2010

Calculations of protein-protein interactions with the fast multipole method

Bongkeun Kim
Iowa State University

Calculations of protein-protein interactions with the fast multipole method

by

Bongkeun Kim

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Physical Chemistry

Program of Study Committee:
Xueyu Song, Major Professor
Mark S. Gordon
Robert Jernigan
Mei Hong
Theresa Windus

Iowa State University

Ames, Iowa

2010

DEDICATION

To my wonderful wife,

Kyoungmin Roh,

who lifted me up, prayed for me

and made all of this possible,

for her endless cheering and patience.

And also to

my family and my friends,

for their advices and guidance.

Finally to God,

for the wisdom, grace and endurance bestowed upon me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to have this opportunity to express my thanks to those who helped me with many aspects of conducting research and writing of this dissertation. First, Dr. Xueyu Song for his advice, patience and support throughout my whole research and writings. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Mark Gordon, Dr. Robert Jernigan, Dr. Therasa Windus and Dr. Mei Hong. I would also like to thank Dr. Jiming Song for his guidance and advice for building algorithms of my research works, and Dr. Burnett for his guidance of my teaching style.

ABSTRACT

I present a physical model to calculate protein-protein interactions. General formulations to calculate the electrostatic and the van der Waals free energies are brought by the boundary element method of solving linearized Poission-Boltzmann equation in an electrolyte solution, then further expanded to the application of the Fast Multipole Method(FMM). We built an efficient solver to investigate how the mutations on the active site of the protein-protein interface affect changes in binding affinities of protein complexes. Calculated results in addition to the structural analysis help us to understand the protein-protein interaction energy and provide a model to the important applications such as protein crystallization. The osmotic second virial coefficient B_2 is directly related to the solubility of protein molecule in electrolyte solution and determined by molecular interactions involving both solvent and solute molecules. Calculations of interaction energies account for the electrostatic and the van der Waals interactions with the structural anisotropic properties of protein molecules. The orientation dependence of interaction energies between two proteins is determined by the crystal space operations and small number of protein-protein pair configurations according to the anisotropic patch model are required to calculate B_2. With the extended FMMs, double-tree and single-tree algorithms, the boundary element formulations of interaction energies can be applied with low computational cost to the proteins. B_2 Calculations of Bovine Pancreatic Trypsin Inhibitor are firstly performed to validate our model and the results of lysozyme protein under different salts, concentrations, pH and temperatures are correlated to the experimental B_2. The reduced number of pair interaction energies between

two proteins are interpolated to predict all pair interaction energies in the patch model as a precursor of the protein phase diagram calculation.

CHAPTER 1. Introduction

1.1 Background

To understand the structure of a protein by either X-ray crystallography or NMR the production of diffraction quality crystals is important. According to Chayen and Saridakis (Chayen and Saridakis, 2008) and Table 1.1 even though the purified proteins are successfully obtained, only 18% of them can be crystallized with suitable quality. Thus this crystallization process is still a bottleneck for all steps in protein structure studies. In order to achieve the successful rate of the protein crystallization, we need to screen the optimal solution condition from the traditional extensive trial-and-error screening. The choice of pH, buffer, temperature, salt concentration and precipitating agents should be guided to narrow down the large set of possibilities.

The correlation between slightly negative second virial coefficient of a protein solution and its successful crystallization condition is observed by George and Wilson (George et al., 1997). There is also a correlation between the solubility of a protein in an electrolyte solution and the osmotic second virial coefficient B_2 of the solution (Veesler et al., 1996; Boistelle et al., 1997). Observing the second virial coefficients with various solution conditions can narrow down the large set of parameters to guide the protein crystallization to the optimal values. The calculation of the osmotic second virial coefficient of a protein in an electrolyte solution requires computation of the pair interaction energy between two protein molecules. The computed pair interaction energy is also useful to calculate the phase diagram of a protein as a first step of this study.

Table 1.1 Statistics summary report of structural genomics projects. The numbers of targets and normalized percentage to Cloned step show the result of structural genomics projects until April 19 2010. The TargetDB website(targetdb.pdb.org) updates the current information providing the production of structures.

Step	Number of targets	% Normalized
Cloned	176,710	100
Expressed	123,905	70.1
Soluble	47,572	26.9
Purified	43,609	24.7
Diffracting Crystal	7,708	4.4
Structure Defined	7,265	4.2
In PDB	7,569	4.3

Our early studies (Song, 2003; Song and Zhao, 2004) of calculations of the electrostatic interaction energy and the val der Waals interacntion energy based on the conventional Boundary Element Method(BEM) require too much computational cost, both memory and time consuming. Only a small protein molecule can be used because of this cost problem. To avoid the high cost demand of our solver and to apply it to the various protein interaction system, we implement the Fast Multipole Method(FMM) algorithm to our BEM solution. With the application of FMM, we are able to build a model to compute the binding affinities of large protein complexes and the second virial coefficients of protein molecules with various solution conditions. Finally, this method is applied to the study of predicting a phase diagram of a protein.

1.2 Organization of Thesis

The organization of this thesis follows:

Chapter 2

In chapter 2, we build a residual model to calculate the electrostatic and the van der Waals contribution to the binding affinity of a protein complex. To avoid its high

computing cost depending the number of discretized surface elements on a protein surface, the Fast Multipole Method(FMM) is applied to the Boundary Element Method(BEM) to calculate the interaction energy of a protein in an electrolyte solution. The changes in binding affinities between the wild-type complex and the P_1 mutant complexes made by Swiss PDB viewer and molecular dynamics simulation are compared to the experimental data. For calculations, Bovine pancreatic trypsin inhibitor(BPTI)-trypsin, barnase-barstar and Streptomyces griseus protease B(SGPB)-turkey ovomucoid third domain(OMTKY3) complexes are used.

Chapter 3

In chapter 3, we expand the residual model of a single protein to the anisotropic patch model to compute the effective interaction energy between two protein molecules whose orientations are defined by two nearest patches on the center-to-center displacement. Two extended FMM algorithms, double-tree FMM and single-tree FMM, are applied to calculate the electrostatic and van der Waals interaction energies between two proteins. With this model, the second virial coefficients of a BPTI protein in sodium chloride solution are first calculated to validate our anisotropic patch model. The second virial coefficients of a lysozyme protein in various salt, concentration, pH and temperature are calculated and compared to the experiments.

Chapter 4

In chapter 4, the anisotropic patch model to calculate the interaction energies between two proteins is used to compute the pair potential as a starting point of predicting the phase diagram of a protein. The pair interaction potentials between many orientations of two proteins defined by the surface patches are interpolated from the small number of calculated pair interactions.

Chapter 5

In chapter 5, we describe a general idea of how the FMM algorithm can accelerate to solve a system of linear equations and reduce the computational cost without storing matrix elements produced by BEM. Applications of the FMM to our systems of linear equations are described with step-by-step procedures and the comprehensive study of reduced cost is proven.

Chapter 6

In chapter 6, we state general conclusions.

Appendix

In the appendix, the analytic expression of the electrostatic interaction free energy based on the Derjaguin-Landau-Verwey-Overbeek(DLVO) theory with two charged identical spheres in electrolyte solution is derived to validate our solution of the electrostatic interaction energy calculation between two proteins.

