National University of Singapore
School of Computing
Dept. of Computer Science

Graduate Research Paper

Protein Docking

by

Lu Haiyun

Supervisor: Dr. Leow Wee Kheng (Associate Professor)

2005

1 Introduction

1.1 Motivation

Research focus on proteomics has increased in recent years. Many efforts are 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 and prevent the disease-causing activities to happen. As shown in Figure 1, a small drug molecule (as shown in green) binds to a protein (as shown in red) to form a new complex. The binding mode of both molecules shown in the figure 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 lab 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 costs. Therefore, there is a need to predict possible binding using computational algorithm.

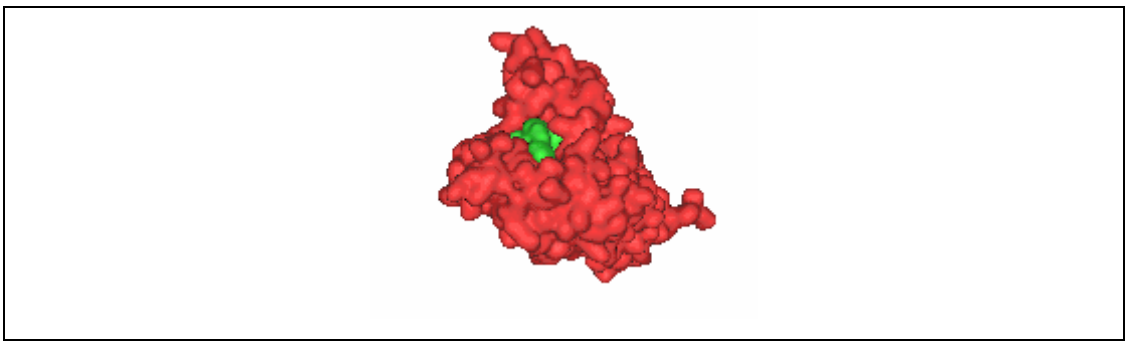


Figure 1. A small molecule drug (green) binds to a disease-causing protein (red).

1.2 What is Protein Docking?

Protein docking is a computational problem to predict the binding of a protein with potential interacting partners. The docking problem can be defined as: Given the atomic coordinates of two molecules, predict their correct bound association [14], which is the relative orientation and position after interaction. In the most general form, no additional data are used for the prediction. In practice, additional biochemical information may be provided, such as salt concentration and temperature of the solution. In particular the knowledge of the binding sites simplifies considerably the docking problem.

In the protein docking problem, the two molecules are named as receptor and ligand. Usually, the smaller molecule is chosen as the ligand. There are two variation of protein docking. The simpler problem is bound docking. It attempts to reconstruct a complex using the bound structures of the receptor and the ligand. A bound structure is extracted from a structure containing more than one molecule, typically a complex of the receptor and the ligand.

The more difficult problem is predictive docking, also referred to as the unbound docking. It attempts to reconstruct a complex using the unbound structures of the receptor and the ligand. A protein in its unbound structure usually undertakes conformation changes to bind with the other molecule. That is, the 3D shape of the protein changes. Thus, the difficulty of the problem increased. 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 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.

There are three key components in protein docking: (1) representation of the molecules, (2) searching and (3) scoring of the potential solutions. Since protein docking is a computational problem, the two molecules involved are to be represented computationally. The searching algorithms are used to find a set of candidate docking solutions and the scoring algorithms are used to rank the solutions. Usually scoring algorithm computes the cost of a solution, and the best solution is the one with minimum cost. The three aspects are mutually interrelated. The choice of the molecule representation decides the types of search algorithms, and the ways to rank potential solutions. In the following sections, we review the principles of the representation, available search algorithms, and scoring schemes. Based on these, we highlight some possible research topics. Before those, let us discuss some background information.

2 Background

Protein is made from a long chain of amino acids, each links to its neighbor through a covalent peptide bond. 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. In average, protein contains around 350 amino acids. Therefore, protein is also known as polypeptides.

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

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

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

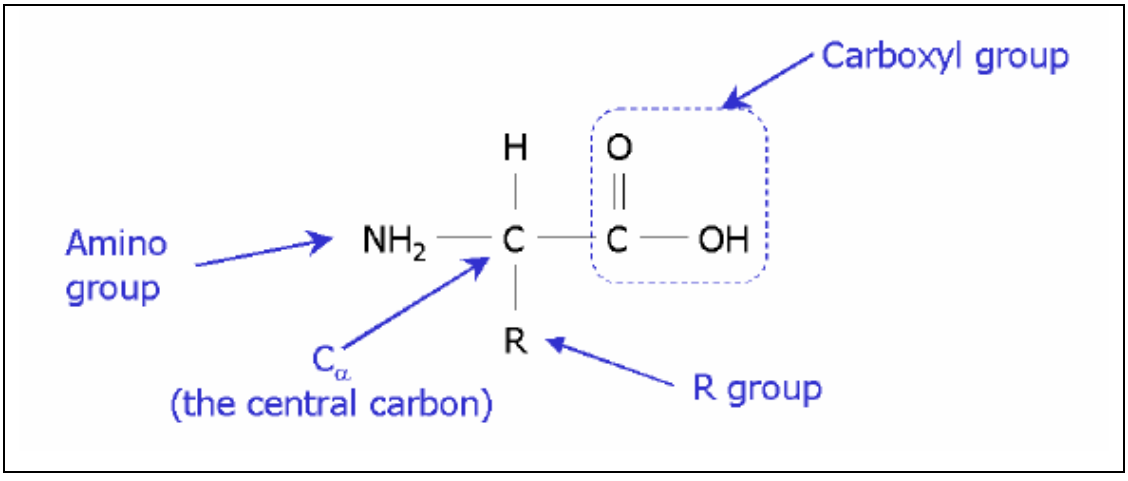


Figure 2. Structure of amino acid

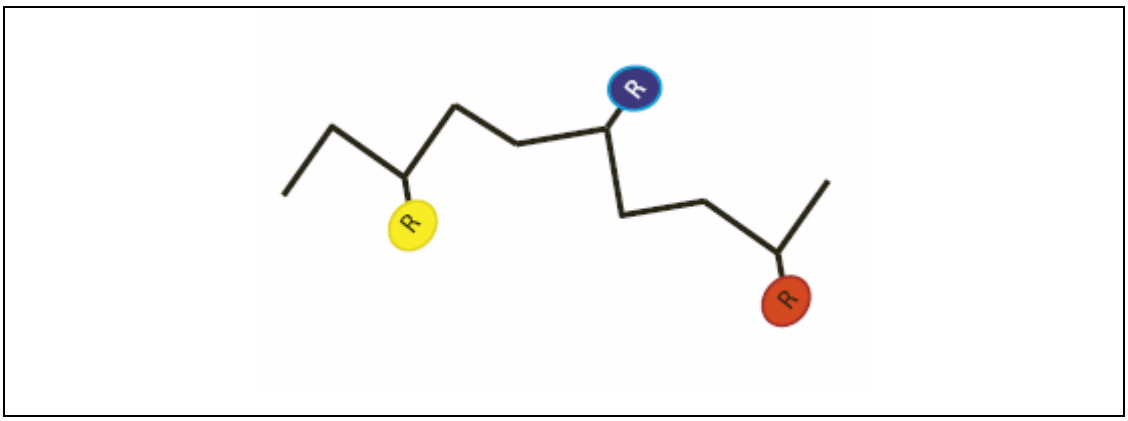


Figure 3. Backbone (black line) and side-chains (R) of protein structure

In order to perform its chemical function, proteins need to fold into certain three-dimensional shapes. A free protein can exist in a range of conformational substates. In each substate, the protein structure is slightly different. Experimentally determined 3D structure for a protein is available, in most cases, for only one conformational substate. As the protein-protein interaction will stabilize both proteins and force them into equilibrium, it usually alters the structures of both participant of the interaction and makes them into another conformational substate.

3 Related Work

3.1 Physical Models of Molecule

Docking essentially simulates the interaction of two protein surfaces. Therefore, the first question is how to define a protein surface. 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. This also suggests that atoms must occupy a well-defined molecular volume. The simplest model of an atom is a sphere. The radius of the sphere depends on the complexity of the atom, i.e. the number of electrons. Assuming that the spheres of two atoms just touch, the measured inter-atomic distance equals the sum of their radii (Figure 4). The radii are called van der Waals radii, and the van der Waals surface of the molecule is the boundary of the union of spheres of each atom in the molecule. [34] Van der Waals radii for different atoms are different.

The solvent accessible surface model and the Connolly surface model use a probe to define the surfaces. The probe is a sphere with an adjustable radius, usually radius of a water molecule. It rolls over the van der Waals surface. As shown in Figure 5, the solvent accessible surface [19] is the trace of the probe center. It takes into account the surface of the molecule which can come into contact with the molecules of the solvent. The Connolly surface [8] is the boundary of the volume which the probe cannot penetrate. Connolly surface is composed of many surface patches, some of which are the van der Waals surface of the molecule, while others come from the surface of the solvent molecule. One of the advantages of Connolly model is the smoothness of the surface.

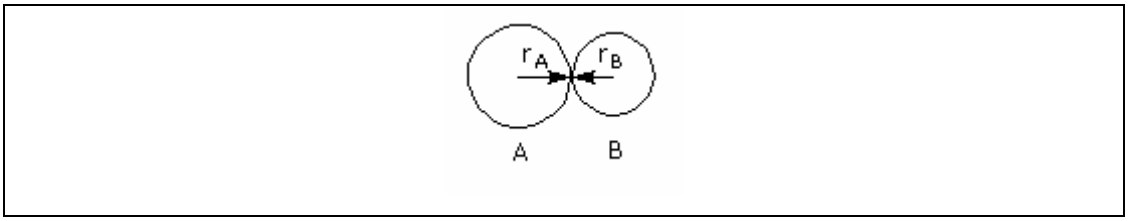


Figure 4. Two unbonded atoms where r_A and r_B are the van der Waals radii of atom A and B.

