of each test case.

Test Case	Katchalski-Katzir's Model		Gabb's Model		New Model	
	RMSD (Å)	FFT Rank	RMSD (Å)	FFT Rank	RMSD (Å)	FFT Rank
1ABI	7.152	14	4.885	45	0.243	2
1ACB	6.894	47	0.582	5	0.476	9
1CHO	15.257	22	12.12	20	0.59	4
1CSE	6.053	41	6.835	1	0.583	5
1FDL	10.44	32	16.527	44	14.823	27
1TGS	5.675	31	0.405	1	0.239	5
2KAI	5.998	12	5.821	51	0.637	10
2MHB	8.163	21	0.752	1	1.052	2
2PTC	5.648	11	2.589	20	0.511	4
2SIC	7.042	10	11.338	18	0.857	6
3HFM	8.118	39	9.397	41	0.552	4
4HVP	14.087	1	9.554	46	0.688	1
4SGB	5.907	13	0.552	1	0.44	1
4TPI	5.233	13	2.404	6	1.149	3
9LDT	25.954	55	17.23	26	0.898	1
9RSA	12.888	43	2.989	22	0.984	1

16 test cases have best candidate solutions ranked within the top 10 according to the FFT score.

5.2 Evaluation of Van der Waals Potential and Electrostatic Complementarity

In order to find an effective fitness function F, two possible factors that can be used in F are evaluated: van der Waals potential and electrostatic complementarity. The methods of calculating these two factors are implemented and tested on candidate docking results produced by the FFT docking program described in the previous section. Various tests are performed to study their properties.

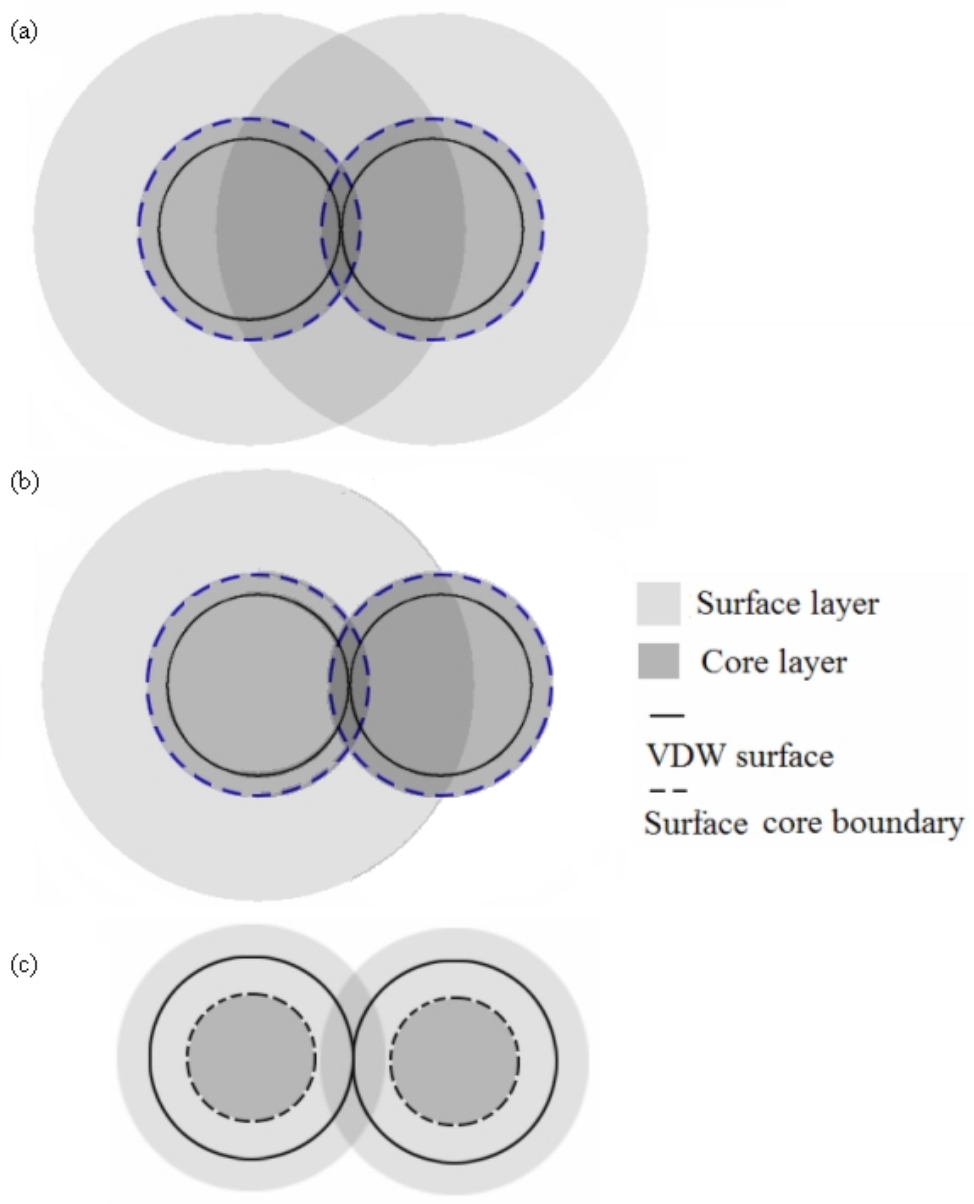

Figure 17: Various atom models used in FFT docking program. (a) The model used in Katchalski-Katzir's work [21]. (b) The model used in Gabb's work [13]. (c) The new model.

Table 3: Equilibrium separation r_{eqm} (Å) and well depth ϵ (kcal mol^{-1}) used in computing van der Waals potential.

r_{eqm} , ϵ	C	N	O	S	H
C	4.00 , 0.150	3.75 , 0.155	3.60 , 0.173	4.00 , 0.173	3.00 , 0.055
N	3.75 , 0.155	3.50 , 0.160	3.35 , 0.179	3.75 , 0.179	2.75 , 0.057
O	3.60 , 0.173	3.35 , 0.179	3.20 , 0.200	3.60 , 0.200	2.60 , 0.063
S	4.00 , 0.173	3.75 , 0.179	3.60 , 0.200	4.00 , 0.200	3.00 , 0.063
H	3.00 , 0.055	2.75 , 0.057	2.60 , 0.063	3.00 , 0.063	2.00 , 0.020

5.2.1 Van der Waals Potential

Van der Waals potential is an important aspect in evaluating the goodness of the possible solutions to protein docking. It is computed according to Equation 2 (Section 3.2.4), summed over all pairs of atoms at the receptor-ligand interface, each consists of one atom from the receptor and one atom from the ligand. The parameters used in the computation are listed in Table 3 [29]. A negative potential corresponds to attraction between receptor and ligand, and a positive potential corresponds to repulsion.