CHAPTER 2. Calculations of the Binding Affinities of Protein-Protein Complexes with the Fast Multipole Method

2.1 Abstract

In this paper, we present a simple physical model to calculate binding free energies of protein-protein complexes. General formulations to calculate the electrostatic free energy and the van der Waals free energy are brought by the boundary element method of solving a linearized Poission-Boltzmann equation in an electrolyte solution, then further expanded to the application of the fast multipole method to reduce the computational cost. The residual model with the fast multipole method allows us to build an efficient solver to investigate how the mutations on the active site of the protein-protein interface affect changes in binding affinities of protein complexes. The calculated results in addition to the structural analysis help us to understand the dominant contribution to the protein-protein interaction free energy and provide a model to important applications such as protein crystallization.

2.2 Introduction

The atomic resolved structure of proteins from X-ray crystallography relies on the production of diffraction quality crystals. Recent extensive studies from structural genomic project clearly indicates that even though purified proteins can be successfully obtained, only about 16% of them are crystallized with suitable quality for diffrac-

tion (TargetDB, 2010). Thus the crystallization process is still the bottleneck for protein structure determination using crystallography. Currently protein crystallization conditions are screened from traditional trial-and-error procedure. Namely, the optimal condition is obtained from extensive searches of a large parameter space of protein solutions such as pH, buffer, temperature, salt concentration and precipitating agent. Even with remarkable progress of automation the poor successful rate of obtaining protein single crystals is in no small part due to the lack of understanding of the crystallization process (Service, 2005). For example, recent studies using time-controlled micro-fluidic seeding heavily rely on knowledge of solution conditions during the nucleation stage and crystal growth stage (Gerdts et al., 2006). In other recent high-throughput experimental studies using micro-fluidics, it is clearly demonstrated that knowledge of the phase behavior of a protein allows one to create a rational screen that increases the success rate of crystallization of challenging proteins (Anderson et al., 2006). It is therefore useful to understand what kind of solution conditions might lead toward the optimal crystallization conditions and why.

As a first step toward a reliable and practical theory of protein crystallization, a realistic model of protein-protein interaction needs to be developed. In general, there are two types of protein-protein interactions in nature. One type of interactions is responsible for the protein-protein recognition to perform specific biological functions. In this case there are complimentary regions on both proteins to recognize each other and hydrophobicity is the dominating factor (Elcock et al., 2001). On the other hand, the protein-protein interaction in the protein crystallization does not necessarily involve complimentary regions to establish protein-protein contacts. For example, we have analyzed the protein contacts for five lysozyme protein crystals from the protein data bank under five different crystallization conditions (Song, 2002b). What we found is that the protein contacts can be formed from different parts of lysozyme surface residues depending upon the solution conditions. Similar conclusions can be drawn from other studies as well. For

example, Crosio et al. found that pancreatic ribonuclease uses nearly the entire protein surface residues to establish crystal-packing contacts under various crystallization conditions (Crosio et al., 1992). An extensive analysis on 78 protein crystals indicates that the amino acid composition involved in the protein contacts is indistinguishable from that of the protein surface accessible to the solvent (Carugo and Argos, 1997). These studies also suggest that crystal-packing contacts formed are sensitive to the solution conditions in contrast to the type of protein-protein interaction where hydrophobic residues are favored (Janin and Rodier, 1995; Dasgupta et al., 1997). Therefore a universal model to capture the effective interaction between protein molecules in solutions can be developed based upon the Derjaguin-Landau-Verwey-Overbeek (DLVO) picture given the protein-protein interaction at short range can be accounted for appropriately since the DLVO picture will fail when two protein molecules are separated by several solvent molecular layers.

The key feature of such a model is that it should be based on the generic properties of twenty amino acids in nature and experimentally accessible properties of electrolyte solutions and crystallization agents, and is therefore portable to all of the protein-protein interactions in aqueous solutions. Our recent studies (Song, 2003; Song and Zhao, 2004) are such efforts toward this goal.

In this model, each residue of a protein is represented by a sphere located at the geometric center of the residue determined by its native or approximate structure. The diameter of the sphere is determined by the molecular volume of the residue in solution environment (Zamyatnin, 1984). The molecular surface of our model protein is defined as the Richard-Connolly surface spanned by the union of these residue spheres using MSMS program (Sanner, 1996). Each residue carries a permanent dipole moment located at the center of its sphere and the direction of the dipole is given by the amino acid type from a protein's native structure. If a residue is charged, the amount of charge is given by the Henderson-Hasselbalch equation using the generic pK_a values of residues,

thus the local environmental effects on pK_a values are neglected. For each residue there is also a polarizable dipole at the center of the sphere, whose nuclear polarizability had been determined from our recent work (Song, 2002a) and the electronic polarizability is estimated from optical dielectric constants augmented with quantum chemistry calculations (Millefiori et al., 2008). There are three kinds of interactions in this model: the electrostatic interaction due to electric double layer effect, the van der Waals attraction due to the polarizable dipoles and a short range correction term to account for the short range interactions such as the desolvation energy, hydrophobic interaction and so on. In this report, we consider the electrostatic interactions which give the most contribution to the protein-protein interaction (Dong and Zhou, 2006; Brock et al., 2007), and the van der Waals interactions, which are the major contributors to the binding affinity calculations.

The electrostatic problem in the electrostatic interaction and the van der Waals interaction is solved using the Poisson-Boltzmann equation where the realistic shape of protein molecules are considered. The boundary element method (BEM) in combination with the fast multipole method (Greengard, 1988; Greengard and Rokhlin, 1997) is implemented to circumvent the extensive memory requirements similar to the recent work (Lu et al., 2007). In order to test the validity of our model, the binding affinities of several protein-protein complexes are calculated using our model and direct comparisons are made against experimental measurements. Reasonable agreements from these comparisons provide the first concrete evidence that our model can be used as a universal model for studies of non-specific protein-protein interactions in aqueous solutions.

2.3 Theoretical developments

2.3.1 The Statistical Thermodynamics of Binding Affinities

To set up the computational framework for calculating the binding affinities, the statistical thermodynamic analysis of "double-decoupling" method from Gilson and his coworkers (Gilson et al., 1997) is used. This approach is based on the change in the free energy of protein-protein binding when one protein and the other react and, then a single complex is formed. The final result for the binding affinity is given as,

$$\Delta G^{\circ} = \Delta G^{\circ}_{\mathrm{sol}}\left(\mathrm{AB}\right) - \Delta G^{\circ}_{\mathrm{sol}}\left(\mathrm{A}\right) - \Delta G^{\circ}_{\mathrm{sol}}\left(\mathrm{B}\right) \quad , \tag{2.1}$$

where ΔG°_{sol} is the free energy change when a molecule is introduced into a solution from vacuum. In this report, the binding free energy calculations are for single mutations at the binding site, thus free energy change due to translational, rotational and vibrational contributions of the proteins upon binding remains relatively constant. In our model the free energy change is the sum of the electrostatic solvation energy and the van der Waals solvation energy of the molecule. For protein A,

$$\Delta G^{\circ}_{\mathrm{sol}}\left(\mathrm{A}\right) = \Delta G^{\mathrm{elec}}_{\mathrm{sol}}\left(\mathrm{A}\right) + \Delta G^{\mathrm{vdw}}_{\mathrm{sol}}\left(\mathrm{A}\right). \tag{2.2}$$

2.3.2 The electrostatic solvation energy calculation

The electrostatic binding free energy between the protein A and B is defined as,

$$\begin{aligned} \Delta G^{\mathrm{binding}}_{\mathrm{elec}} &= \Delta G_{\mathrm{elec}}\left(\mathrm{AB}\right) \\ &- \Delta G_{\mathrm{elec}}\left(\mathrm{A}\right) - \Delta G_{\mathrm{elec}}\left(\mathrm{B}\right). \end{aligned} \tag{2.3}$$