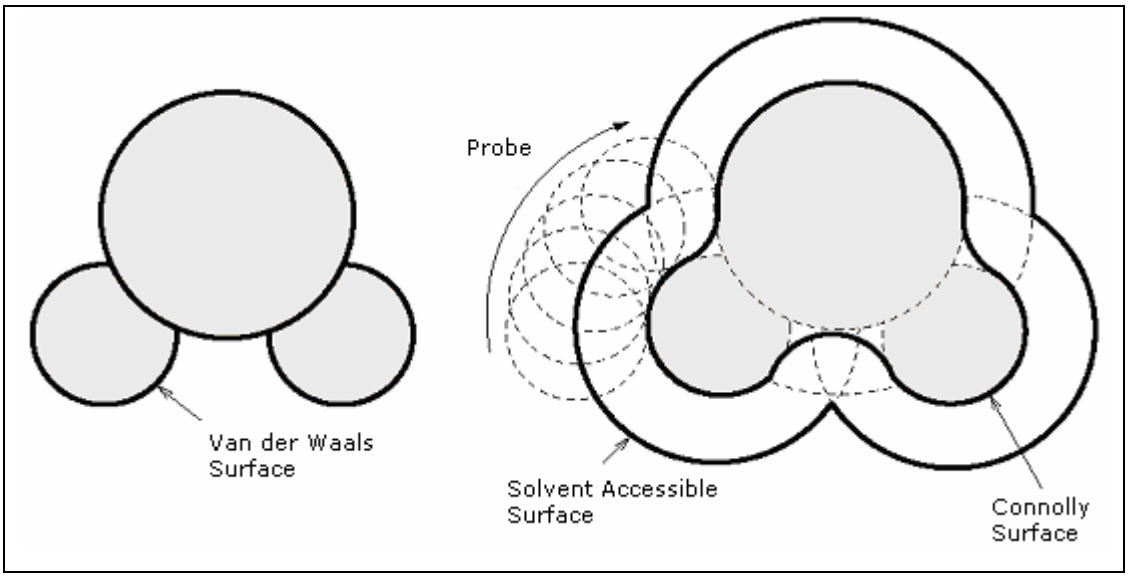


Figure 5. Various molecular surface models

3.2 Computational Models of Molecule

As protein docking is a computational problem, the docking methods have to model receptor and ligand computationally. One intuitive representation is using a set of coordinates of each atom, as well as rotation angles, translations, and torsion angle of each atom bond.

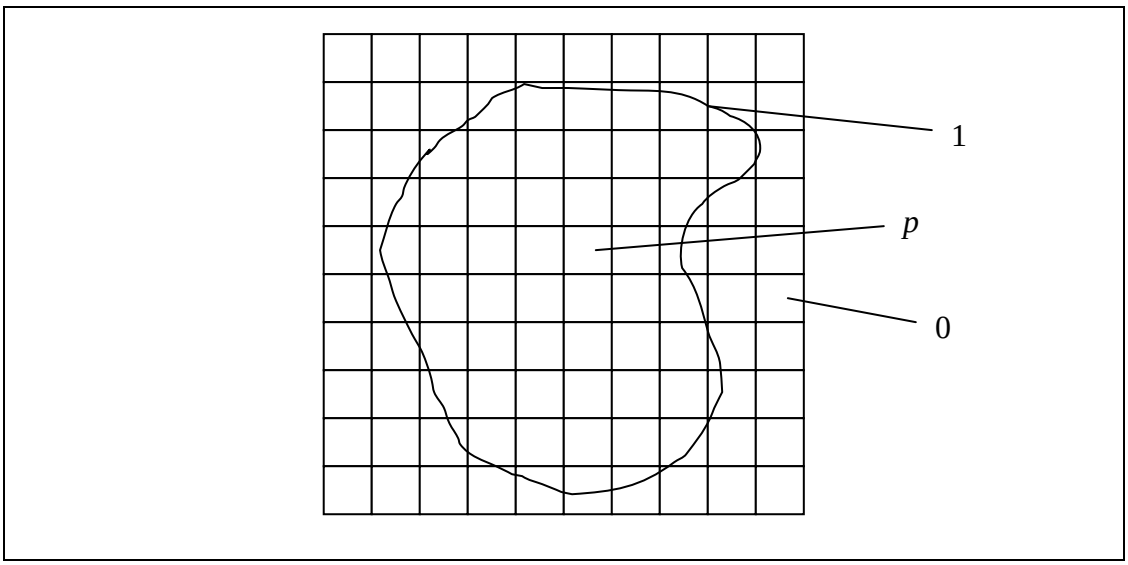


Figure 6. Mapping of the surface of a molecule onto grid

Another representation is mapping the surfaces of the molecules onto three dimensional grids. The surface of molecule is first modeled as var der Waals surface. Then molecules are represented by discrete functions, where 1 denotes grid points on the surface of molecules, *p* denotes grid points inside the molecule, and 0 denotes grid points outside the molecule (Figure 6). This model is used by Fourier correlation search algorithms [18, 36, 4], discussed in section 3.3.1.

Another approach is to regard the molecules as articulated objects. There are two variations. One way is to set one or two hinge points in the molecule, and then the molecule is divided into several domains (Figure 7) [25-29]. The hinge-bending movements of domains are then allowed. This model is used in algorithm of domain movement discussed in section 3.3.2. The other way is to model a small ligand as an articulated robot (Figure 8) [31]. Each atomic bond of the ligand molecule maps to a joint of the robot with torsion freedom of motion. Bond angles and bond lengths are kept constant. Bonds involved in a ring are modeled as rigid. The root atom, which represents the free base of the robot, is an arbitrarily chosen terminal atom from the ligand. This model is used in motion planning algorithms discussed in section 3.3.2.

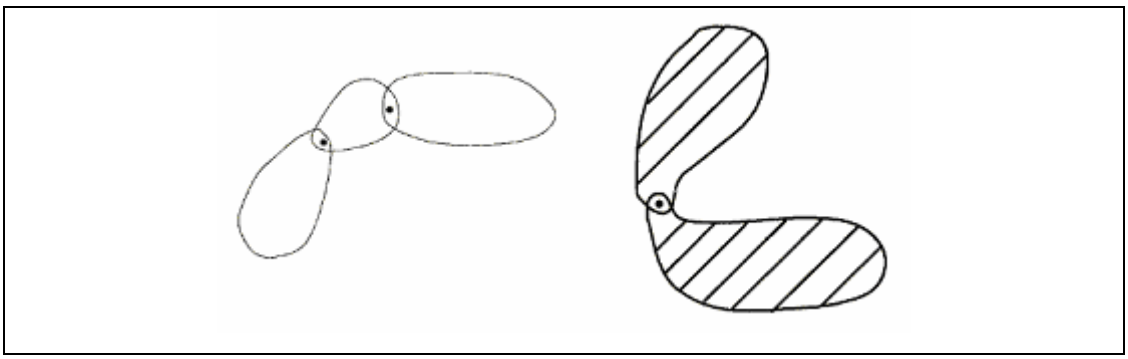


Figure 7. Hinge point(s) is added into molecules.

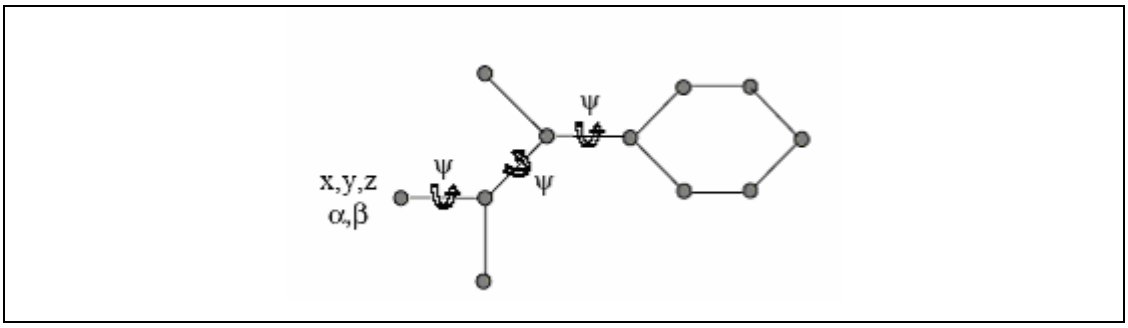


Figure 8. A ligand with 8 degrees of freedom (3 coordinates (x, y, z) and 2 angles (α, β) for the root atom plus one torsion angle (ψ) for each non-terminal atom).

3.3 Search Algorithms

Depending on the extent of flexibility taken into account, docking algorithms can be classified to three categories [11]: (1) Rigid body docking. Both molecules are regarded as rigid solid bodies. No conformational changes will happen. (2) Semi-flexible docking. One of the molecules, usually the smaller ligand, is considered flexible, while the receptor is considered as rigid. In this case, ligand may have conformational changes. (3) Flexible docking. Both molecules are considered flexible and may have conformational changes.

Protein docking uses searching algorithms to find the solution with most stable state in the energy landscape. There are two different approaches of searching for candidate solutions: (1) a full solution space search, and (2) a gradual guided progression through solution space. The first one scans the entire solution space in a predefined systematic manner. The second one can be further classified into two approaches: scans part of the solution space in random or criteria-guided manner, or generates solutions incrementally.

3.3.1 Rigid Body Docking

One approach is to perform exhaustive search on six degrees of freedom, rotational and translational. In this approach, the molecules are represented in terms of three-dimensional grid, as discussed in section 3.2. The matching of surfaces is then computed by the correlation function of two discrete representations of molecules. When two molecules has 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, as the p value is positive for one molecule and negative for the other. When the geometric match is good (Figure 9d), the correlation value is a high positive peak.

Fourier transformation is used to calculate the spatial correlation more efficiently [18, 36, 4]. Multiplication in Fourier domain corresponds to the translational search in sample domain, and it can be done very fast. However, correlation must be calculated for all relative orientation of two molecules. This exhaustive shape-based algorithm works well when both molecules are considered as rigid bodies and only when shape complementarity is essential in the docking.

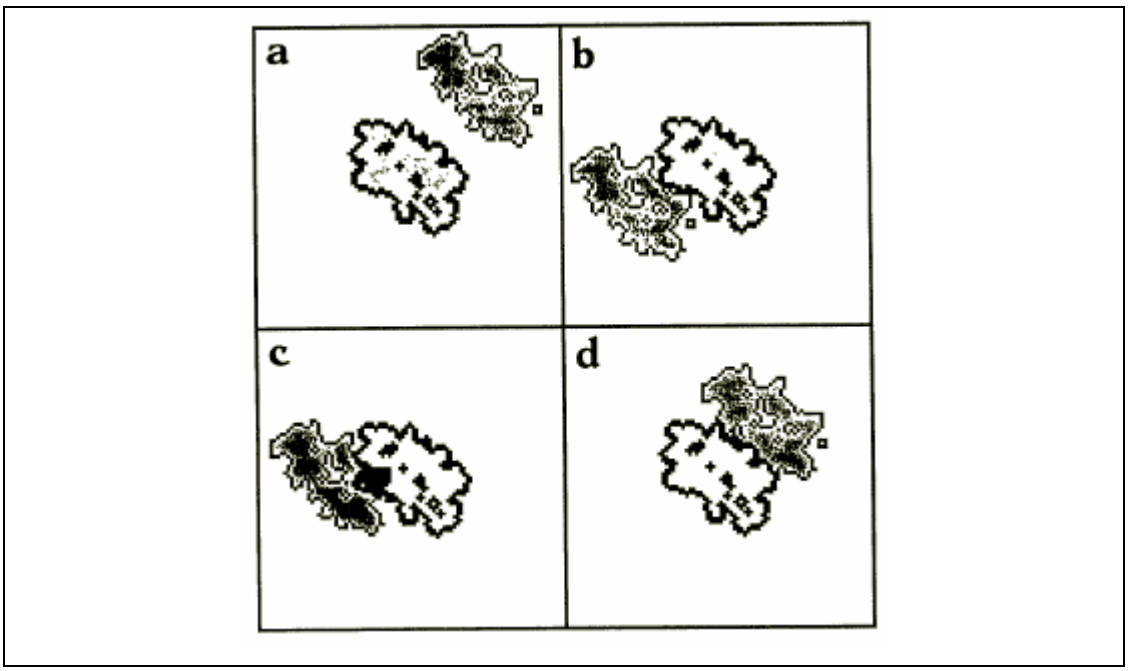


Figure 9. Different relative positions of two molecules. (a) No contact. (b) Limited contact. (c)Penetration. The penetrated part is represented in black. (d) Good geometric match.