Van der Waals potentials of the candidate docking results produced by the FFT program are computed. Table 4 lists the van der Waals potential and the corresponding rankings of the best candidate solutions. Out of 16 test cases, there are 9 cases with their best candidate solutions ranked within the top 10 according to van der Waals potentials. It shows that vad der Waals potential can be used in the fitness function to select good candidate solutions, but it is not as effective as the FFT score in bound docking.

The potentials of the bound complexes are also listed in Table 4 as a reference. For the best candidate solution, almost all of their van der Waals potentials are higher than those of the actual bound complexes (the ground truth). There are two test cases, 1ABI and 1CHO, with positive van der Waals potentials for the ground truth. This is due to a single pair of sulfur atoms that are closer to each other than predicted by the theory. Removing this pair of atoms in the computation will lower the van der Waals potentials to negative values. This situation may be caused by errors in measuring the atom coordinates using x-ray diffraction or errors in the parameters used in the computation (Table 3).

Table 4: Candidate results ranked using van der Waals potential. The data listed are the ranks and van der Waals potentials of the best candidate solutions, and the van der Waals potentials of the bound complexes (the ground truth).

Test Case	Best Candidate Solution			Bound Complex
	FFT Rank	VDW Rank	VDW Potential	VDW Potential
1ABI	2	26	826.69	533.616
1ACB	9	23	721.98	-73.37
1CHO	4	4	185.63	480.005
1CSE	5	1	-61.21	-104.529
1FDL	27	21	2057.51	-83.901
1TGS	5	4	-26.02	-125.114
2KAI	10	2	-8.35	-101.7
2MHB	2	12	764.19	-109.168
2PTC	4	3	70.87	-106.137
2SIC	6	6	224.12	-111.557
3HFM	4	3	186.65	-91.582
4HVP	1	1	-104.38	-213.975
4SGB	1	13	342.34	-93.457
4TPI	3	3	70.93	-117.802
9LDT	1	11	1246.71	-336.207
9RSA	1	15	681.08	-78.321

Because FFT docking method employs a surface layer on the atom model, which allows for penetration, the candidate solutions produced may have intermolecular overlaps and result in higher van der Waals potential (repulsion). This shows that the candidate solutions are less stable than the bound complexes.

Candidate docking results are not perfect solutions, but can be very close to the ground truth. As the bound complexes of the test cases are known, candidate results can be further transformed to the correct position and orientation. Such transformations can be divided into small steps to study the changes of van der Waals potential as the candidates get closer to the ground truth.

Figure 18 shows the RMSD of the candidates at each steps of the transformation. It is trivial that the RMSD decreases as the candidates get closer to the ground truth. Figure 19

shows the van der Waals potentials of the candidates at each steps of the transformation. In most cases, the van der Waals potentials also decrease as the candidates get closer to the ground truth. This property can be useful in the refinement of the transform matrix given an initial guess. The refinement can be guided to seek a minimum van der Waals potential which corresponds to the correct solution.

There are three cases that do not have a decreasing van der Waals potential. For cases 1ABI and 1CHO, the bound complexes have high van der Waals potentials as explained in the previous paragraph. For case 1FDL, the best candidate result produced by FFT docking program has RMSD = 14.82Å, which is a poor result. The candidate needs to be rotated and translated by a large magnitude to become the correct solution, and thus causes the fluctuation of its van der Waals potentials as it gets closer to the ground truth.

5.2.2 Electrostatic Complementarity

Electrostatic complementarity is another important aspect in evaluating the fitness of the possible docking results. The electrostatic complementarity score is computed as the product of the electric charge of the ligand and the electrostatic potential of the receptor, summed over the whole system. The receptor is the source of the potential field and the ligand is a collection of partial charges centered at its atomic coordinates. Let $V(\mathbf{x})$ be the electrostatic potential of the receptor at point $\mathbf{x}$ and $Q(\mathbf{x})$ be the charge of the ligand at point $\mathbf{x}$, then the electrostatic complementarity score E_{elec} is given by

$$E_{elec} = \int V(\mathbf{x})Q(\mathbf{x})d\mathbf{x}. \tag{7}$$

The partial charges for both receptor and ligand are assigned according to an AMBER parameter set [44]. The electrostatic potential of the receptor is generated by solving the linearized Poisson-Boltzmann equation with the program Adaptive Poisson-Boltzmann Solver (APBS) [2]. The potential is evaluated on a $128 \times 128 \times 128$ discrete grid with 0.78Å spacing, a solvent dielectric of 78.54, a protein dielectric of 1.0, a temperature of 298.15K and a solvent radius of 1.4Å. The partial charges of the ligand are mapped to the same grid. A negative value of E_{elec} reflects the complementarity of electrostatic potentials and thus the solution with negative value of E_{elec} is better than the one with positive value.

The candidate docking results produced by the FFT program are used to evaluate E_{elec}. Table 5 lists electrostatic complementarity scores and the corresponding rankings of the