The electrostatic interaction is estimated from the Poisson-Boltmann(PB) equation. To solve the PB equation, we use the boundary element method based on the integral equa-

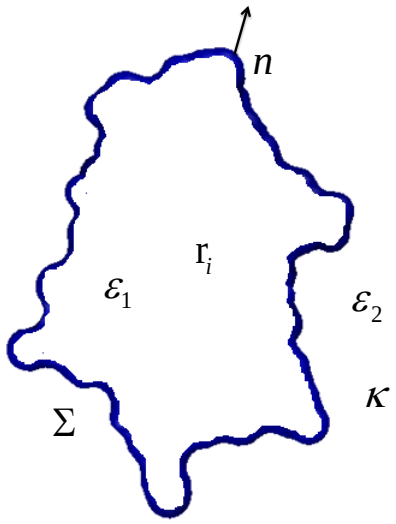


Figure 2.1 Schematic illustration showing the electrostatic formulation of single protein. $\sum$ is the molecular surface of a protein, n is outward unit normal, ε_1 and ε_2 are dielectric constant inside the cavity and outside solvent respectively. κ is the inverse Debye screening length for the electrolyte solution. r_i stands for residue center and charge q_i and dipole μ_i are located on each residue center.

tion formulation of the linearized PB equation for a single protein (Yoon and Lenhoff, 1990; Juffer et al., 1991).

Consider the molecular surface $\sum$ which covers a protein molecule. There are N charges q_i and dipoles $\vec{\mu_i}$ at position $\mathbf{r_i}$ enclosed by the surface $\sum$. Inside this dielectric cavity the dielectric constant is ε_1 and the dielectric constant of the solution is ε_2(see Figure 2.1). The inverse Debye screening length κ is given by the solution's ionic strength. The integral equations for the potential $\varphi(\mathbf{r})$ and its gradient $\partial\varphi(\mathbf{r})/\partial n$ are given by the following integral equations (Song, 2003; Juffer et al., 1991),

$$\frac{1}{2}\left(1+\frac{\varepsilon_2}{\varepsilon_1}\right)\varphi(\mathbf{r}_0)+\iint_{\sum} L_1(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r})\mathrm{d}\mathbf{r}$$
$$+\iint_{\sum} L_2(\mathbf{r},\mathbf{r}_0)\frac{\partial\varphi(\mathbf{r})}{\partial n}d\mathbf{r}$$
$$=\sum_{i=1}^{N}\left\{q_i F(\mathbf{r_i},\mathbf{r}_0)+\vec{\mu_i}\cdot\nabla F(\mathbf{r_i},\mathbf{r}_0)\right\}/\varepsilon_1, \tag{2.4}$$

$$\frac{1}{2}\left(1+\frac{\varepsilon_1}{\varepsilon_2}\right)\frac{\partial\varphi(\mathbf{r}_0)}{\partial n}+\iint_{\sum} L_3(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r})\mathrm{d}\mathbf{r}$$
$$+\iint_{\sum} L_4(\mathbf{r},\mathbf{r}_0)\frac{\partial\varphi(\mathbf{r})}{\partial n}d\mathbf{r}$$
$$=\sum_{i=1}^{N}\left\{q_i\frac{\partial F}{\partial n_0}(\mathbf{r_i},\mathbf{r}_0)+\vec{\mu_i}\cdot\nabla\frac{\partial F}{\partial n_0}(\mathbf{r_i},\mathbf{r}_0)\right\}/\varepsilon_1, \tag{2.5}$$

where

$$L_1(\mathbf{r},\mathbf{r}_0) = \tfrac{\partial F}{\partial n}(\mathbf{r},\mathbf{r}_0)-\tfrac{\varepsilon_2}{\varepsilon_1}\tfrac{\partial P}{\partial n}(\mathbf{r},\mathbf{r}_0), \tag{2.6}$$

$$L_2(\mathbf{r},\mathbf{r}_0) = P(\mathbf{r},\mathbf{r}_0)-F(\mathbf{r},\mathbf{r}_0), \tag{2.7}$$

$$L_3(\mathbf{r},\mathbf{r}_0) = \tfrac{\partial^2 F}{\partial n_0\partial n}(\mathbf{r},\mathbf{r}_0)-\tfrac{\partial^2 P}{\partial n_0\partial n}(\mathbf{r},\mathbf{r}_0), \tag{2.8}$$

$$L_4(\mathbf{r},\mathbf{r}_0) = -\tfrac{\partial F}{\partial n_0}(\mathbf{r},\mathbf{r}_0)+\tfrac{\partial P}{\partial n_0}(\mathbf{r},\mathbf{r}_0)\tfrac{\varepsilon_1}{\varepsilon_2} \tag{2.9}$$

and

$$F(\mathbf{r},\mathbf{r}_0) = \tfrac{1}{4\pi|\mathbf{r}-\mathbf{r}_0|}, \tag{2.10}$$

$$P(\mathbf{r},\mathbf{r}_0) = \tfrac{e^{-\kappa|r-\mathbf{r}_0|}}{4\pi|\mathbf{r}-\mathbf{r}_0|}. \tag{2.11}$$

Although the traditional boundary element method such as Atkinson and his coworkers' (Atkinson and Han, 2009) can be used to solve above integral equations, the memory requirement is too costly even on the newest computers using either a direct linear system solver or iterative solver, such as Generalized minimal residual method (GMRES) (Barrett et al., 1994) for a moderate size protein. In the current work the fast multipole method is implemented and the details will be outlined in chapter 5. Once the above

integral equations are solved the potential inside the dielectric cavity is,

$$\begin{aligned}\varphi(\mathbf{r}_0) = &- \iint_{\sum} L_1(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r})\mathrm{d}\mathbf{r} \\ &- \iint_{\sum} L_2(\mathbf{r},\mathbf{r}_0)\frac{\partial\varphi(\mathbf{r})}{\partial n}\mathrm{d}\mathbf{r}, \qquad (2.12)\\ \nabla_0\varphi(\mathbf{r}_0) = &- \iint_{\sum} \nabla_0 L_1(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r})\mathrm{d}\mathbf{r} \\ &- \iint_{\sum} \nabla_0 L_2(\mathbf{r},\mathbf{r}_0)\frac{\partial\varphi(\mathbf{r})}{\partial n}\mathrm{d}\mathbf{r}. \qquad (2.13)\end{aligned}$$

Finally, the electrostatic solvation free energy is given by,

$$\Delta G_{\mathrm{ele}} = \sum_{i=1}^{N}\left\{\frac{q_i}{\epsilon_1}\varphi(\mathbf{r_i}) + \frac{1}{\epsilon_1}\vec{\mu_i}\cdot\nabla\varphi(\mathbf{r_i})\right\}. \qquad (2.14)$$

2.3.3 The van der Waals energy contribution

The van der Waals binding free energy between proteins A and B is defined as in Eq. (2.15),

$$\begin{aligned}\Delta G_{\mathrm{vdw}}^{\mathrm{binding}} &= \Delta G_{\mathrm{vdw}}\left(\mathrm{AB}\right) \\ &- \Delta G_{\mathrm{vdw}}\left(\mathrm{A}\right) - \Delta G_{\mathrm{vdw}}\left(\mathrm{B}\right). \qquad (2.15)\end{aligned}$$