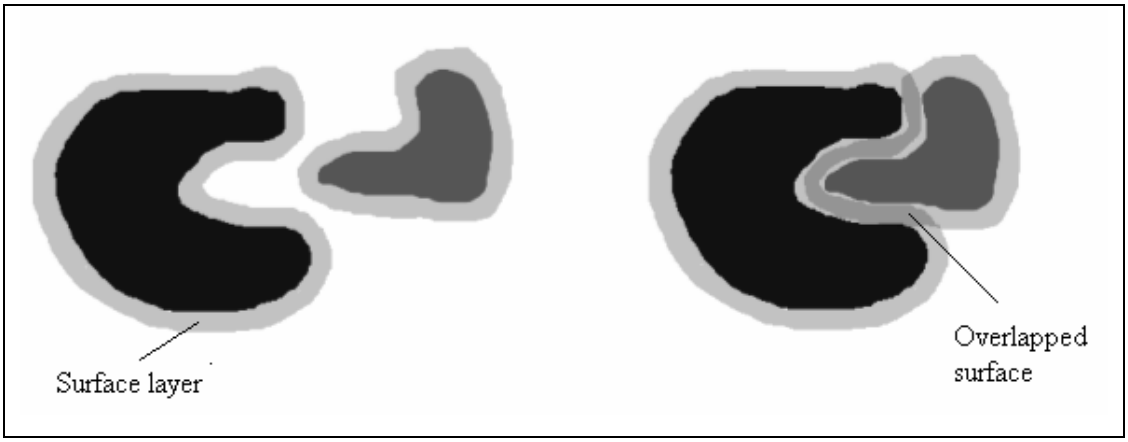


Figure 10. Surface layers of two molecules are allowed to be penetrated to achieve small-scale flexibility.

3.3.2 Semi-flexible & Flexible Docking

Even using a rigid body docking algorithm, small-scale flexibility can be achieved by using a surface layer that allows penetration. As shown in Figure 10, the surface layers of two molecules are allowed to be penetrated, such that it represents some flexibility to a certain extent.

An incremental algorithm is developed for semi-flexible protein docking. In the algorithm, the ligand is divided into fragments. The algorithm works by first placing a base fragment into the pre-defined active site of the receptor, followed by a greedy searching to incrementally add more fragments and grow the base fragment to the final optimal conformation. For each fragment, it is added in such a way that the cost function is minimized. FlexX [24] uses an incremental construction algorithm. This method is fast. Hoffmann proposed a two-stage method of docking [15]. The first stage uses the fast algorithm such as FlexX to generate a large number of plausible ligand conformations, using a simple cost function. The second stage uses a more detailed cost function to re-rank the candidate solutions. Using incremental algorithm, certain 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 the knowledge of binding site is required.

Monte Carlo algorithm may be used in semi-flexible protein docking problem. In the algorithm, the receptor is treated as rigid body, and ligand is considered as flexible. Ligand is represented by a set of variables consisting of rotation angles, translations, and torsion angle of each atom bond. In each Monte Carlo cycle, those variables are assigned random selected values according to a uniform distribution of specified sets of allowed values. Then a cost is computed. Those solutions with cost smaller than the lowest cost found up to current Monte Carlo cycle are saved as candidate solutions. This algorithm was employed by Caflisch et al [3]. Monte Carlo algorithm

is not an exhaustive searching method. It generates potential solutions randomly, and may not generate the correct solution. In order to generate solution that is close to correct one, a large number of cycles are needed, and thus slow down the process.

A motion planning approach to semi-flexible docking also avoids a full solution space search. By modeling the flexible ligand as an “articulated robot” (discussed in section 3.2), the robot motion planning is applicable. 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. It tries to simulate the motion of ligand towards to receptor in real interaction process. Hence, it examines the possible motions of the robot induced by the energy landscape of its immediate environment. The more energetically favorable paths between initial and goal position of ligand are computed. Singh et al proposed to use motion planning algorithm to solve the semi-flexible protein docking problem [31]. The knowledge of binding site is needed, and used as the goal position of ligand. This algorithm is energy-based only and has no concern about shape of molecules.

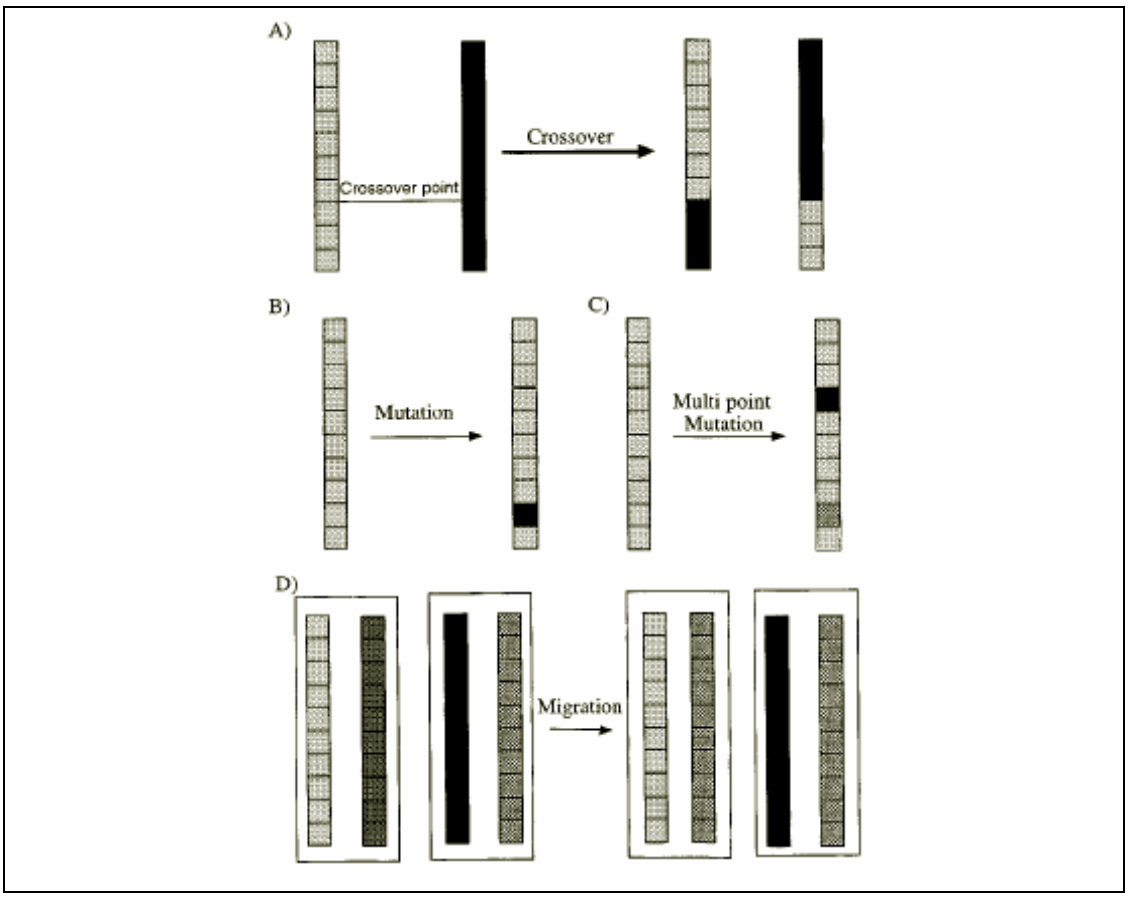


Figure 11. Operations in Genetic Algorithm

Genetic algorithm can also be used in flexible protein docking problem. Genetic algorithm evolves the population of possible solutions through genetic operators to a final population, optimizing a predefined fitness function. For the purpose of flexible docking, the translation, rotation, and torsion angles of internal atom bonds encoded into genes for both molecules. Every individual consists of a collection of genes and it is assigned a fitness value. The number of genes depends on the number of internal bonds. A new population is generated from the old one by the use of three genetic operations (Figure 11). The mutation changes the value of the gene by a random value depending on the type of the gene. Crossover exchanges a set of genes from one parent to another. Migration moves individual from on subpopulation to another. Parents are selected for breeding based on their fitness values. Oshiro [21]

used genetic algorithm for protein docking. GAPDOCK [13] and AutoDock [20] are similar programs developed using genetic algorithms. Genetic algorithm is not an exhaustive searching method; instead, it generates potential solutions, which may not result into the correct solution. Actually, the quality of the solutions usually depends on the starting genes, the number of evolutionary events (mutations, crosses, and migration), and the fitness function to pick the more favorable conformers. One of the drawbacks of genetic algorithm is that it is too slow for extensive flexible docking of two large molecules.

Movements of domains are essential in simulating protein flexibility. In this algorithm, either ligand or receptor is modeled as hinge-articulated object. Rather than dock each of the molecule parts separately, all parts are docked simultaneously. Like pliers closing on a screw, the receptor closes on its ligand and vice versa (Figure 12). Movements are allowed either in the ligand or in the receptor, 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. A method using domain motion algorithm has been presented by Sandak et al. [25-29]. So far, the algorithm was implemented with at most two hinges. The performance of the method using this representation depends largely on the choice of hinge points. By considering only domains of molecules, the flexibility inside the domain is ignored, which limits the level of flexibility achieved.