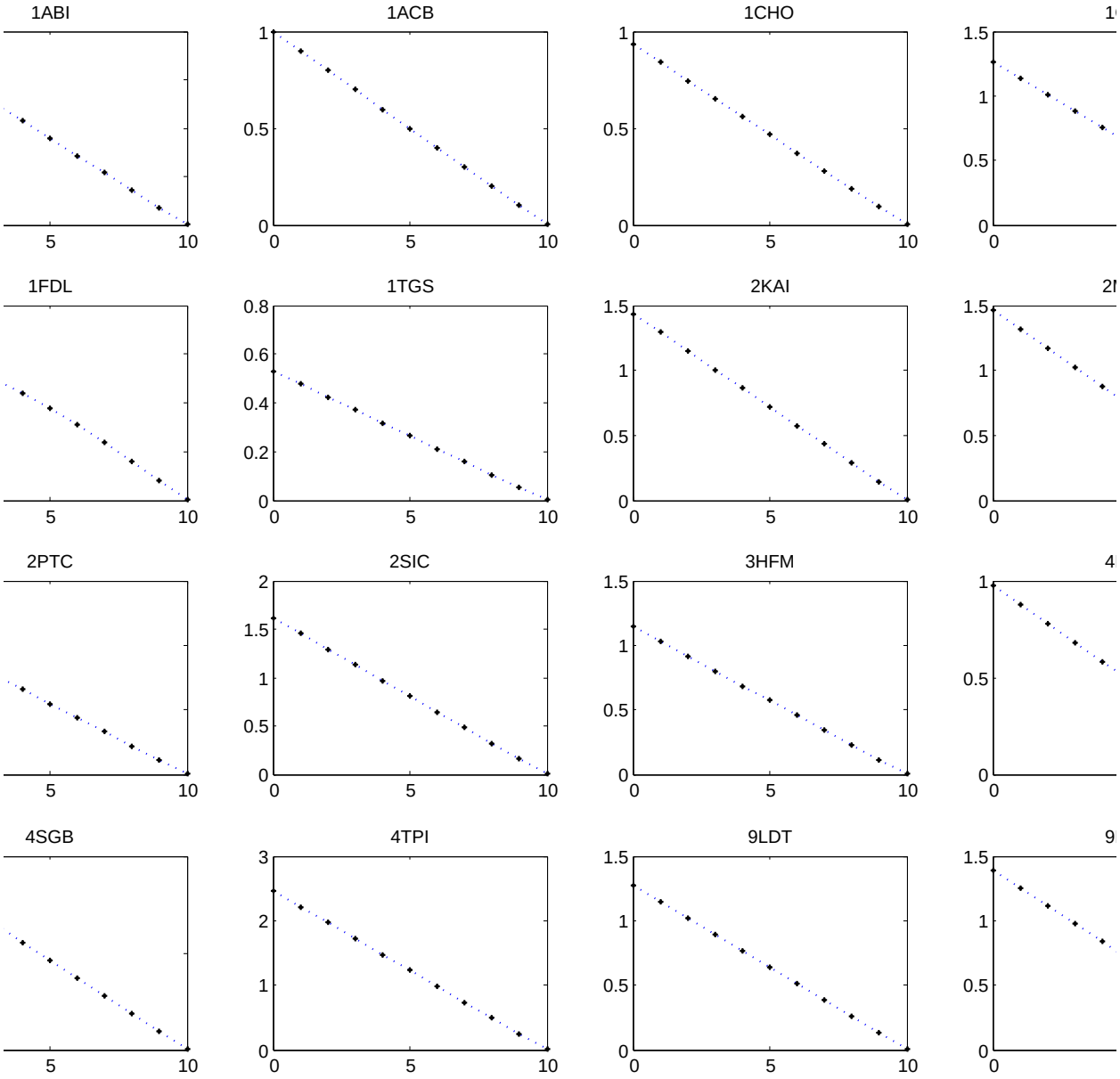

Figure 18: RMSD of the candidate results at each step of the transformation to the ground truth. 10 steps of transformation are computed. The ground truth has RMSD = 0.

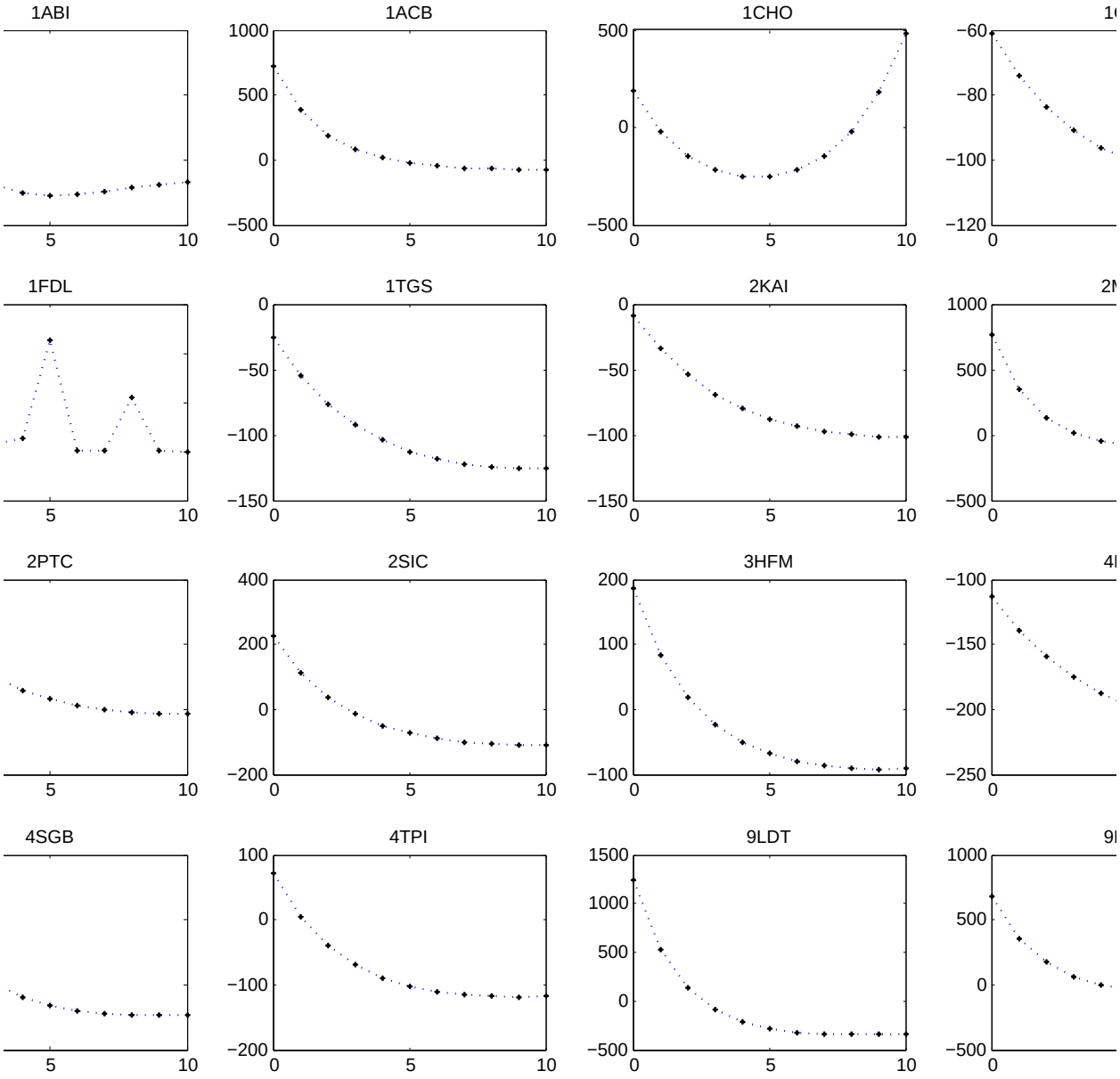

Figure 19: Van der Waals potentials of the candidate results at each step of the transformation to the ground truth. 10 steps of transformation are tested. In theory, the ground truth has a minimum van der Waals potential.

Table 5: Candidate results ranked using electrostatic complementarity. The data listed are the ranks and electrostatic complementarity of the best candidate solutions, and the electrostatic complementarity of the bound complexes (the ground truth).