Song and Zhao had developed a theory to calculate the van der Waals interaction between protein molecules in an electrolyte solution using the following effective action in Fourier space of the polarizable dipoles (Song and Zhao, 2004),

$$\begin{aligned}S\left[\mathbf{m}_{\mathbf{r},n}\right] = &-\frac{\beta}{2}\sum_{\mathbf{r}}\sum_{n=-\infty}^{n=\infty}\frac{1}{\alpha_{\mathbf{r},n}}\mathbf{m}_{\mathbf{r},n}\cdot\mathbf{m}_{\mathbf{r},-n} \\ &+\frac{\beta}{2}\sum_{\mathbf{r}\neq\mathbf{r}'}\sum_{n=-\infty}^{n=\infty}\frac{1}{\alpha_{\mathbf{r},n}}\mathbf{m}_{\mathbf{r},n}\cdot T(\mathbf{r}-\mathbf{r}')\cdot\mathbf{m}_{\mathbf{r},-n} \\ &+\frac{\beta}{2}\sum_{\mathbf{r},\mathbf{r}'}\sum_{n=-\infty}^{n=\infty}\frac{1}{\alpha_{\mathbf{r},n}}\mathbf{m}_{\mathbf{r},n}\cdot R_n(\mathbf{r}-\mathbf{r}')\cdot\mathbf{m}_{\mathbf{r},-n}, \qquad (2.16)\end{aligned}$$

where $\alpha_{\mathbf{r},n}$ is the frequency-dependent polarizability of a residue located at position $\mathbf{r}$. $T(\mathbf{r}-\mathbf{r}')$ is the dipole-dipole interaction tensor between dipoles at $\mathbf{r}$ and $\mathbf{r}'$, where the

retardation is neglected. $R_n(\mathbf{r}-\mathbf{r}')$ is the reaction field tensor at frequency $\omega_n = 2\pi n/\beta\hbar$, which captures the effect of surrounding dielectric medium. If the electrolyte solvent is treated by the Debye-Hückel theory, this reaction field tensor can be calculated by solving the PB equation with dielectric function $\epsilon(i\omega_n)$ and the Debye screening length κ (Song and Zhao, 2004). The quantum partition function from this effective action of the system is

$$Q(\mathrm{A}) = \prod_n \left[\frac{2\pi}{\beta \det\left[A_n(\mathrm{A})\right]}\right]^{1/2}, \tag{2.17}$$

where A represents the protein A, and A_n's matrix element is given by,

$$A_n(\mathbf{r},\mathbf{r}') = \frac{1}{\alpha_{\mathbf{r},n}}\delta_{\mathbf{r},\mathbf{r}'} - T(\mathbf{r}-\mathbf{r}') - R_n(\mathbf{r}-\mathbf{r}'). \tag{2.18}$$

The symbol "det" represents the determinant of the matrix. Finally the van der Waals binding free energy is given by (Song and Zhao, 2004),

$$\begin{aligned}\Delta G_{\mathrm{vdw}}^{\mathrm{binding}} = \frac{1}{2}k_B T \sum_{n=-\infty}^{n=\infty} \big[&\ln\left\{\det\left[A_n\left(\mathrm{AB}\right)\right]\right\} \\ &- \ln\left\{\det\left[A_n\left(\mathrm{A}\right)\right]\right\} - \ln\left\{\det\left[A_n\left(\mathrm{B}\right)\right]\right\}\big].\end{aligned} \tag{2.19}$$

In order to evaluate the van der Waals interaction in our model, the reaction field matrix $R_n(\mathbf{r}-\mathbf{r}')$ has to be calculated with the properties of the proteins and of the solution. The boundary element formulation which is used to evaluate the electrostatic free energy can also be used to calculate the reaction field matrix. Again consider the molecular surface $\sum$ spanned by a protein molecule (Figure 2.2). There are N polarizable dipoles $\mathbf{m_r}$ at position $\mathbf{r}$ enclosed by the surface $\sum$. Inside this dielectric cavity the dielectric constant is one and the dielectric function of the solution is $\varepsilon(i\omega_n)$ at the Matsubara frequency ω_n. The inverse Debye screening length κ is given by the solution's ionic strength. If we recognize that in order to calculate the potential at the molecular surface a dipole $\mathbf{m}$ at position $\mathbf{r}_0$ can be described by an effective charge density $\rho_{\mathrm{eff}}(\mathbf{r}) = -\mathbf{m}\nabla\delta(\mathbf{r}-\mathbf{r}_0)$ (Jackson, 1999), the reaction field matrix involving residues $\mathbf{r}_i$ and $\mathbf{r}_j$ can be written as,

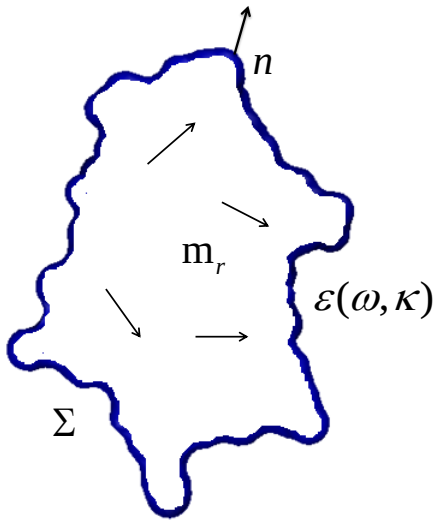


Figure 2.2 Schematic illustration showing the van der Waals energy formulation of a single protein. $\sum$ is the molecular surface of a protein, n is outward unit normal, ε is dielectric constant outside solvent. m_r stands for the polarizable dipole located on the residue center.

$$R_n(\mathbf{r}_i,\mathbf{r}_j) = \iint_{\sum} \left[\nabla_i F(\mathbf{r}_i,\mathbf{r}_j) - \nabla_i P(\mathbf{r}_i,\mathbf{r}_j)\right] \frac{\partial \varphi}{\partial n}(\mathbf{r}_j,\mathbf{r})\mathrm{d}\mathbf{r} + \iint_{\sum} \left[-\nabla_i \frac{\partial F}{\partial n_j} F(\mathbf{r}_i,\mathbf{r}_j) + \nabla_i \frac{\partial P}{\partial n_j}(\mathbf{r}_i,\mathbf{r}_j)\varepsilon\right] \varphi(\mathbf{r}_j,\mathbf{r})\mathrm{d}\mathbf{r}, \quad (2.20)$$

where F and P are defined in Eq.(2.10) and Eq. (2.11). φ and $\partial\varphi/\partial n$ can be obtained by solving the following integral equations at each frequency ω_n (Song and Zhao, 2004; Juffer et al., 1991),

$$\frac{1}{2}\left(1+\varepsilon(i\omega_n)\right)\varphi(\mathbf{r}_i,\mathbf{r}_0) + \iint_{\sum} L_1(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r}_i,\mathbf{r})\mathrm{d}\mathbf{r} + \iint_{\sum} L_2(\mathbf{r},\mathbf{r}_0)\frac{\partial \varphi}{\partial n}(\mathbf{r}_i,\mathbf{r})\mathrm{d}\mathbf{r} = \nabla_i F(\mathbf{r}_i,\mathbf{r}_0), \quad (2.21)$$

$$\frac{1}{2}\left(1+\frac{1}{\varepsilon(i\omega_n)}\right)\frac{\partial\varphi}{\partial n}(\mathbf{r}_i,\mathbf{r}_0) \quad + \iint_{\Sigma} L_3(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r}_i,\mathbf{r})\mathrm{d}\mathbf{r} + \iint_{\Sigma} L_4(\mathbf{r},\mathbf{r}_0)\frac{\partial\varphi}{\partial n}(\mathbf{r}_i,\mathbf{r})\mathrm{d}\mathbf{r} = \nabla_i\frac{\partial F}{\partial n_0}(\mathbf{r}_i,\mathbf{r}_0), \tag{2.22}$$