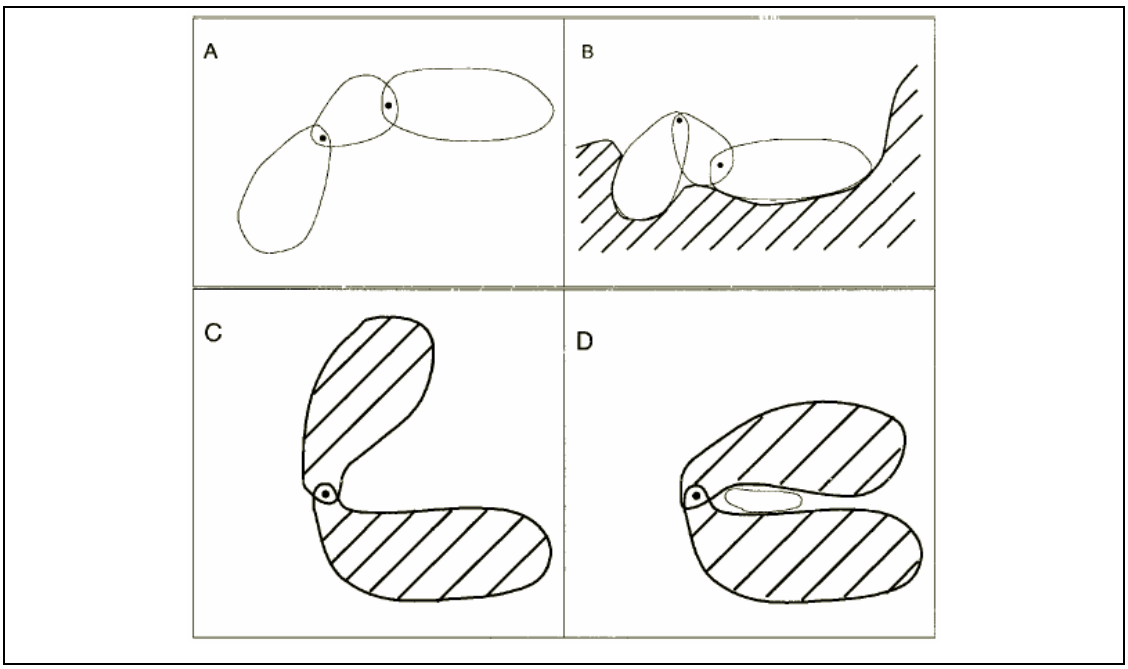


Figure 12. Hinge-bending movements of domains. Shaded domains are from larger receptor and the other is ligand.

A further improvement of existing searching algorithms is to limit the side-chain flexibility. 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 conformation space, whereby a side-chain is only allowed to adopt a discrete set of conformations. 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 [17, 30]. 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 search algorithms mentioned above, such as Monte Carlo algorithm and genetic algorithm. It reduces the searching space on a large scale and thus allows fast sampling of molecules.

3.4 Scoring

A large number of candidate solutions may be produced by a searching algorithm after or during the searching. So, the scoring function is used to access the goodness of the candidate solutions. A scoring function may contain multiple aspects, such as geometric complementarity, 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 it could be used together with searching to prune the solutions. The latter approach is required for some searching algorithms, for example, genetic algorithm needs to apply a fitness function at each generation.

3.4.1 Geometric Complementarity

Geometric complementarity is the measurement of how the 3D structures of two molecules match each other at the contacting interface. It plays an important role in protein docking since most protein-protein interactions presents a good geometric complementarity [33, 18].

There are several definitions available for geometric complementarity. One definition is based on opposite surface normals of contact area between two molecules. Other defines it as the contacting area of two molecules.

3.4.2 Intermolecular Overlap

Intermolecular overlap is the overlap between two different molecules. By allowing some intermolecular overlap, a certain extent of conformational flexibility is taken into account implicitly.

The general approach to intermolecular overlaps is tolerance to slight interface clashes and penalty for large interior clashes. The tolerance is usually implemented by a surface belt of non-penalized penetration area. As shown in Figure 10, a surface layer is used outside the molecules and overlaps of surface layer are allowed.

3.4.3 Intramolecular Overlap

When ligand or receptor flexibility is taken into account, for example, a ligand is divided into fragments, or allowing hinge-movements, the overlaps inside a molecule may occur. Though slight overlap may be considered as flexibility, large scale self collision is not presented in real protein interactions. Usually, penalty for self penetration is given.

3.4.4 Hydrogen Bonds

Polar molecules, such as water molecules, have a weak, partial negative charge at one region of the molecule (the oxygen atom in water) and a partial positive charge elsewhere (the hydrogen atoms in water). 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 13, is called a hydrogen bond.

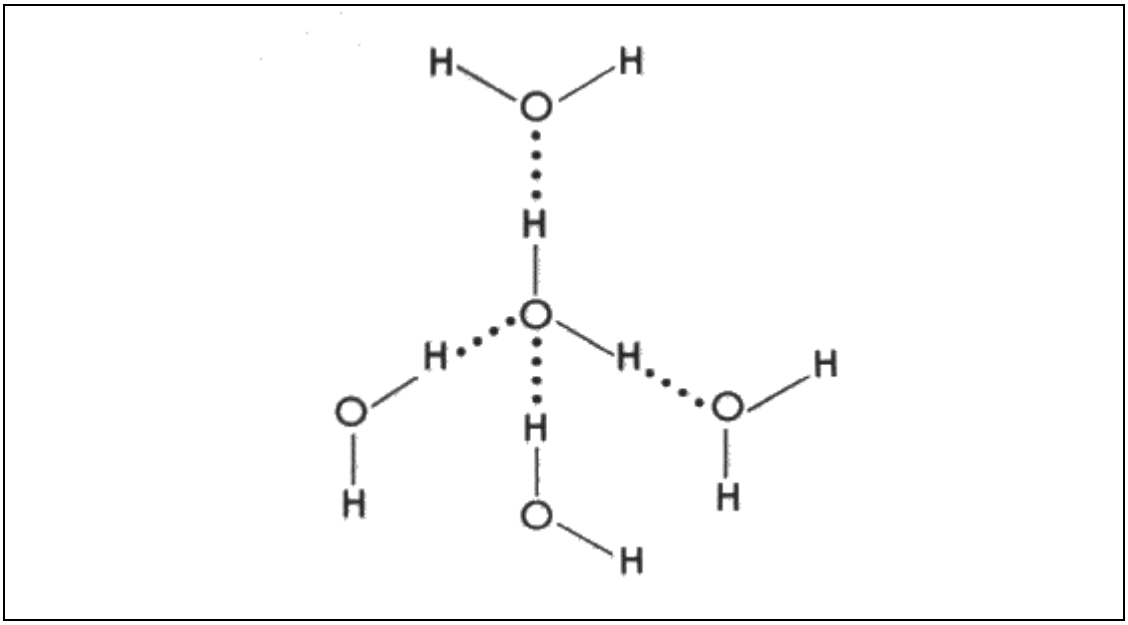


Figure 13. Hydrogen bonds (dotted lines) among five water molecules

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 [2]. (Å stands for Angstrom, 1 Å = 1/10,000,000,000 meter) Thus the number of hydrogen bonds on the interface of two interacting molecules is another important measurement of interaction.

The classification of atoms with respect to hydrogen bonding is: H donor, H acceptor, H donors/acceptors, and non-H bonding. They are matched as following:

- H donor matches H acceptor or H donor/acceptor
- H acceptor matches H donor or H donor/acceptor
- H donor/acceptor matches H donor, H acceptor, or H donor/acceptor
- Non-H bonding matches non-H bonding

Atoms are classified into four types as mention above, and a distance is defined such that the atoms within the distance form matched hydrogen bonding. Different

distances are used by different programs: Gardiner et al [13] used 2 Å, while Ausiello et al [1] used 3.4 Å. Different values of distance represent different level of approximation.

3.4.5 Electrostatic Potential

The common definition of electrostatic potential is: 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 low electron density exists and the nuclear charge is incompletely shielded. Water molecule is a good example to help understanding the electrostatic potential (Figure 14). The hydrogen 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 oxygen atomic nucleus. As the 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.

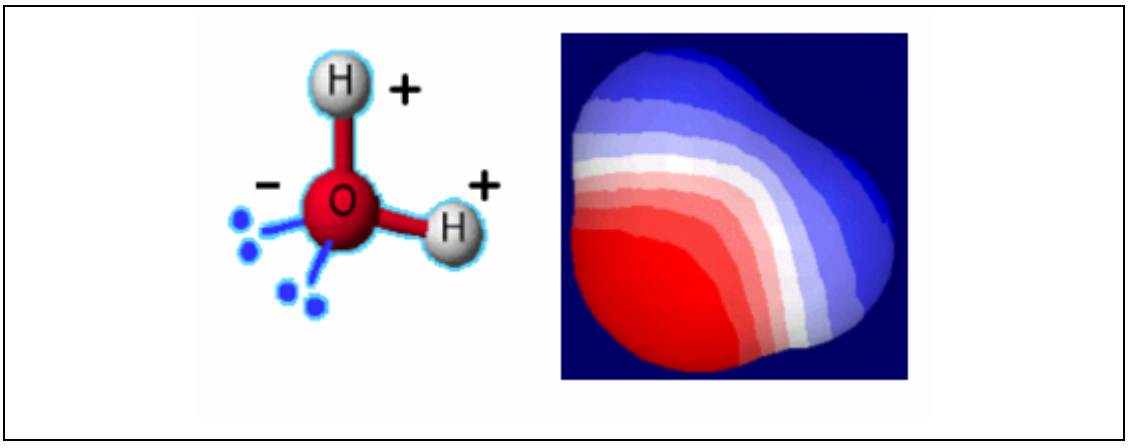


Figure 14. Electrostatic potential on the surface of a water molecule

Electrostatic interactions play an important role in the 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 good association interface [16].

The classical treatment of electrostatic interactions in solution is based on the Poisson-Boltzmann equation (PBE):

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

where $\phi(\mathbf{r})$ is the dimensionless electric potential, ε is the dielectric constant, and ρ is the fixed charge density. The term $\kappa^2 = 1/\lambda^2 = 8\pi q^2 I / \varepsilon k_B T$ where λ is the Debye length, q is he charge on a proton, T is the absolute temperature, and I is the ionic strength of the bulk solution. ϕ, ε, κ, ρ are functions of the position vector **r**. Analytical solutions to the PBE are only available for a limited number of cases involving idealized geometries such as spheres and cylinders. The numerical methods of computing electrostatic potentials can be categorized into two approaches: Finite difference method (FDM), Boundary element method (BEM).

3.4.6 Van der Waals Potential

When two non-bonded atoms are at short distance, the van der Waals attraction occurs; however 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 zero when two molecules are bind stably. In practice, this term should be minimized.

The van der Waals potential is modeled as

$$\sum_{i,j} 4D_{ij}\left[\left(\frac{C_{ij}}{r_{ij}}\right)^{12} - \left(\frac{C_{ij}}{r_{ij}}\right)^{6}\right] \tag{2}$$

where r_{ij} is the distance between two atoms i and j, C_{ij} is the collision constant, and D_{ij} is the value at the unique minimum.

3.4.7 Other Energy Terms

There are other energy terms used in many existing scoring functions, namely bond potential, bond angle potential, torsion angle potential, hydrophobicity, and etc. Please refer to [22] and [5] for more details.

4 Possible Research Topics

The protein docking problem 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.

Rigid body docking is the relatively simpler subset of the protein docking problem. An exhaustive searching method, such as Fourier correlation algorithm, is capable to find the correct solution. However, rigid body docking does not always give correct solutions in the practice, because proteins usually undertake conformational changes when they are bind to the partner.

Semi-flexible and flexible docking problems are more difficult as the search space increases dramatically. No existing method attempts an exhaustive search. In order to reduce the search space, random sampling or criteria guided generation are used. Another major approach is to add certain flexibility into the rigid body docking, such as allowing penetration on the surface or dividing a molecule into small number of parts. The results of those methods may be acceptable for a small set of test cases; however, it is still far from using in practice.