Test Case	Best Candidate Solution				Bound Complex
	FFT Rank	VDW Rank	Elec Rank	Elec Complm	Elec Complm
1ABI	2	26	7	-64.82	-66.74
1ACB	9	23	1	-130.59	-15.41
1CHO	4	4	6	-112.66	-24.5
1CSE	5	1	8	-21.23	-4.32
1FDL	27	21	13	-8.06	-5.32
1TGS	5	4	9	-44.92	-20.02
2KAI	10	2	7	-56.1	-19.5
2MHB	2	12	12	-18.7	-4.99
2PTC	4	3	2	-98.15	13.9
2SIC	6	6	2	-88.11	-4.22
3HFM	4	3	35	-0.4	1.07
4HVP	1	1	11	-4.72	-6.71
4SGB	1	13	11	-15.54	-2.76
4TPI	3	3	3	-80.16	-12.76
9LDT	1	11	2	-68.91	-33.02
9RSA	1	15	3	-45.78	-46.52

best candidate solutions. Out of 16 test cases, there are 11 cases with their best candidate solutions ranked within the top 10 according to E_{elec}. It shows that electrostatic complementarity can be used to evaluate the fitness of candidate results. However, similar to the van der Waals potential, it is not as effective as the FFT score in bound docking.

Table 5 also shows the electrostatic complementarity of the bound complexes. In theory, E_{elec} of the ground truth should be smaller than those of the candidate solutions. However, our test results show that almost all the best candidate solutions have E_{elec} smaller than those of the ground truth. Many factors contribute to this result, for example, the use of discrete grids in the computation, and the intermolecular overlaps in the candidate solutions produced by the FFT docking program. The overlaps cause the charges of the ligand to be inside the boundary of the receptor and result in very small E_{elec}.

Since the intermolecular overlaps affect the evaluation of E_{elec}, three different ways of handling overlaps in the computation were tested:

1. Exclude the overlaps in the computation. The ligand atoms that overlap the receptor are discarded.

2. Displace the ligand's overlapping atoms to randomly selected empty neighboring grid points.

3. Displace the ligand's overlapping atoms to empty neighbor grid points in the opposite direction of penetration. Let $\mathbf{x}$ be the coordinates of a ligand atom that is overlapping the receptor. Let $\{\mathbf{y}_i\}$ be the set of coordinates of receptor atoms that are overlapping the ligand atom at $\mathbf{x}$. The direction $\mathbf{V}$ of penetration is given by

$$\mathbf{V} = \sum_i (\mathbf{y}_i - \mathbf{x}) \,. \tag{8}$$

   After computing $\mathbf{V}$, the ligand atom at $\mathbf{x}$ is displaced to a neighboring empty grid point in the direction of $-\mathbf{V}$.

Table 6 lists the results of these tests. In all tests, 13 of 16 test cases have higher (or the same) ranks compared to the case before treatment of the overlaps. The guided displacement of overlapping atoms has better performance than the other crude treatments. This shows that when using electrostatic complementarity, overlap treatment is an important aspect for increasing the reliability of the fitness score.

5.3 Robust Registration

Conformational changes happen when proteins bind to each other. To study these changes, an unbound protein is spatially registered with its bound version and the displacements of the atoms are measured by rigid transformation (rotation and translation).

Rigid registration has a drawback in comparing two conformational states of the same protein. For a protein that binds with another molecule, conformational changes may occur only in some parts of it. Rigid registration will result in misalignment everywhere so that the overall error is minimized. Therefore, to study conformational changes, robust registration is preferred. Robust registration aligns the unchanged parts of the two molecules as much as possible such that conformational changes are distinguished.

Table 6: Results of tests on different treatments of intermolecular overlaps in evaluating electrostatic complementarity score.

Test Case	E_{elec} of Bound Complex	FFT Solution		Excluding Overlaps		Random Displacement		Guided Displacement	
		Rank	E_{elec}	Rank	E_{elec}	Rank	E_{elec}	Rank	E_{elec}
1ABI	-66.74	7	-64.82	4	-14.47	7	-14.28	5	-14.03
1ACB	-15.41	1	-130.59	1	-29.78	1	-31.61	1	-31.58
1CHO	-24.5	6	-112.66	4	-29.87	4	-36.76	5	-34.72
1CSE	-4.32	8	-21.23	2	-20.1	3	-20.33	3	-20.75
1FDL	-5.32	13	-8.06	49	5.42	29	2.32	44	6.09
1TGS	-20.02	9	-44.92	4	-23.95	4	-34.51	4	-33.24
2KAI	-19.5	7	-56.1	3	-21.09	2	-36.64	2	-27.01
2MHB	-4.99	12	-18.7	3	-6.17	4	-8.21	4	-8.81
2PTC	13.9	2	-98.15	1	-47.87	1	-56.58	2	-51.14
2SIC	-4.22	2	-88.11	6	-6.84	8	-8.34	5	-8.12
3HFM	1.07	35	-0.4	37	0.67	32	0.19	33	0.16
4HVP	-6.71	11	-4.72	5	-2.94	5	-4.85	5	-4.28
4SGB	-2.76	11	-15.54	9	-3.45	11	-3.82	6	-4.29
4TPI	-12.76	3	-80.16	3	-12.12	4	-14.21	4	-12.26
9LDT	-33.02	2	-68.91	2	-34.85	2	-36.28	2	-35.13
9RSA	-46.52	3	-45.78	1	-51.37	1	-50.3	1	-50.36

Proteins are made of residues and each residue contains several atoms. The correspondence between the atoms in unbound protein and those in bound protein is known. A residue contains enough atoms for determining a transform matrix. So, we can register each residue and determine the residues with the smallest conformational changes. The robust registration algorithm can be summarized as follows:

1. For each residue, compute the transformation T between its bound and unbound version using the atoms in this residue. Then, register the whole molecule using T and measure the global registration error of the whole molecule.

2. Collect a set of residues one at a time in increasing order of the global registration error. Compute the transformation of this set, and use it to register the whole molecule. Also compute a local registration error of the set of residues. Stop adding residues when the error increases significantly.

Figure 20 shows the global registration error curves for 18 test cases. The residues that produce high global registration errors are possibly those having conformational changes. Figure 21 shows the local registration errors of sets of 5%, 10%, ... 100% of all residues for 18 test cases.

Using this algorithm, we identify those parts of the molecule that can be registered with a small error. These are the residues that have little or no conformational changes. Then, the same transformation is applied to register the whole molecule. Thus, conformational changes are distinguished. Figure 22 to 24 show comparisons of robust registration with rigid registration.