where L_1, L_2, L_3, and L_4 are defined in the electrostatic free energy calculation in Eqs. (2.6), (2.7), (2.8) and (2.9). To evaluate the van der Waals binding free energy in Eq. (2.19), the reaction field matrix is built using the dielectric function $\varepsilon(i\omega_n)$ for each frequency ω_n. And the polarizability of a residue in a protein is given by,

$$\alpha_n = \alpha(i\omega_n) = \frac{\alpha_{nu}}{1+\omega_n/\omega_{rot}} + \frac{\alpha_{el}}{1+(\omega_n/\omega_I)^2}, \tag{2.23}$$

where α_{nu} is the static nuclear polarizability of a residue (Song, 2002a) and ω_{rot} is a characteristic frequency of nuclear collective motion from a generalization of the Debye model. α_{el} is the static electronic polarizability of a residue and ω_I is the ionization frequency of a residue as in the Drude oscillator model of electronic polarizabilities. $\omega_{rot} = 20cm^{-1}$ for this calculation which is typical rotational frequency of molecules (Israelachvili, 1985). Other properties are listed in Table 2.1 based on the calculated result (Millefiori et al., 2008). An accurate parametrization of the dielectric function $var\epsilon(i\omega)$ of water based on the experimental data is taken from Parsegian's work (Parsegian, 1975).

2.3.4 Implementation of the Fast Multipole Method to the Boundary Element Method

The major drawback of the traditional boundary element method (BEM) is the order $O(N^2)$ dependence of the matrix size on the number of surface elements N. The large size of a matrix not only requires larger usage of memory but also takes longer time to solve the corresponding linear system. An efficient algorithm developed by Greengard and Rokhlin, the fast multipole method (FMM), is implemented to avoid storing

Table 2.1 Intrinsic nuclear polaizability(α_{nu}), electronic polarizability(α_{el}) and ionization frequency of amino acids in unit of $\mathring{A}^3$ (Millefiori et al., 2008).

Amino acid	α_{nu}	α_{el}	ω_I
Ala	2.09	8.25	75650
Arg	4.38	18.01	63880
Asn	5.05	11.66	71740
Asp	3.08	10.86	75250
Cys	2.18	11.40	70900
Gln	4.40	13.54	70450
Glu	3.10	12.79	73400
Gly	2.01	6.44	77110
His	2.53	15.14	65730
Ile	1.98	13.67	73710
Leu	1.90	13.80	74320
Lys	2.96	15.39	67510
Met	2.07	15.33	66380
Phe	2.01	18.33	67550
Pro	1.47	11.07	71340
Ser	3.52	8.94	74890
Thr	4.19	10.72	73640
Trp	3.06	23.35	58430
Tyr	3.50	19.25	63070
Val	1.99	11.82	74160

matrix elements and speed up matrix-vector multiplications which is the most time consuming step in solving linear equations (Greengard and Rokhlin, 1997). To apply the FMM algorithm to the BEM, surface elements on a protein surface are distributed to different 3D rectangular boxes at different levels based on a hierarchical oct-tree, and a divide-conquer strategy is applied to the far-field interactions at each level in the tree structure (see Figure 5.4 in chapter 5). The fundamental observation in FMM is that the multipole moment expansion of the far field interaction, which is roughly O(N^2) in the direct BEM, can be approximated by the low number of summation depending on the designated accuracy to lower computational cost. The integral elements of matrices in the electrostatic and the van der Waals interaction formulations are described by two different interactions, Columbic interaction and Debye-Hückel (screened Columbic) interaction. The detail formulations of the fast multipole method is described in chapter 5.

2.3.5 Preparation of protein complex structures

Three protein complex systems, where extensive experimental data are available, are used to test our protein-protein interaction model. The Bovine Pancreatic Trypsin Inhibitor (BPTI)-trypsin system where the crucial P_1 residue had been mutated to various residues, the binding affinities and mutated protein complex structures have been extensively documented (Krowarsch et al., 1999). The other two systems are well studied barnase-barstar complex (Schreiber et al., 1997) and the *Streptomyces griseus* proteinase B (SGPB)-turkey ovomucoid third domain complex (OMTKY3) (Lu et al., 1997).

In our preparations for the mutant structure without the experimental one, the Swiss-PDB Viewer (Guex and Peitsch, 1997) is first used to make a single mutant on the binding site and select the best rotamer based on its lowest score according to the

formula.

$$\begin{aligned} \text{score} &= (4 \times \text{NbClash with backbone N, C}\alpha \text{ and C atoms}) \\ &+ (3 \times \text{NbClash with backbone O atoms}) \\ &+ (2 \times \text{NbClash with side-chain atoms}) \\ &- \text{NbHbonds} - 4 \times \text{NbSSbonds}, \end{aligned}$$

where the "Nb" is the abbreviation of "number". Then molecular dynamics (MD) simulations using CHARMM force field (Brooks et al., 1983) are performed to determine the final mutant structure used in our calculations. In order to test the validity of the simulated mutant structures as compared to the experimental mutant structures, 10 P_1 mutants of BPTI-trypsin complexes based on the wild-type PDB (PDB ID=3BTK) are used to validate our procedure for the simulated mutant structures. For the BPTI-trypsin system, the crystal structures of complexes between Bovine $\beta-$trypsin and ten P_1 variants of BPTI (Helland et al., 1999) are known experimentally (PDB code: 3BTD, 3BTE, 3BTF, 3BTG, 3BTH, 3BTK, 3BTM, 3BTQ, 3BTT, 3BTW). The RMSD studies between simulated structures and experimental structures are within $1.3\mathring{A}$ (the average value from all 10 mutants). Figure 2.3 shows the correlation of calculated binding affinities between experimental PDB structures and simulated PDB structures. Thus, the simple mutant PDB structure from Swiss PDB viewer mutation followed by MD simulations can be used as mutant structure to calculate the binding free energy of mutant complexes. This method is used to generate all mutant structures for barnase-barstar complexes and SGPB-OMTKY3 complexes for calculations.

For the barnase-barstar complex system, the crystal structure of the pseudo wild-type barnase-barstar complex (Vaughan et al., 1999) (PDB code=1B27) is used as a template of the mutant complexes. This complex contains three sets of barnase-barstar complex, chain A is used for barnase and its binding site mutations and chain D is used for modeling the wild-type barstar. In order to make a wild-type protein, A40 and A82

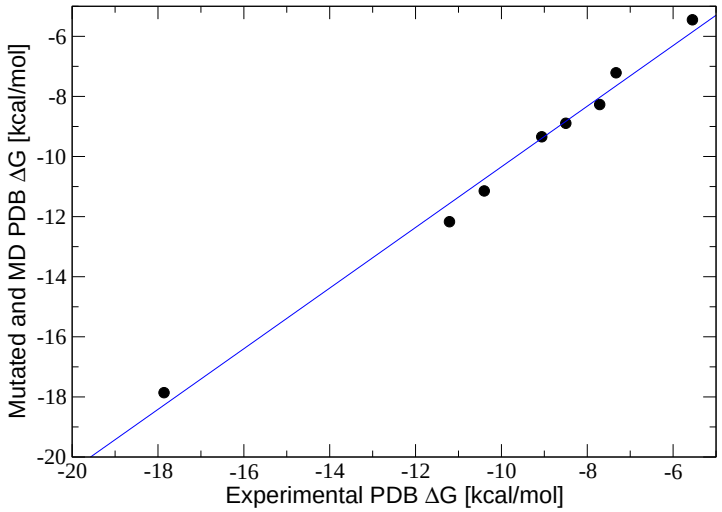


Figure 2.3 The experimental PDB versus MD simulated PDB comparison for the electrostatic binding free energy of the P_1 variants of BPTI-trypsin complexes. The linear fit correlation coefficient is 0.989.

in barstar are mutated to Cys. As the experimental results of barnase-barstar binding measurements indicate that the deletion of N-terminal Met residue in barstar, thus in our calculation the N-terminal Met is deleted in the template protein structure. So the final template has 110 barnase residues, 89 barstar residues. To make a comparison with experimental binding affinities (Schreiber et al., 1997), seven mutant complexes (Ala, Cys, Phe, Gln, Ser, Trp and Tyr) on the Glu73 residue in barnase are made by the Swiss PDB viewer and followed by MD simulations.