The other bottleneck is the lack of selective and efficient scoring function. Though many of existing methods can have the correct solution in their top hundred or even top ten possible solutions, many solutions are false-positive.

One possible research topic is semi-flexible or flexible protein docking. For example, the topic could be how to do semi-flexible or flexible protein docking efficiently, completely and automatically. We could either try to develop a novel algorithm or make improvements could be made to some existing algorithms. We could use rigid

body docking to quickly prune away poor candidate solutions and then use flexible docking methods to perform further searching. This could be a coarse to fine process.

Another possible topic is to research on the additional bio-chemical information which could more efficiently solve the docking problem. For example, a possible topic could be to identify the potential binding site by examining the energy distribution on the molecule surface. For another example, the topic could be to limit the conformational space of a protein by examining the bound state of its homologous proteins.

Scoring function is another possible research topic. Various evaluation functions, such as shape complementarity and energy terms, could be studied to develop a combination which is more efficient and selective.

5 Preliminary Work

Deformation of 3D model is related to flexible protein docking problem. Though there are differences between protein conformational change and 3D object deformation, some aspects are shared, such as collision detection. A project on constrained deformation of 3D model has been developed.

Electrostatic potential is an important term of the scoring functions used in many current methods for protein docking problem. A project has been developed to calculate the electrostatic potential on the surface of a protein molecule.

5.1 Constrained Deformation of 3D Model

Consider two 3D objects: a large deformable target object and a small rigid probe. The probe object is orientated in a selected manner and pushed into the target object at a selected spot along a selected direction. The target object then undertakes deformation to wrap around the probe object.

5.1.1 Method

Collision detection is one of the major components in the project. Since the 3D models used in the project are surface mesh models with triangle patches, the collision detection is focused on triangle-triangle collision. When two triangles collide with each other, there exists at least one edge of a triangle that penetrates the other triangle. Thus, the collision detection is used to detect triangle-edge collision.

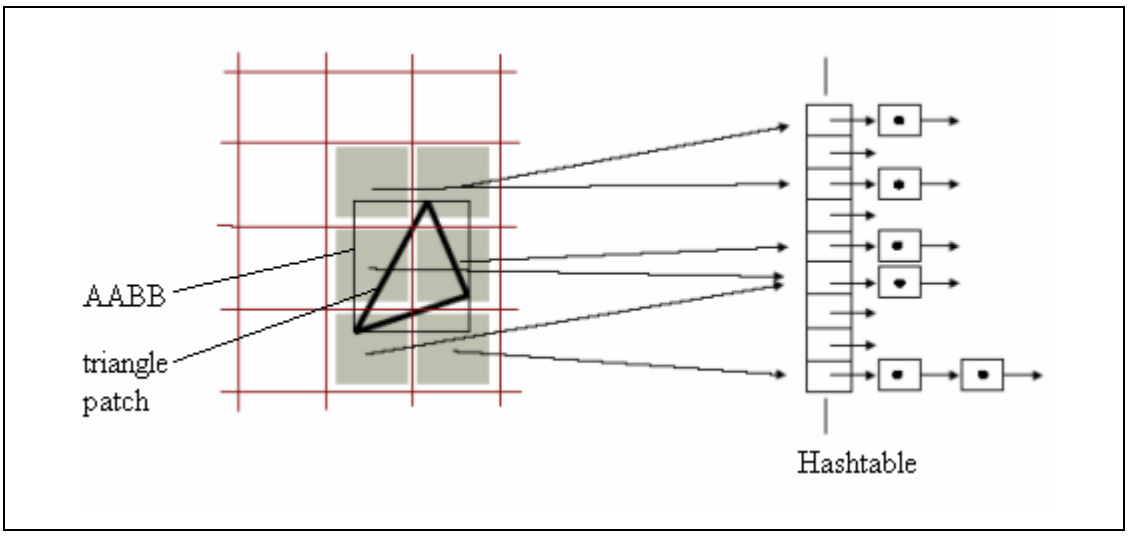


Figure 15. Spatial hasing

It is not efficient to compare all edges of one object against all triangles of the other object. Hence, spatial hashing [32] (Figure 15) is used to accelerate the collision detection process. First, a uniform grid is placed into the space. The size of grid can be user-defined. Each grid cell has grid coordinates (*i*, *j*, *k*), where *i*, *j*, *k* are integers. Then, the Axis Aligned Bounding Box (AABB) of each triangle is computed and mapped to the grid. Usually the AABB of a triangle resides in more than one grid cells. Finally, each triangle is stored into the hashtable. For each grid cell where the triangle's AABB resides, a hash value is computed according to the hash function (Equation 3) and the triangle is stored into the corresponding entry in the hashtable. A triangle may be stored in several entries.

$$Hash\ (\ i, j, k\ ) = (\ i\ P1\ \mathbf{xor}\ j\ P2\ \mathbf{xor}\ k\ P3\ )\ mod\ n \qquad (3)$$

where *P1*, *P2*, *P3* are large prime numbers, *n* is the size of hashtable and is usually a prime number, **xor** is the bitwise exclusive-or operator.

After the hashtable is built for triangles, edges are examined to detect collision. For an edge, the same spatial hashing procedure is applied to find those triangles whose AABB resides in the same grid cell as the edge's AABB. Then penetration test is performed to judge whether the edge and the triangle collides. In this project, edges are tested against triangles from the other object. The self collision is not considered.

The response to a collision is another major component of this project. The purpose is to deform the target object to remove the collision. As the probe is moving towards the target, if the target's parts in collision are moved along the same direction of probe's motion and moved by same displacement, it is guaranteed that collision will be removed. This is an intuitive yet effective approach of deformation.

Figure 16 illustrates two possible cases. First case is when a triangle of the probe (red) intersects an edge of the target (black). The edge is moved at the same direction and by the same displacement as the probe (blue arrow). Collision is removed after such the edge is moved. Second case is when an edge of the probe intersects a triangle of the target. One vertex of the triangle is moved similarly as the previous case. It is possible that moving one vertex of the triangle may not remove the collision, and then other vertices of the triangle should be moved. It is intuitive that the collision will be removed when all three vertices are moved.

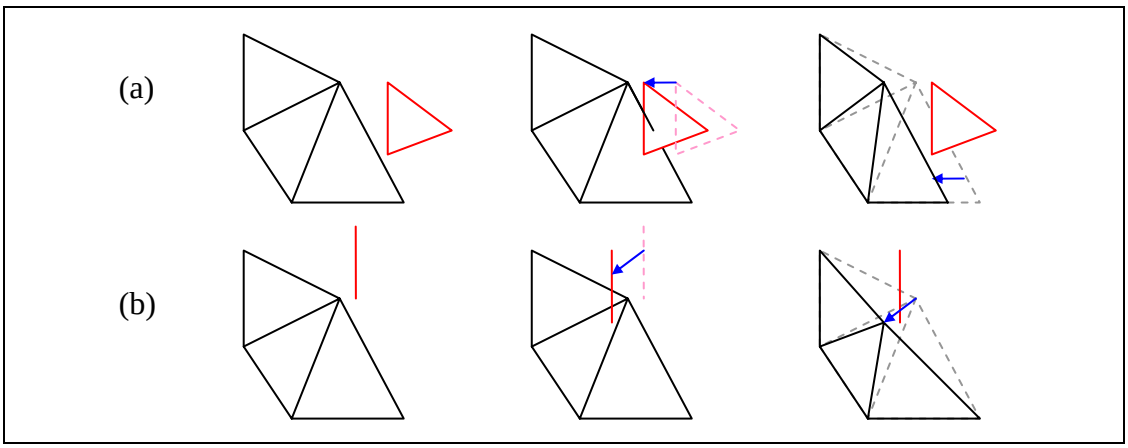


Figure 16. Two possible cases of collision response

Surface smoothing can be added to refine the deformation. Many existing algorithms could be applied to smooth the surface after deforming. Catmull-Clark subdivision method [6] is applied in the test.

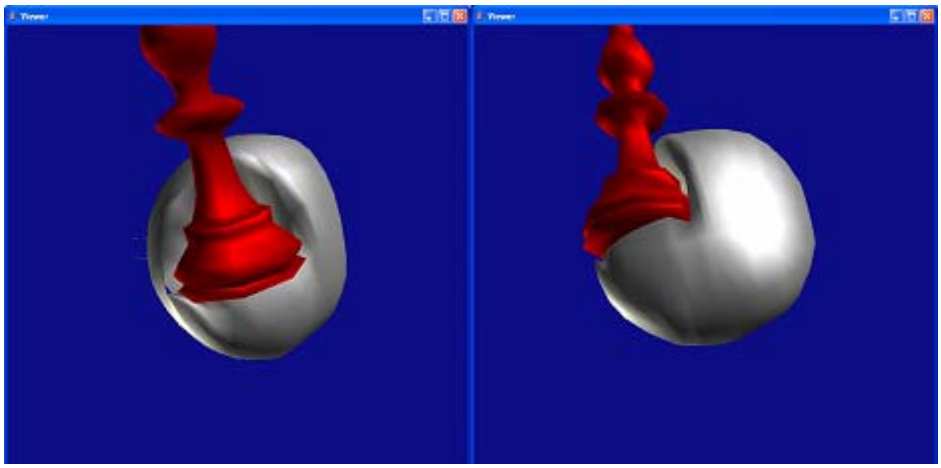
5.1.2 Result

Figure 17 shows two test results using same set of target and probe objects but different orientation and movement of probe. Each result is shown in two different viewing angles. It can be seen from the result that the deformation of the target is satisfactory and roughly fit the shape of the probe.

Figure 18 (a) shows the deformation without applying surface smoothing, and (b) shows the result after smoothing. Comparing the portion highlighted by red circle in both image, we can see that the sharp angle becomes more rounded.

There are possibilities of further improvement of the project. More constrains could be added, such as volume preserving deformation. Boundary element method (BEM) could be used to implement volume preserving property. Other approaches of collision response could be used. For example, adding more vertices on the deforming target yields better wrap around effect.