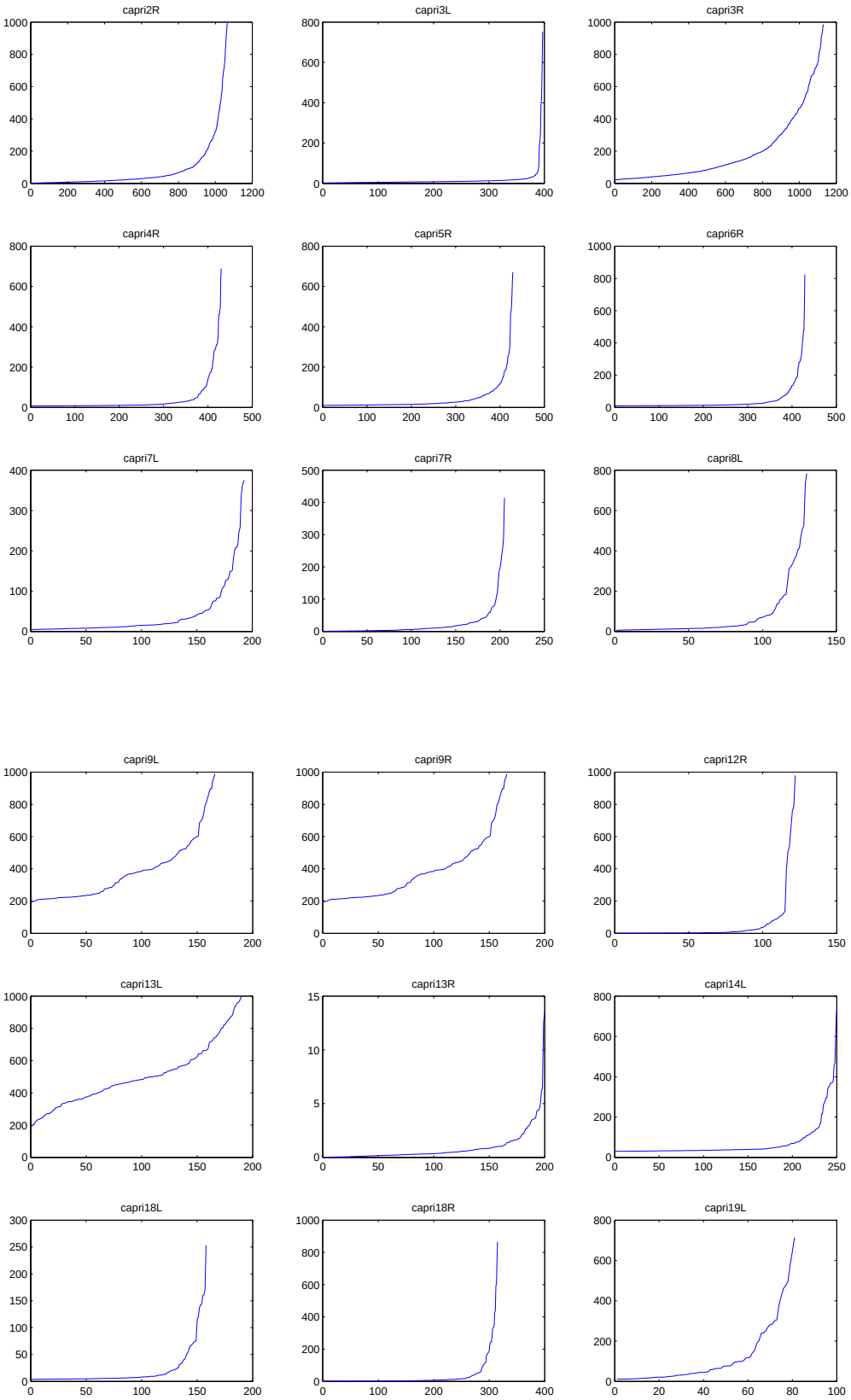

Figure 20: Sorted global registration errors corresponding to each residue.

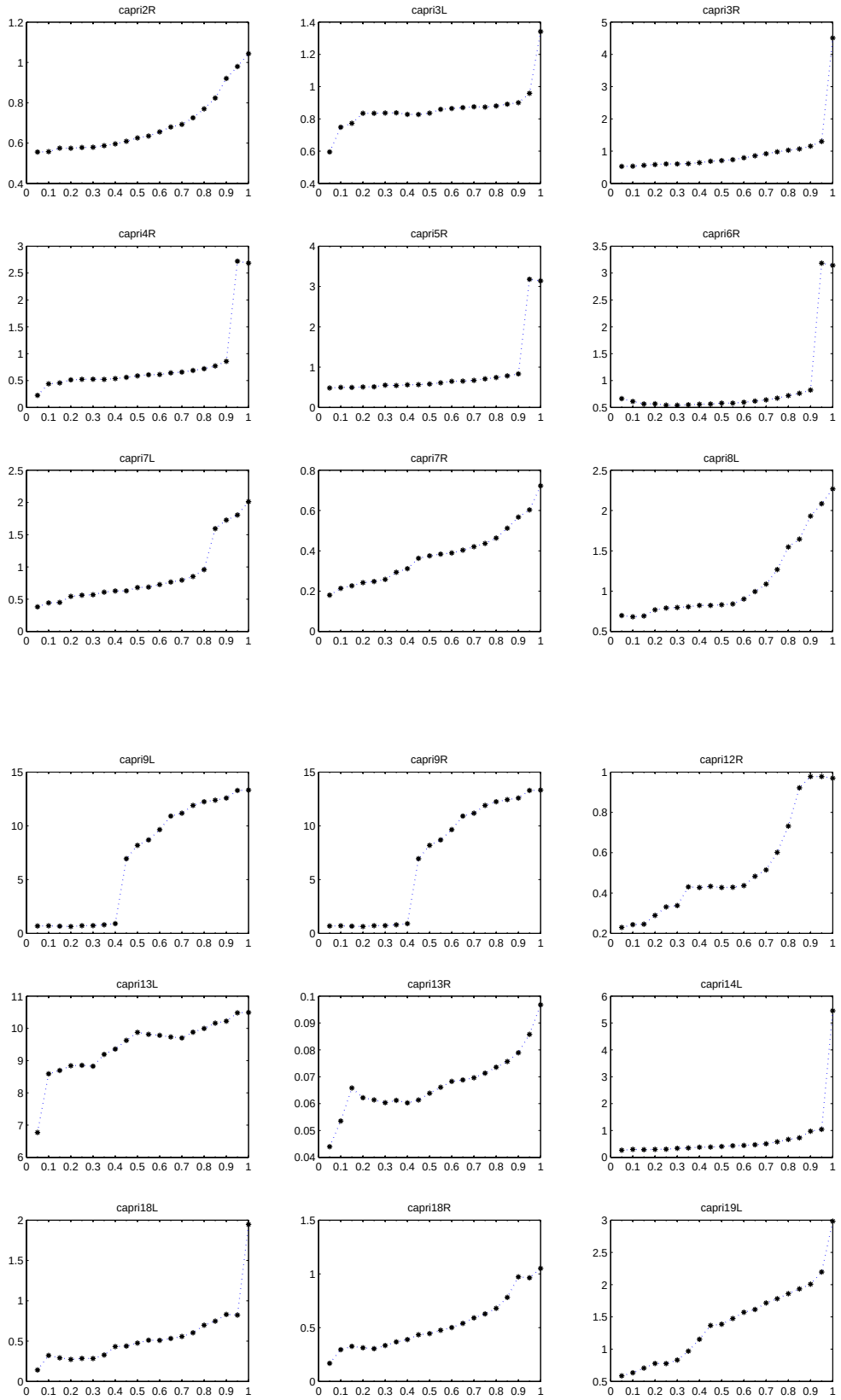

Figure 21: Local registration errors of sets of 5%, 10%, ... 100% of all residues.