Finally, the crystallographic structure of the SGPB and OMTKY3 complex (Read et al., 1983) (PDB code=3SGB) is used as the wild-type template for the mutant complexes. The following experimental PDB structures, PDB code 1CSO, 1CT0, 1CT2 and 1CT4 (Bateman et al., 2000) for P_1 Ile, Ser, Thr and Val mutant complexes and PDB code 1SGP (Kui et al., 1995) for P_1 Ala mutant complex and PDB code 2NU0, 2NU1, 2NU2, 2NU3, 2SGF for P_1 Trp, His, Arg, Lys and Phe mutant complexes are already ex-

isted but these structures were not used to calculate the binding affinities of the mutant complexes because hand-mutated structures from the wild-type template are well fitted to experimental mutant PDB structures whose RMSDs with experimental PDB structures are all within 0.49$\mathring{A}$ and calculations of BPTI-trypsin complexes already show the validity of binding affinities from hand-mutated and MD simulated mutant complexes as shown on Figure 2.3. For the cross-reference from the wild-type template to the protein complexes used in the binding affinity measurements, the first 6 residues in OMTKY3 inhibitor chain on the wild-type structure, PDB code 3SGB, are deleted. The final templates for The SGPB and OMTKY3 complex contains 185 SGPB residues, 50 OMTKY3 residues.

2.4 Results and Discussion

2.4.1 Binding energy calculations of BPTI-trypsin complexes

The binding free energies of BPTI-trypsin complexes are calculated according to our model. Firstly, the binding free energy, ΔG, is calculated and the change of binding affinity from the mutation of P_1 residue, $\Delta\Delta G = \Delta G_{\text{bind}}(\text{mutant}) - \Delta G_{\text{bind}}(\text{wild-type})$, is obtained. The correlation between calculated and experimental data (Krowarsch et al., 1999). of the binding free energy, ΔG, is shown on the Figure 2.4(a) and the relation of changes in the binding free energy with a single mutation is also drawn on Figure 2.4(b) and the values are also listed in Table 2.2. In Figure 2.4(a), there are two mutants data which give the positive binding affinity (repulsion) instead of small negative affinity as the experiment shows.

The calculation of the binding free energy of BPTI-trypsin complexes shows the positive binding energy for the acidic P_1 Asp and P_1 Glu variants in BPTI-trypsin complexes. Considering the binding arrangement of the P_1-S_1 site in BPTI-trypsin complex, the electrostatic repulsion between S_1 Glu and acidic P_1 makes the binding

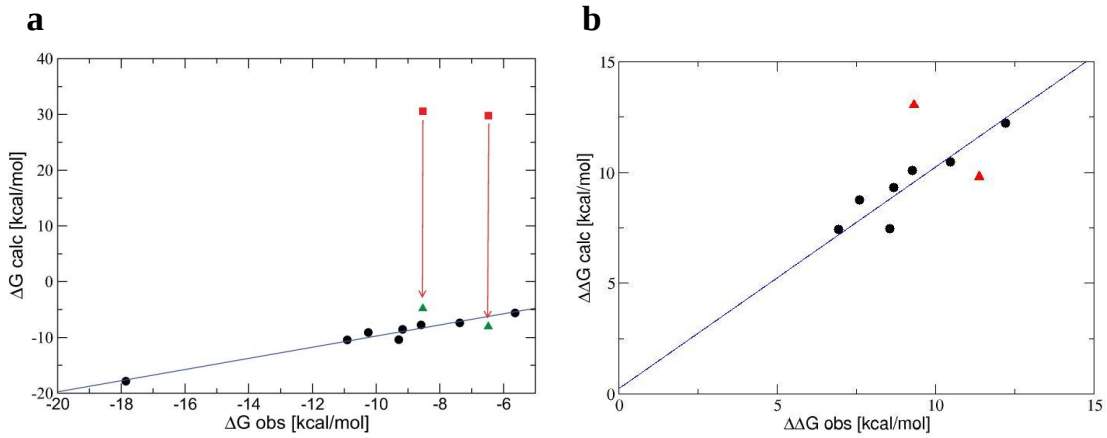


Figure 2.4 The binding free energy changes of BPTI-trypsin complexes after applying the pK_a shifts of P1 Asp and P1 Glu (a). The red square boxes represent the negative charged acidic P_1 variants and the green up-triangles do the protonated and neutral P_1 after using using PROPKA 2.0 (Delphine et al., 2008). The red arrows indicate the binding affinity shifts from positive to negative one. Y-axis is the observed (experimental) binding free energies of 10 P_1 variants (Krowarsch et al., 1999). After considering the pK_a shifts for the acidic P_1 residues, the correlation of $\Delta\Delta G$ between observed and calculated data is shown on (b). The linear fit correlation coefficient from all mutants is 0.912. The linear fit excluding the mutations making water mediated hydrogen bonds (two acidic P_1 mutants as red triangles) yields 0.982.

too unfavorable when both residues are negatively charged. The experimental result still indicates favorable binding affinities of P_1 Asp and P_1 Glu mutants, $\Delta G = -6.478$ and -8.534 respectively. The pK_a shift of P_1 Glu mutant in OMTKY3-SGPB complex from 4.46 (unbound) to 8.74 (bound) was measured experimentally (Qasim et al., 1995). By Brandsdal et al. the pK_a shift of P_1 Glu mutant in OMTKY3-SGPB complex was calculated to 13.1 and they also calculated the pK_a shift of P_1 Glu mutant of BPTI-trypsin complex upon binding from 4.3 to 14.3 (Brandsdal et al., 2006). Even though their calculations of pk_a shifts are too overestimated, we obtained an idea that acidic P_1 Glu in BPTI-trypsin complex is protonated. Therefore negative charge no more exists in the reference pH=8.3 condition. To make consistent data of pK_as in unbounded state and bounded state, we use the PROPKA 2.0 (Delphine et al., 2008) because this program is known as the most accurate one to predict the pK_a values of amino acids compared with a large data set of experimentally determined pK_as (Davies et al., 2006). For our calculations it gives pK_a values 8.7 and 8.8 for P_1 Asp and P_1 Glu mutants in BPTI-trypsin complexes and also gives pK_a values 8.8 for P_1 Glu mutant in OMTKY3-SGPB complex which is close to 8.74 from the experiment (Qasim et al., 1995). With the shifted pK_as, the calculated binding free energies are much more improved to fit to the binding affinity trend which is indicated on Figure 2.4(a). The red arrows show the binding free energy changes from the positive one (using generic pK_a) to the negative one (using pK_a that accounts for the local environments).