(a)



(b)

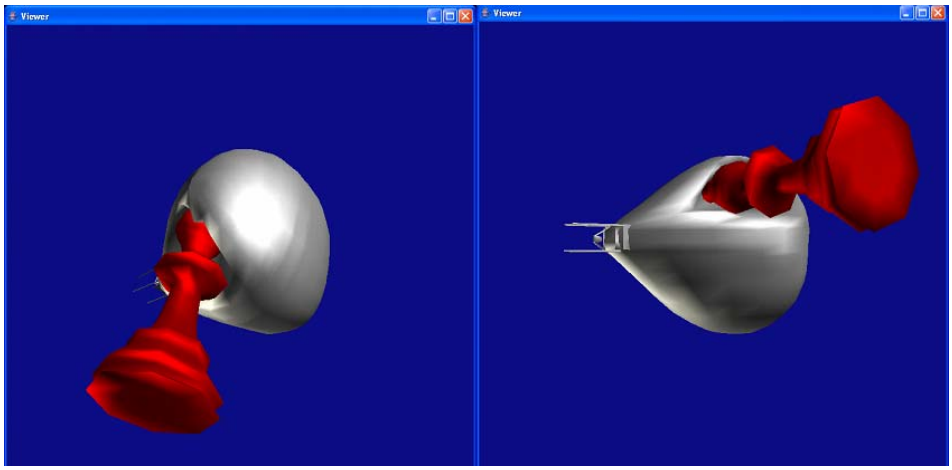


Figure 17. 3D model deformation

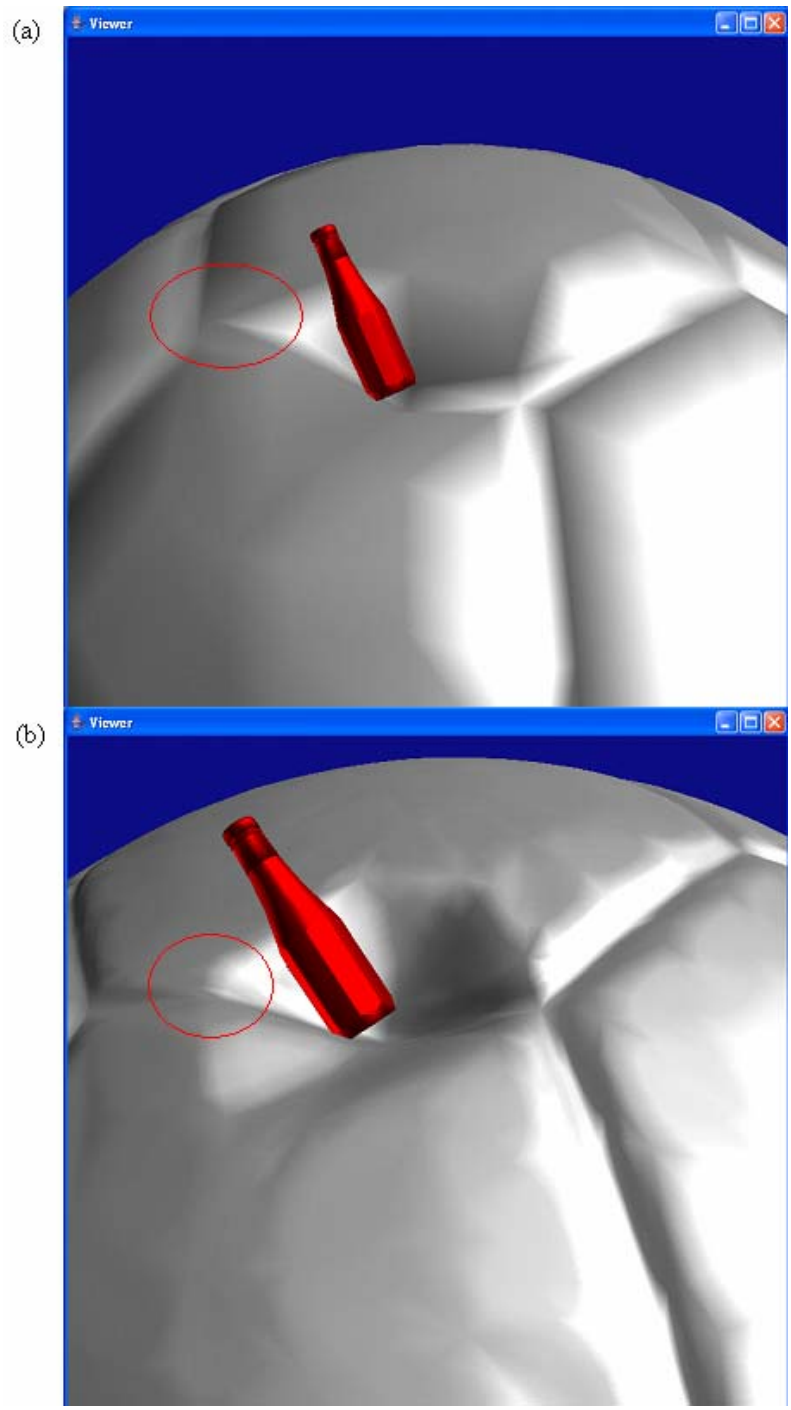


Figure 18. Smoothing of the surface

5.2 Electrostatic Potentials on Molecular Skin Surface

Under the assumption of continuous linear dielectric media with zero salt concentration, such as water, boundary element method (BEM) is applied to calculate the electrostatic potentials on the skin surface of a molecule.

The charges inside the molecule establish a polarization field throughout the volume of the dielectric medium, and it is a well-known classical electrostatic theory that the effects of this field can be exactly reproduced by appropriate distributions of induced polarization charge at dielectric interface (commonly assumed to be molecular surface). So, the focus in BEM approach is to compute the induced charge distribution on the molecular surface [23, 37]. Afterwards, the electrostatic potentials can be obtained.

5.2.1 Method

In the BEM method, the surface of a molecule is represented by a discrete mesh with uniform charge density within each surface element. The molecular skin surface defined by Edelsbrunner [9, 10] is used and the skin triangulation is refined by Cheng et al [7]. Figure 19 shows examples of skin surfaces. Each triangular patch of the mesh is regarded as a boundary element in BEM method.

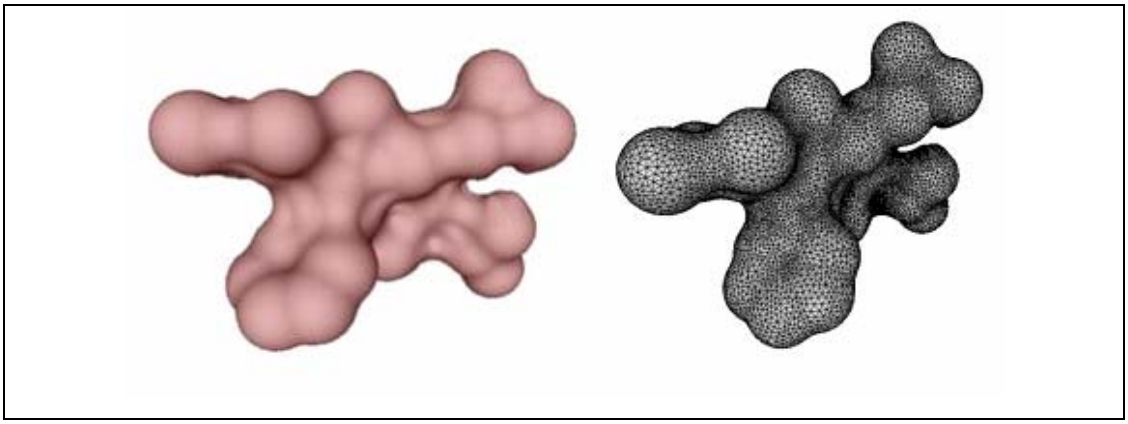


Figure 19. Molecular surface

The induced charge density of the molecular surface can be computed using the following equation [37]:

$$(\mathbf{I} - f\,\mathbf{K})[\sigma] = f\,[E] \tag{4}$$

where $[\sigma]$ is the column vector of charge density on each boundary element (triangular patch), $\mathbf{I}$ is the $n \times n$ identity matrix, n is the number of patches, f is a scalar depending on dielectric constants in the solute and solvent. $[E]$ is a column vector of the normal component of the electric field at each of the patch centers due to the solute charge distribution. $\mathbf{K}$ is the $n \times n$ matrix of coefficients between each pair of patches. The coefficicients in matrix $\mathbf{K}$ represent the impact of induced charge on one surface element to the other and are only based on the geometry. Details about Equation (4) are described in Zauhar and Varnek's paper [37]. Equation (4) is a linear equation system, but it is computationally difficult to solve $[\sigma]$ because $(\mathbf{I} - f\,\mathbf{K})$ is $n \times n$ matrix, where n can be hundreds of thousands.

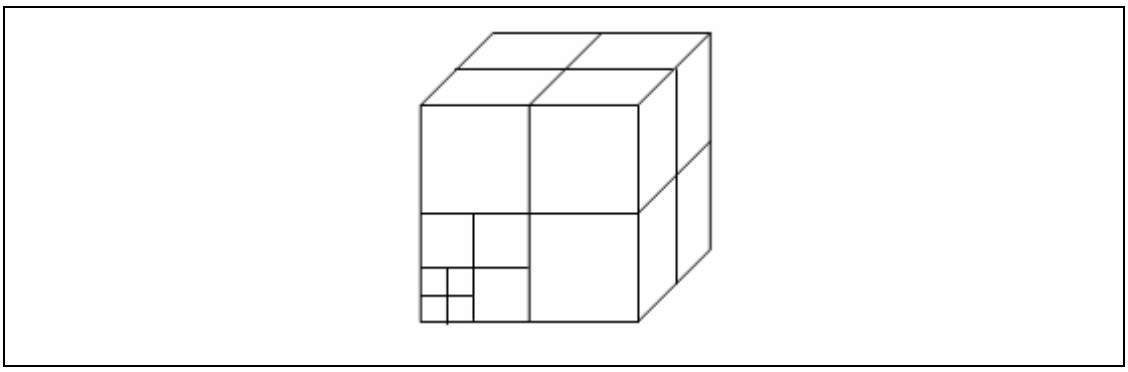


Figure 20. Cube cells

The method used to solve Equation (4) is to build a cube cell to enclose the entire molecule, and then subdivide it to make an orc-tree (Figure 20). An orc-tree is a tree whose cells have either eight children or none. Boundary elements in a leaf cell are considered near to each other; boundary elements in different leaf cells are considered far to each other. Thus, **K** can be decomposed into two components as Purisima described [23]:

$$\mathbf{K} = \mathbf{K}_{near} + \mathbf{K}_{far} \tag{5}$$

Inserting Equation (5) into Equation (4) yields,

$$(\mathbf{I} - f\,\mathbf{K}_{near})[\sigma] = f\left([E] + \mathbf{K}_{far}[\sigma]\right) \tag{6}$$

In matrix $\mathbf{K}_{far}$, coefficient between two faraway elements is approximated by considering one element and the cell in which the other element resides. If the leaf cells are smaller enough, there is a small number of pairs of near elements. Thus, $\mathbf{K}_{near}$ is a sparse matrix, and we can use well-known algorithms to solve the linear system (Equation 6) efficiently.

With induced surface charge distribution computed as described previously, electrostatic potentials at any point can be easily calculated by applying the following equation as describe by Varnek et al [35].

$$\Phi(r) = \frac{1}{D_{in}} \sum_k \frac{q_k}{|r - r_k|} + \sum_j \frac{\sigma_j A_j}{|r - r_j|} \tag{7}$$

where $\mathbf{r}_j$ is the position of the center of *j*-th patch, A_j is the area of the *j*-th patch, D_{in} is the dielectric constants in the solute, q_k and $\mathbf{r}_k$ are the partial charge and position of *k*-th atom, respectively.