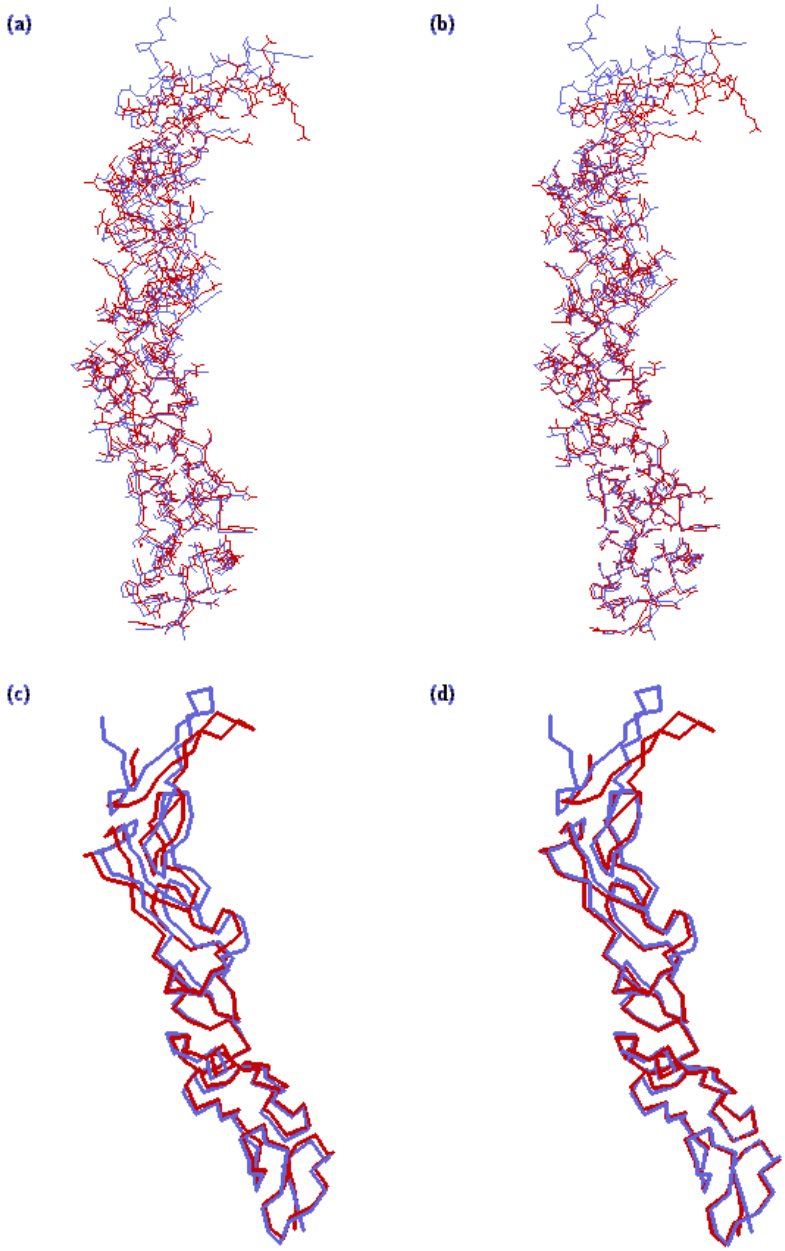

Figure 22: Registration of bound and unbound protein Nidogen-G3. (a) Rigid registration shown as wireframe. (b) Robust registration shown as wireframe. (c) Rigid registration shown as backbone. (d) Robust registration shown as backbone. The bound molecule is in blue and the unbound molecule is in red.