After considering the pK_a shifts in BPTI-trypsin compelxes, P_1 Asp and P_1 Glu participate into the correlation of changes in the binding affinities between the observed and calculated data. The overall linear fit coefficient is 0.912. But two acidic P_1 data is still relatively far from the correlation curve because without those two data the linear fit correlation coefficients improved to 0.982. The reason is from the stabilization effect of water molecules in the interface of the two proteins when binding. In our model, water molecules are not explicitly represented. Therefore the effect of the hydrogen

bonds between water molecules and the side chains of interfacial amino acids are not considered in the binding energy calculations. We already considered the pK_a shifts on BPTI P_1 Asp and P_1 Glu as a stabilization effect between P_1-S_1 binding interface. Additional stabilization effect is observed by Helland et al. The solvent molecules Sol653 and Sol654 participate into forming the hydrogen bonds with the carboxylate group of P_1 Asp and P_1 Glu and the interfacial interaction between P_1-S_1 is stabilized by the bridge-forming water molecules (Helland et al., 1999). The RMSD studies between MD simulated PDB structures and experimental structures describe this water effect. The RMSDs of C_α from all other 7 mutants except the wild-type are ranged between $1.13\mathring{A} \sim 1.31\mathring{A}$. But the RMSD of P_1 Glu is $1.53\mathring{A}$ and the RMSD of P_1 Asp is even worse $1.75\mathring{A}$. This is the indication that the simulated PDB structures of acidic P_1 mutants are not stabilized by the bridge-forming water molecules.

2.4.2 Binding energy calculations of barnase-barstar complexes

The residual model is also applied to a set of barnase-barstar complexes. As comparisons, the experimental data set from Schreiber et al. for barnase-barstar is used to make a correlation between our calculations and the experimental values (Schreiber et al., 1997). In Figure 2.5, linear fit yields the correlation coefficient, 0.890 for barnase-barstar complex set. The calculated binding free energies, ΔG, for this set and the changes in binding free energies, $\Delta\Delta G$, from the single mutations on the active site are listed in Table 2.2.

Excluding a mutant which may involve additional hydrogen bonds, the linear fit correlation improves from 0.890 to 0.932. The mutants from the wild-type Glu73 in barnase-barstar complexes show the loss of hydrogen bonds and insertion of additional water molecules reducing the loss in binding energy. Especially for the Ser73E mutant in our calculation (the red triangle in Figure 2.5) the loss of the hydrogen bond and insertion of a water molecule causing destabilization may be more severe than other.

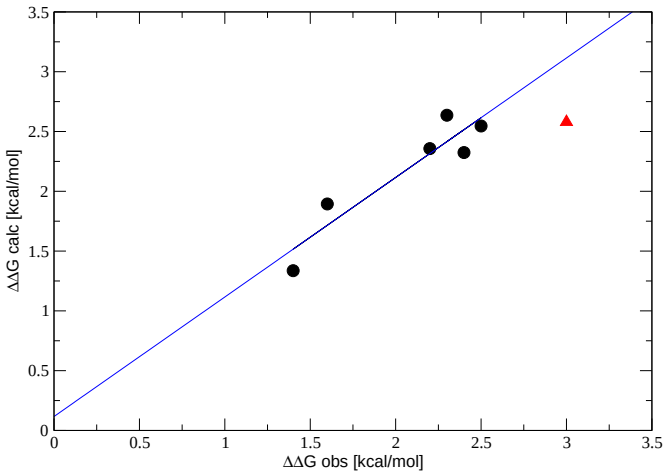


Figure 2.5 Calculated versus observed changes in the binding free energy brought by P_1 mutants of barnase-barstar complexes. The linear fit from all mutant sets yields 0.890. A residue (E73Ser) reported to form hydrogen bonds with water molecules in the complex is indicated by red up-triangles in barnase-barstar complexes. The linear fit excluding this mutation making water mediated hydrogen bonds yields 0.932 of barnase-barstar complexes.

Lo Conte et al. and Bahadur et al. analyzed the interface of protein-protein complexes with the interfacial atomic structures and classified the ratios of water molecule participation (Lo Conte et al., 1999; Bahadur and Zacharias, 2008). According to this analysis, all the interfacial residues in the barnase-barstrar complex are buried with water molecules. Again, given the simplicity of our model without explicit water modeling the correlation between the observed and the calculated binding energies is quite good for this system.

2.4.3 Binding energy calculations of OMTKY3-SGPB complexes

The experimental pK_a shift of OMTKY3 P_1 Glu bound to SGPB is 8.7 (Qasim et al., 1995) and also calculated to 13.1 (Brandsdal et al., 2006). We also used the PROPKA 2.0 (Delphine et al., 2008) and it gives pK_a values 8.7 and 8.8 for OMTKY3 P_1 Asp and P_1 Glu. With these shifted pK_as, the binding free energies, ΔG, of protein complex set are calculated and changes in binding free energies, $\Delta\Delta G$, from the single mutations on the active site are listed in Table 2.2. The correlation of $\Delta\Delta G$ data of our calculations with the observed data (Lu et al., 1997) yields the linear fit correlation coefficient, 0.828.

After taking into account the pK_a shifts of the acidic P_1 mutants, there are four additional exceptional data points in the correlation fitting in Figure 2.6(b). If all exceptional data points are excluded, the correlation between our calculations and the experimental results of $\Delta\Delta G$ yields an improved linear fit from 0.828 to 0.945.

The SGPB protein prefers hydrophobic P_1 side chain which are not branched at β-carbon (Lu et al., 1997). The wild-type Leu18I fits into the S_1 pocket of SGPB binding site in Figure 2.8(a) (Bateman et al., 2000). This pocket has narrow top entrance and broadening cavity toward the bottom. This narrow top structure causes that the β-branched residues cannot fit into the pocket. Thus, the β-branched side chains are not complementary to the shape of the S_1 binding site. The observed χ_1 angles of these residues in S_1 pocket are approximately 40° (Ile18I, 33°; Val18I, 47°; Thr18I, 39°;

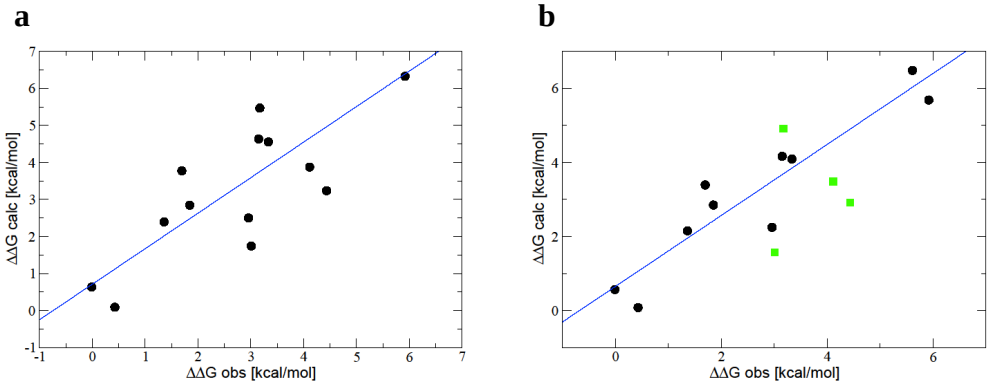
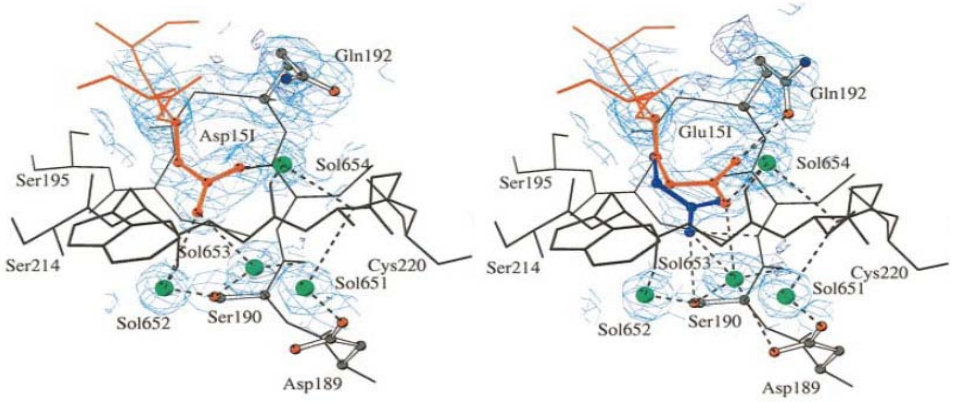