5.2.2 Result

We first tested a shape of two spheres, with one +1.0 charge and one -1.0 charge placed at the center of each sphere respectively. We also tested a shape of three spheres, with two +0.5 charges and one -1.0 charge placed at the center of each sphere respectively. These results are shown in Figure 21. For the case of two spheres, a neutral (white) region is presented in between the two charges, and the potentials on the surface gradually changes from weak (lighter) to strong (darker). The case of three spheres is similar. These results are as expected and are considered correct.

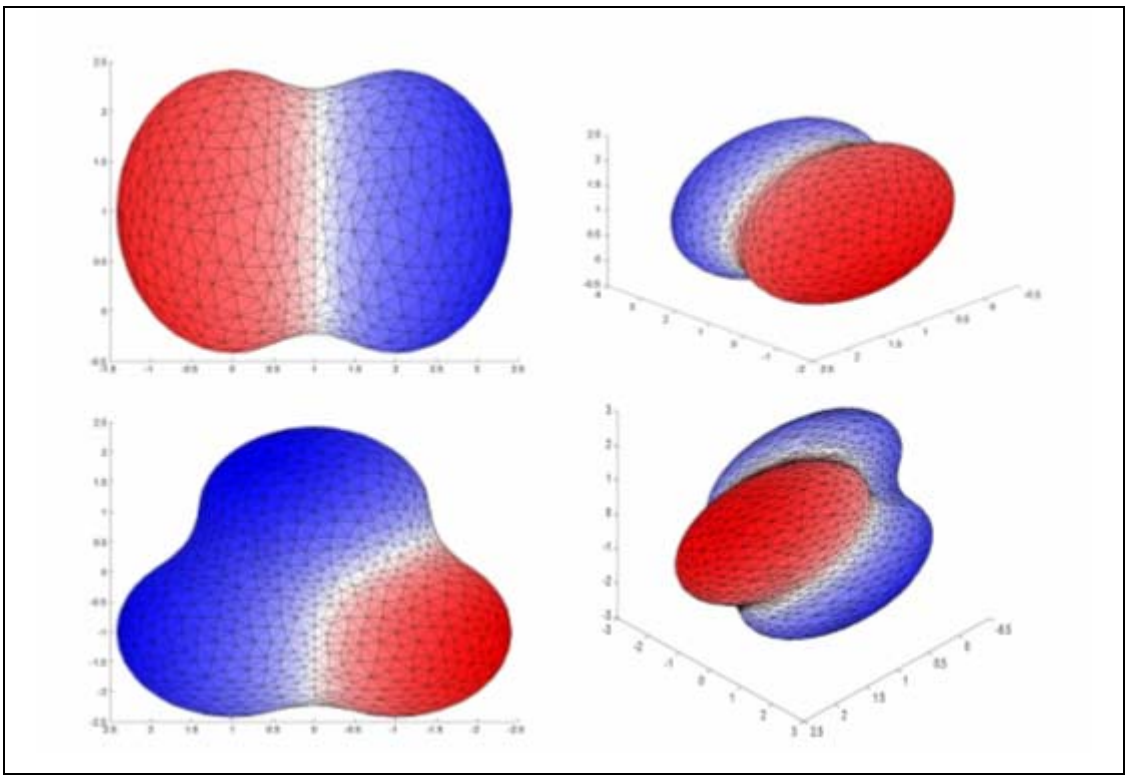


Figure 21. Surface electrostatic potentials of two-sphere shape and three-sphere shape. Negative potentials are colored red, and positive potentials are colored blue.

Table 1. Charges assigned to atoms

Atoms	Charges
Terminal-N	1.0
Terminal-O	-1.0
N	0.5
O	-0.5
C	0.0

Table 2. Performance of the tests

Protein	Number of atoms	Number of surface patches	CPU Time
7tmn	33	30520	4.5 hours
Crambin	678	75780	7.8 hours
Thrombin	5867	72104	8.8 hours

Further tests are carried on three real protein molecules, 7tmn, Crambin and Thrombin. Charges are assigned to the atoms of protein according to Table 1 [12]. The results are shown in Figure 22, 23, 24.

It is hard to evaluate the correctness of surface electrostatic potentials. However, compared with space filling model, we can see that all results approximately reflect the charge distribution of the molecule. The space filling model uses van der Waals surface, and atoms colored red are oxygens of negative charge and those colored blue are nitrogens of positive charge. Since we use molecular skin surface, the result surface model looks slightly different from the space filling model.

Table 2 shows that the computation time increases with number of boundary elements on the molecular surface. The molecular surface adopted in the test is refined and contains a large number of triangle patches. This helps to model the electrostatic potential more accurately but also increases the computational cost considerably.

There are possibilities for further optimization of the method. For example, instead of taking surface patches as boundary elements, we could consider the points (corners of the patch). Usually the number of points is much smaller than the number of patches.

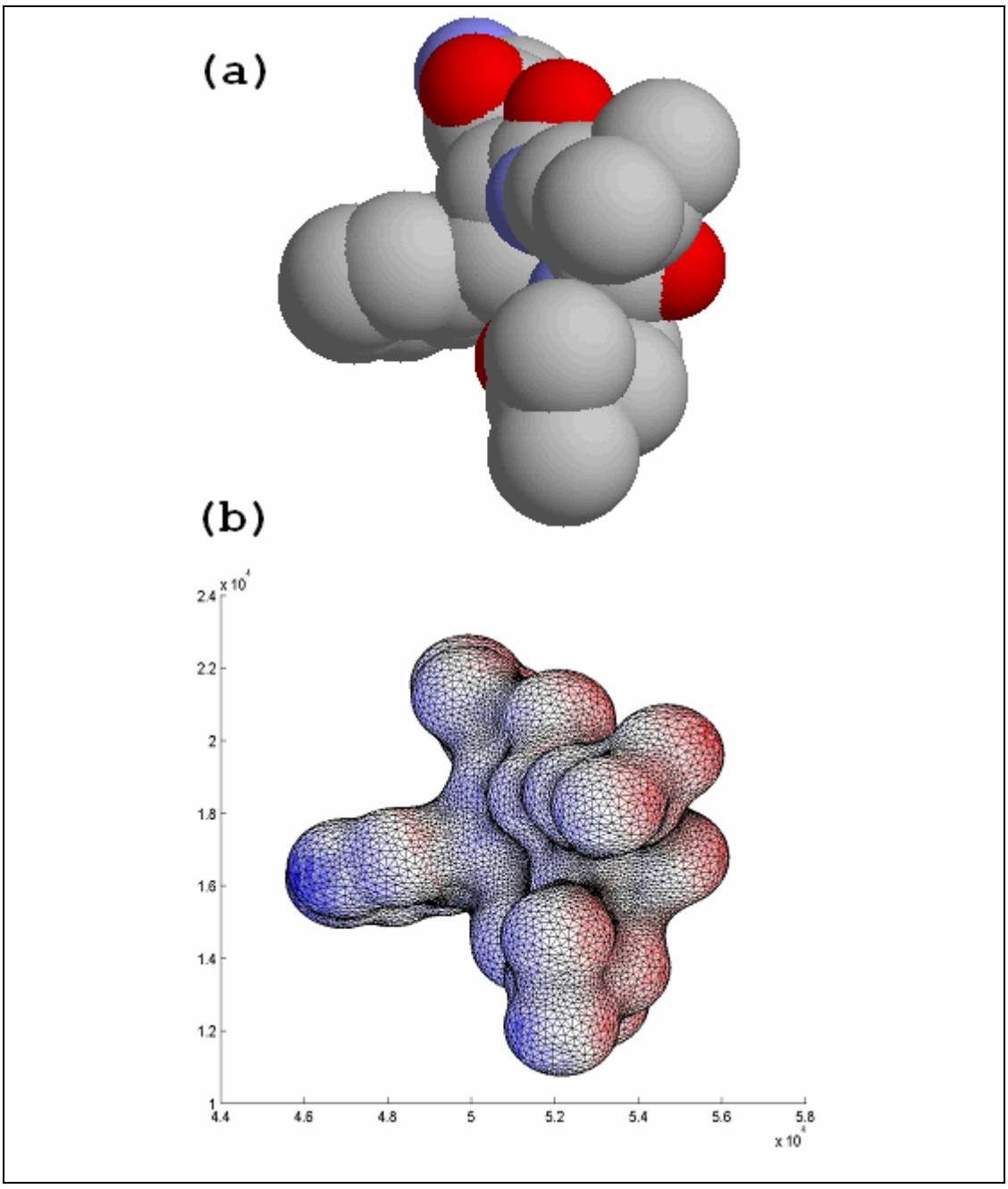


Figure 22. (a) Space filling model of 7tmn. Atoms colored red are oxygens of negative charge and those colored blue are nitrogens of positive charge. (b) The surface electrostatic potentials of 7tmn. Negative potentials are colored red, and positive potentials are colored blue.