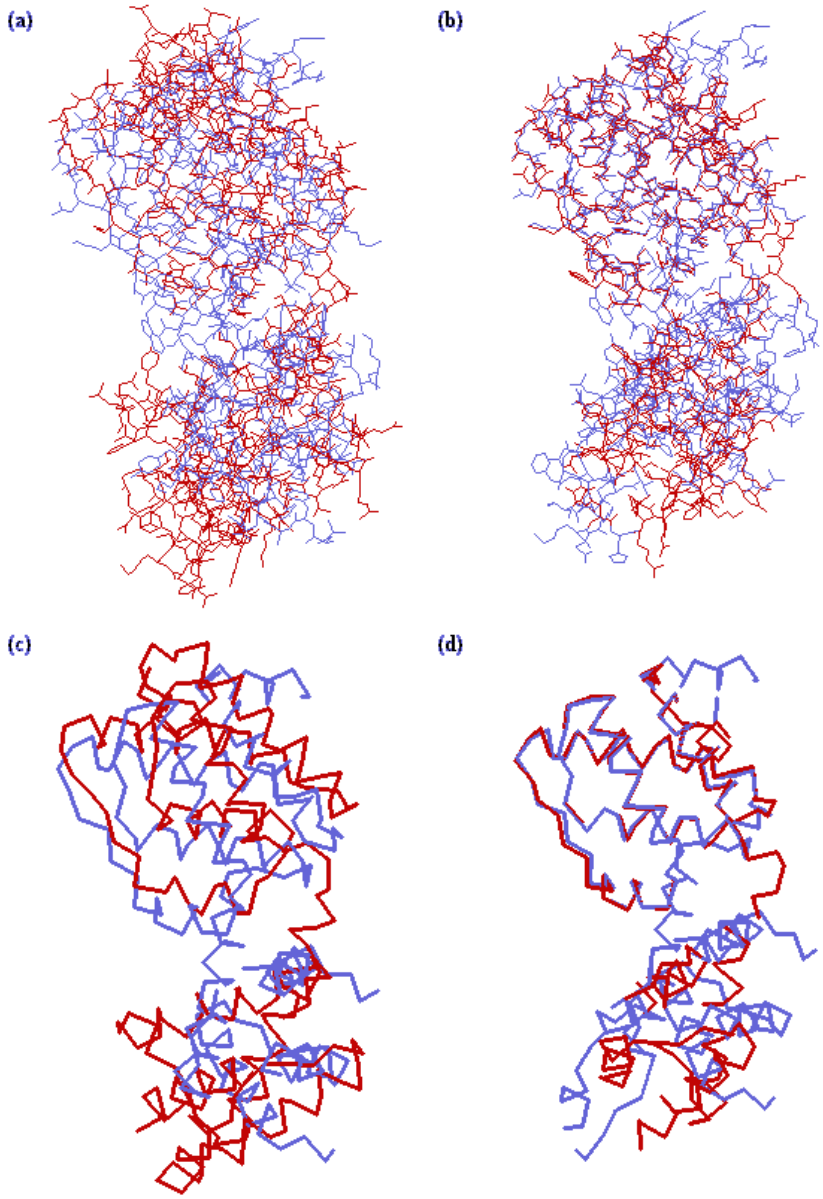

Figure 23: Registration of bound and unbound protein LicT homodimer. (a) Rigid registration shown as wireframe. (b) Robust registration shown as wireframe. (c) Rigid registration shown as backbone. (d) Robust registration shown as backbone. The bound molecule is in blue and the unbound molecule is in red.

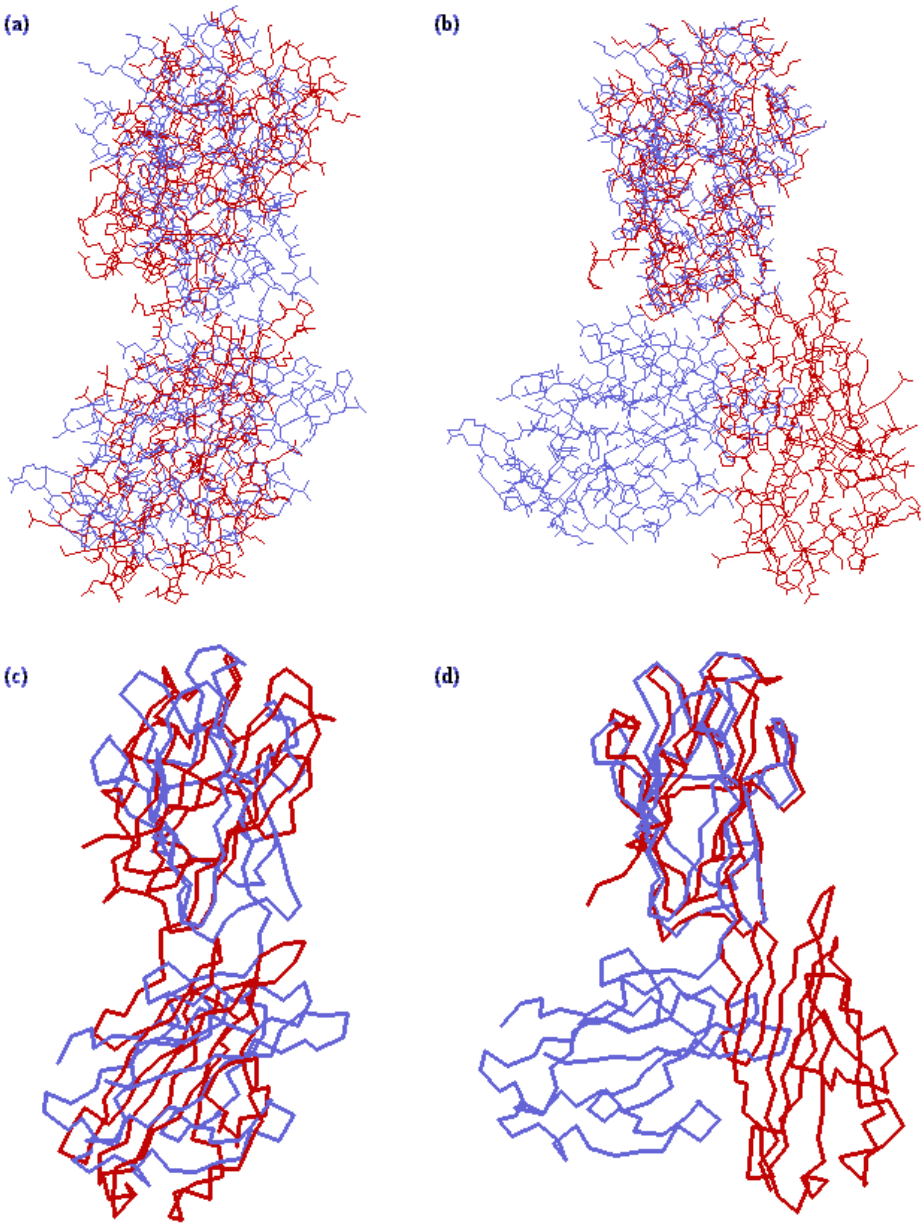

Figure 24: Registration of bound and unbound protein SAG1. (a) Rigid registration shown as wireframe. (b) Robust registration shown as wireframe. (c) Rigid registration shown as backbone. (d) Robust registration shown as backbone. The bound molecule is in blue and the unbound molecule is in red.

6 Conclusion

In this proposal, we propose to model the conformational changes in protein interactions. An assumption is made that the receptor is rigid and the ligand may have conformational changes. In order to assess the goodness of a conformation, an effective fitness function is required. Therefore, two research problems are proposed: the first one is to determine an effective fitness function, and the second one is to model the conformational changes given an initial transformation of the ligand at a possible binding site.

The research plan consists of four stages. The first stage of the research plan is done. The second stage has been explored by studying the properties of van der Waals potential and electrostatic complementarity. Van der Waals potential can be a good factor for the fitness function because it decreases as the candidate docking solution gets closer to the ground truth. Electrostatic complementarity can not be used directly before the intermolecular overlaps are handled properly. The third stage can be inspired by van der Waals potential and its property. The refinement can be guided to seek a minimum van der Waals potential which corresponds to the correct solution. The fourth stage remains to be done.

The research problem can be further extended to unbound docking problem. An unbound docking problem can be divided into two sub-problems: (1) determine the possible binding site, and (2) determine the transformation and conformational changes of the ligand.

Conformational changes are the key aspect in protein interactions and handling them is important for an unbound docking problem. With knowledge of possible binding site, simulation of conformational changes can lead to prediction of a complex made by flexible proteins. Therefore, the proposed research problem is useful for pharmaceutical applications such as drug design that require unbound docking techniques.

References

[1] O. Bachar, D. Fischer, R. Nussinov, and H. Wolfson. A computer-vision based technique for 3d sequence independent structural comparison of proteins. *Protein Engineering*, 6:279–288, 1993.

[2] N. A. Bakern, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, 98:10037–10041, 2001.