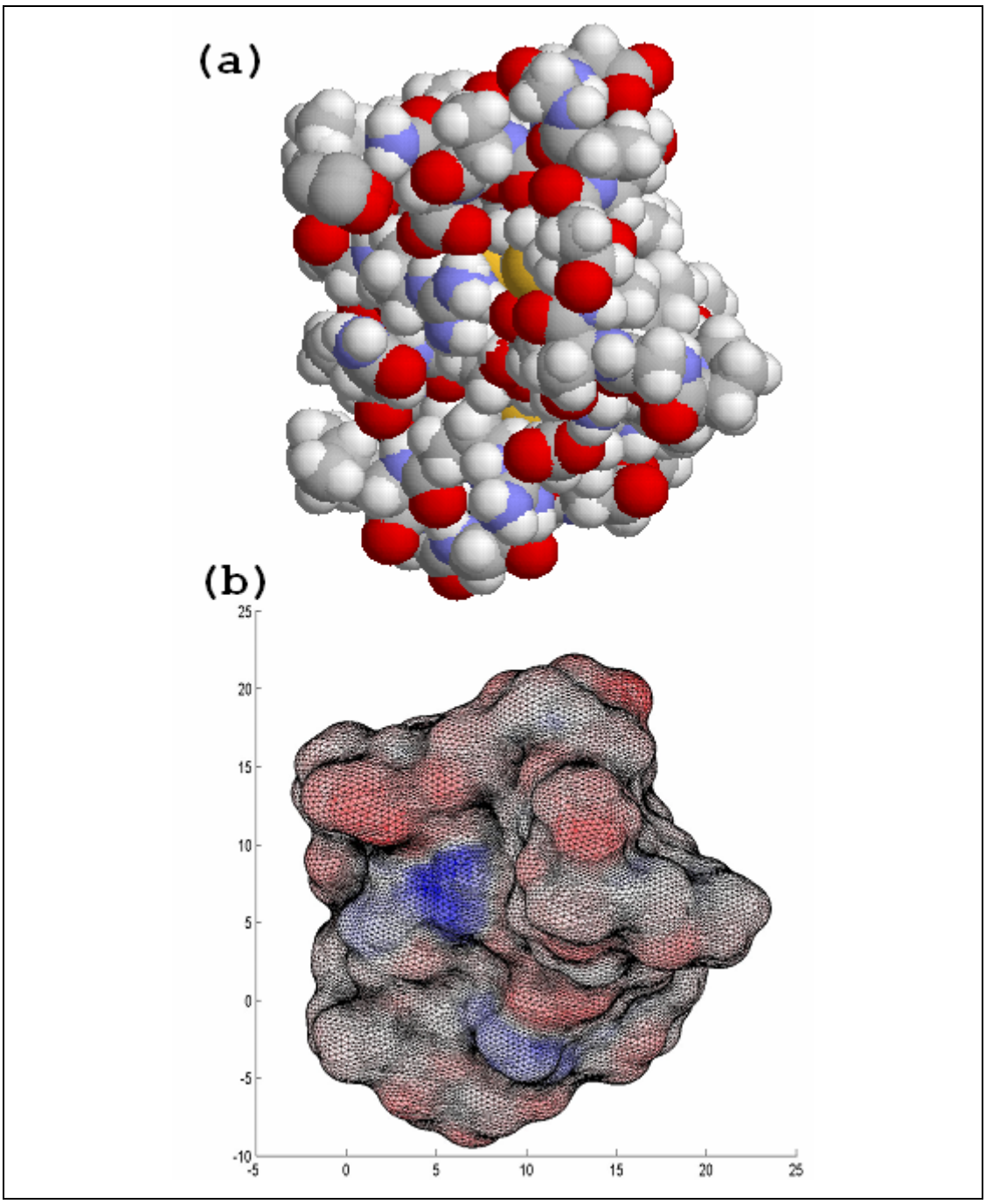


Figure 23. (a) Space filling model of Crambin. Atoms colored red are oxygens of negative charge and those colored blue are nitrogens of positive charge. (b) The surface electrostatic potentials of Crambin. Negative potentials are colored red, and positive potentials are colored blue.

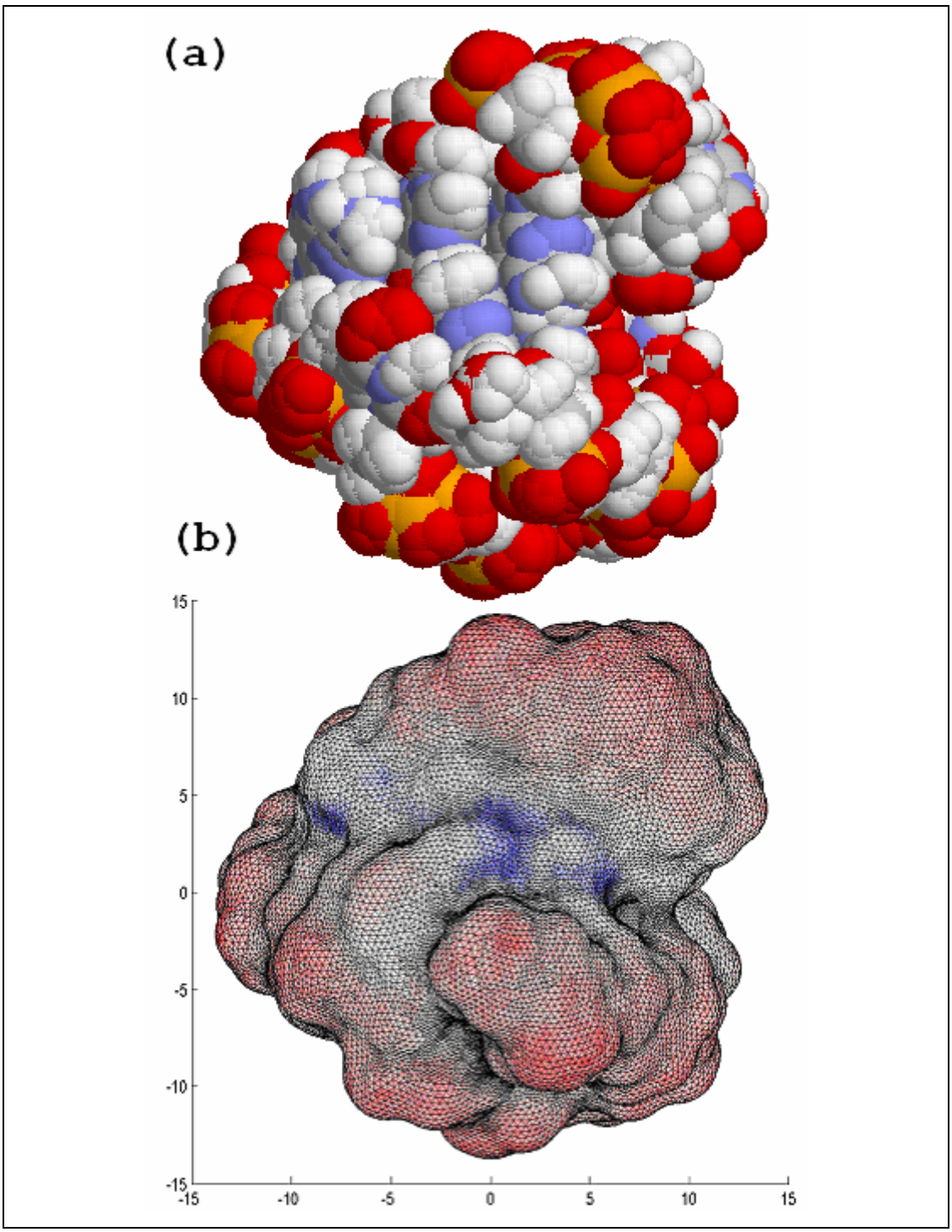


Figure 24. (a) Space filling model of *thrombin*. Atoms colored red are oxygens of negative charge and those colored blue are nitrogens of positive charge. (b) The surface electrostatic potentials of *thrombin*. Negative potentials are colored red, and positive potentials are colored blue.

6 Conclusions

The protein docking problem is a difficult computational problem. In the past decade, many methods have been proposed, and significant progress has been made. Rigid body protein docking is relatively simple and it is considered to be solved quite successfully. However, rigid body docking is not useful in the practice, because proteins usually undertake conformational changes when they are bind to the partner.

Semi-flexible and flexible docking problems are more difficult as the search space increases dramatically. Many methods have been proposed to reduce the search space or add flexibility into rigid body docking. The results of those methods may be acceptable for a small set of test cases; however, it is still far from using in practice.

Another bottleneck is the lack of selective and efficient scoring function. Scoring function is crucial in the protein docking problem as it evaluates all the possible solutions and selects best solutions as results. Many elements could be included in the scoring function; however, the problem is how to combine and use them in the most general cases.

Protein docking problem is an open and prominent area of research. There are many possible research topics that are important and difficult. Progress in the area will have great impact on the life science.

References

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

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

[3] Caflisch A, Fischer S, and Karplus M. Docking by Monte Carlo Minimization with a Solvation Correction: Application to an FKBP-Substrate Complex. *J Comput Chem*, 1997, 18:723-743.

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

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

[6] Catmull E and Clark J. Recursively generated B-spline surfaces on arbitrary topological surfaces. *Computer-Aided Design*, 1978, 10(6):350-355.

[7] Cheng H, Dey T K, Edelsbrunner H and Sullivan J. Dynamic Skin Triangulation. *Discrete Computational Geometry*, 2001, 25: 525-568.

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

[9] Edelsbrunner H. Smooth Surfaces for Multiscale Shape Representation. *Proc Sympos Found Software Techn Theoret Comput Sci*, 1995, 391-412.

[10] Edelsbrunner H. Deformable Smooth Surface Design. *Discrete Comput Geom*, 1999, 21:87-115.

[11] Fraga S, Parker J M, and Pocok J M. Computer simulations of protein structures and interactions. New York: Springer Verlag, 1995, p.2081.

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

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

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

[15] Hoffmann D, Kramer B, Washio T, Steinmetzer T, Rarey M, Lengauer T. Two-stage method for protein-ligand docking. *J Med Chem*, 1999, 42:4422–4433.

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

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

[18] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A, Aflalo C, and Vakser I. Molecular surface recognition: determination of geometric fit between protein and their ligands by correlation techniques. *Proc National Academic Science*, USA, 1992, 89:2195–2199.

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

[20] Morris G M, Goodsell D S, Halliday R S, Huey R, Hart W E, Belew R K, and Olson A J. Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function. *J Comput Chem*, 1998, 19:1639-1662.

[21] Oshiro C M, Kuntz I D, and Dixon J S. Flexible ligand docking using a genetic algorithm. *J Comput-Aided Mol Design*, 1995, 9:113-130.

[22] Pearlman D A and Charifson P S. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *J Med Chem*, 2001, 44:3417–3423.

[23] Purisima E O. Fast summation boundary element method for calculating salvation free energies of macromolecules. *J Comput Chem*, 1998, 19:1494-1504.

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

[25] Sandak B, Nussinov R, and Wolfson H J. An automated computer vision and robotics based technique for 3-D flexible biomolecular docking and matching. *Comp Appl BioSci*, 1995, 11:87–99.

[26] Sandak B, Wolfson H J, and Nussinov R. Hinge-bending at molecular interfaces: Automated docking of a dihydroxyethylene-containing inhibitor of the HIV-1 protease. *J Biomol Struct Dyn, Proceedings of the Ninth Conversation*, Sarma RH, Sarma MH, editors. New York: Adenine Press, 1996, 1:233–252.

[27] Sandak B, Nussinov R, and Wolfson H J. Docking of conformationally flexible proteins. *Seventh Symposium on Combinatorial Pattern Matching*, Laguna Beach, California. Lecture Notes in Computer Science. New York: Springer Verlag 1996, 1075:271–287.

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

[29] Sandak B, Nussinov R, and Wolfson H J. A Method for biomolecular structural recognition and docking allowing conformational flexibility. *J Comput Biol*, 1999, 5:631–654.

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

[31] Singh A P, Latombe J C, and Brutlag D L. A motion planning approach to flexible ligand binding. *Proceedings of the 7th Conference on Intelligent Systems in Molecular Biology (ISMB)*. Menlo Park, CA: AAAI Press, 1999, p 252–261

[32] Teschner M, Heidelberger B, Muller M, Pomeranets D, and Gross M. Optimized spatial hashing for collision detection of deformable objects, *Proc. Vision, Modeling, Visualization VMV'03*, 2003, pp. 47-54.

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

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

[35] Varnek A, Wipff G, Glebov A S and Feil D. An application of the Miertus-Scrocco-Tomasi salvation model in molecular mechanics and dynamics simulations. *J Comput Chem*, 1995, 16:1-19.

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

[37] Zauhar R J and Varnek A. A fast and space-efficient boundary element method for computing electrostatic and hydration effects in large molecules. *J Comput Chem*, 1995, 17:864-877.