[3] M. Betts and J. E. Sternberg. An analysis of conformational changes on protein-protein association: implcations for predictive docking. *Protein Engineering*, 12(4):271–283, 1999.

[4] A. Caflisch, S. Fischer, and M. Karplus. Docking by Monte Carlo minimization with a solvation correction: Application to an FKBP-substrate complex. *Journal of Computational Chemistry*, 18:723–743, 1997.

[5] J. C. Camacho, D. W. Gatchell, S. R. Kimura, and S. Vajda. Scoring docked conformations generated by rigid body protein protein docking. *Proteins*, 40:525–537, 2000.

[6] J. C. Camacho, Z. Weng, S. Vajda, and C. DeLisi. Free energy landscapes of encounter in protein-protein association. *Biophysical Journal*, 76:1166–1178, 1999.

[7] M. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16:548–558, 1983.

[8] L. Ehrlich, M. Nilges, and R. Wade. The impact of protein flexibility on protein-protein docking. *Proteins: Structure, Function, and Genetics*, 58:126–133, 2005.

[9] J. Fernandez-Recio, M. Totrov, and R. Abagyan. Identification of protein-protein interaction sites from docking energy landscapes. *Journal of Molecular Biology*, 335:843–865, 2004.

[10] S. Fraga, J. M. Parker, and J. M. Pocok. *Computer Simulations of Protein Structures and Interactions.*, page 2081. Springer Verlag, New York, 1995.

[11] H. Frauenfelder and E. Gratton. Protein dynamics and hydration. *Methods In Enzymology*, 127:207–216, 1986.

[12] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscape and motions of proteins. *Science*, 254:1598–1603, 1991.

[13] H. A. Gabb, R. M. Jackson, and M. J. E. Sternberg. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology*, 272:106–120, 1997.

[14] E. J. Gardiner, P. Willett, and P. J. Artymiuk. Native protein docking using a genetic algorithm. *Proteins: Structure, Function, and Genetics*, 44:44–56, 2001.

[15] 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. *Journal of Molecular Biology*, 331:281–299, 2003.

[16] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Genetics*, 47:409–443, 2002.

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

[18] R. M. Jackson, H. A. Gabb, and M. J. E. Sternberg. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *Journal of Molecular Biology*, 276:265–285, 1998.

[19] J. Janin and Chothia C. The structure of protein-protein recognition sites. *Journal of Biological Chemistry*, 265:16027–16030, 1990.

[20] J. Janin, S. Wodak, M. Levitt, and B. Maigret. Conformation of amino-acid side-chains in proteins. *Journal of Molecular Biology*, 125:357–386, 1978.

[21] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. Friesem, C. Aflalo, and I. Vakser. Molecular surface recognition: determination of geometric fit between protein and their

ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, 89:2195–2199, 1992.

[22] A. Kitao, S. Hayward, and N. Go. Energy landscape of a native protein: jumping-among-minima model. *Proteins*, 33:496–517, 1998.

[23] F. S. Kuhl, G. M. Crippen, and D. K. Friesen. A combinatorial algorithm for calculating ligand binding. *Journal of Computational Chemistry*, 5:24–34, 1984.

[24] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161:269–288, 1982.

[25] A. R. Leach and I. D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *Journal of Computational Chemistry*, 13(6):730–748, 1992.

[26] B. K. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55:379–400, 1971.

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

[28] R. Méndez, R. Leplae, M. F. Lensink, and Wodak S. J. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Structure, Function, and Bioinformatics*, 60:150–169, 2005.

[29] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and and empirical binding free energy function. *Journal of Computational Chemistry*, 19:1639–1662, 1998.

[30] D. Mustard and D. W. Ritchie. Docking essential dynamics eigenstructures. *Proteins*, 60:269–274, 2005.

[31] C. M. Oshiro, D. Kuntz, and S. Dixon. Flexible ligand docking using a genetic algorithm. *Journal of Computer-Aided Molecular Design*, 9:113–130, 1995.

[32] V. Pande. Simulations of the villin headpiece. http://folding.stanford.edu/villin/, 2002.

[33] D. A. Pearlman and P. S. Charifson. Are free energy calculations useful in practice? a comparison with rapid scoring functions for the p38 MAP kinase protein system. *Journal of Medicinal Chemistry*, 44:3417–3423, 2001.

[34] M. Rarey, S. Wefing, and T. Lengauer. Placement of medium-sized molecular fragments into active sites of proteins. *Journal of Computer-Aided Molecular Design*, 10:41–54, 1996.

[35] D. Ritchie and G. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins: Structure, Function, and Genetics*, 39(2):178–194, 2000.

[36] B. Sandak, H. J. Wolfson, and R. Nussinov. Hinge-bending at molecular interfaces: Automated docking of a dihydroxyethylene-containing inhibitor of the HIV-1 protease. *Journal of Biomolecular Structure and Dynamics*, 1:233–252, 1996.

[37] B. Sandak, H. J. Wolfson, and R. Nussinov. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins*, 32:159–174, 1998.

[38] H. Schrauber, F. Eisenhaber, and P. Argos. Rotamers: to be or not to be? an analysis of amino acid side-chain conformations in globular proteins. *Journal of Molecular Biology*, 230:591–612, 1993.

[39] A. P. Singh, J. C. Latombe, and D. L. Brutlag. A motion planning approach to flexible ligand binding. In *Proceedings of the 7th Conference on Intelligent Systems in Molecular Biology (ISMB)*, pages 252–261, 1999.

[40] C. J. Tsai, D. Xu, and R. Nussinov. Protein folding via binding, and vice versa. *Fold Design*, 3:R71–R80, 1998.

[41] J. D. van der Waals. *Over de Continuiteit van den Gas-en Vloeistoftoestand [On the Continuity of the Gaseous and Liquid States].* PhD thesis, Leiden, A, W, Sijthoff., 1873.

[42] P. H. Walls and M. H. J. Sternberg. New algorithm to model protein-protein recognition based on surface complementarity. *Journal of Molecular Biology*, 228:277–297, 1992.

[43] C. Wang, O. Schueler-Furman, and D. Baker. Improved side-chain modeling for protein-protein docking. *Protein Science*, 14:1328–1339, 2005.

[44] P. K. Weiner and P. A. Kollman. Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, 2:287, 1981.

[45] Huang Wenfan. Rigid body protein docking by Fast Fourier Transform. Honours Year Project Report. School of Computing, National University of Singapore, 2005.

[46] D. Whitford. *Proteins: structure and function.* J. Wiley & Sons, 2005.

[47] H.J. Wolfson and R. Nussinov. Geometrical docking algorithms. A practical approach. *Methods in Molecular Biology*, 143:377–397, 2000.