Figure 2.6 Calculated versus observed changes in the binding free energy brought by P_1 mutants of OMTKY3-SGPB complexes (a) and (b). The linear fit from all mutant sets yields 0.828. Deleterious effects of β-branched residues are indicated by green rectangles in Figure (b) in OMTKY3-SGPB complexes. Excluding this data makes the linear fit coefficient raise to 0.945.



(a) P_1 Asp **(b) P_1 Glu**

Figure 2.7 Structure of the P_1-S_1 binding site in BPTI-trypsin with P_1 Asp (a) and P_1 Glu (b). Only the hydrogen bonds between interfacial water molecules and P_1, S_1 residues are indicated by dashed lines. Figures were taken from Helland et al. (Helland et al., 1999)

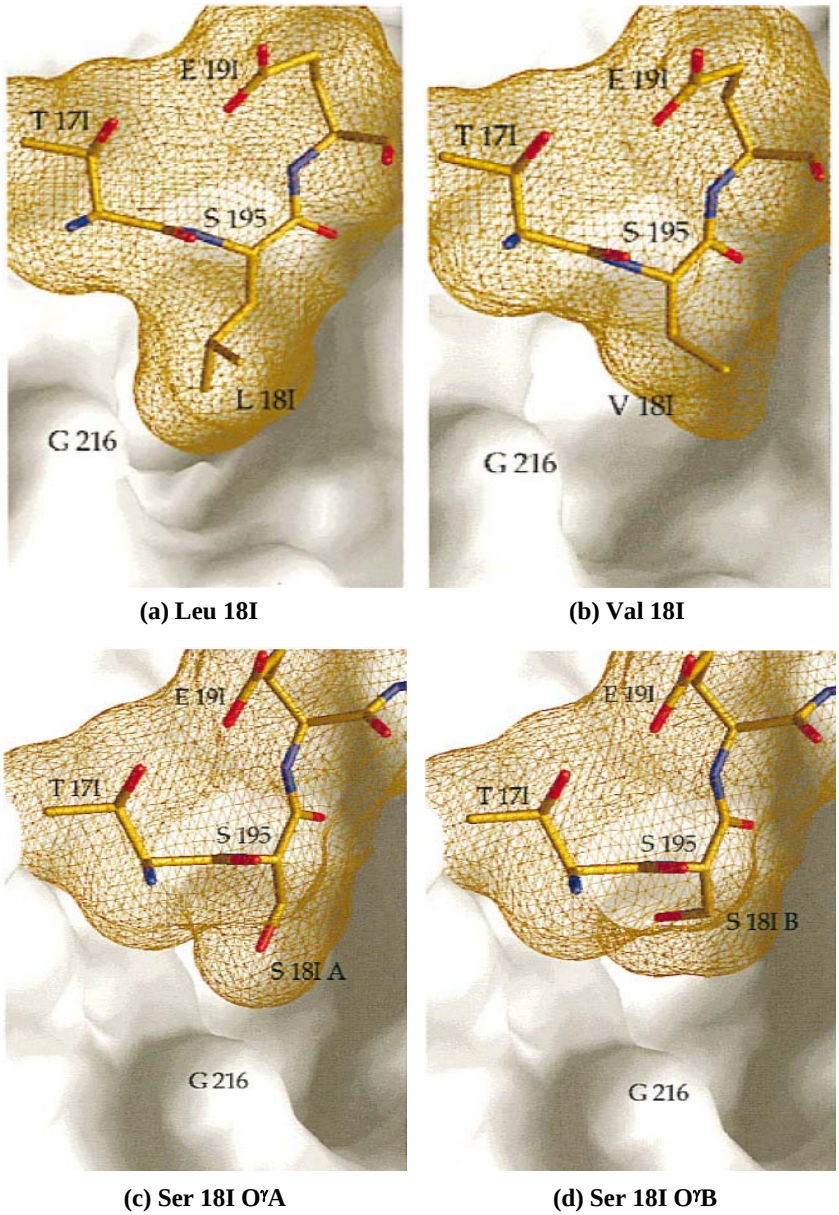
Table 2.2 Comparison of the binding free energy between the experimental data and calculated data ΔG. The first set is the result of P_1 mutants of BPTI-trypsin complexes, the second and third sets are the results of p_1 mutants of barnase-barstar complexes and SGPB-OMTKY3 complexes respectively. The PDB codes used here are based on the PDB code of the wild-type template for each complex set: the first 4 letter code is the experimental PDB code of the wild-type and 5th code is the one-letter code of the mutated amino acid. The wild-type itself is shown as bold character. $kcal/mol$ unit is used for all energy terms. The calculated binding free energies in each complex set are linearly scaled by setting the calculated binding free energy of the wild-type equal to the experimental binding free energy because of our model which includes only the electrostatic and van der Waals energy contributions.

PDB	ΔG_{obs}	ΔG_{calc}
3BTK	-17.86	-17.86
3BTKM	-10.25	-9.11
3BTKQ	-8.59	-7.76
3BTKT	-7.37	-7.39
3BTKW	-9.29	-10.40
1B27A	-16.70	-16.36
1B27C	-16.50	-16.45
1B27F	-16.80	-16.64
1B27Q	-17.60	-17.66
1B27S	-16.00	-16.42
1B27W	-17.40	-17.11
1B27Y	-16.60	-16.68
1B27	-19.00	-19.00
3SGBA	-11.55	-12.01
3SGBC	-14.52	-13.88
3SGBD	-8.90	-7.29
3SGBE	-8.59	-8.18
3SGBF	-13.15	-12.11
3SGBH	-12.81	-10.73
3SGBI	-10.07	-11.27
3SGBK	-11.36	-9.87
3SGB	-14.51	-14.51
3SGBM	-14.08	-14.42
3SGBR	-11.17	-9.95
3SGBS	-10.39	-10.63
3SGBT	-11.34	-9.04
3SGBV	-11.50	-12.76
3SGBW	-12.66	-11.66

Ser18I, $-46°$ or $40°$) that are rotated $\cong 180°$ away from their actual orientations (for Val mutation, see Figure 2.8(b)). The alternate conformations for Ser18I O^{γ} are also observed as shown on Figures 2.8(c) and 2.8(d). When the β-branched residues involve binding the bottom of the pocket is left relatively empty to avoid the steric crashes in contrast to the wild-type Leu18I whose side chain tightly fits into the bottom. Finally the empty cavity which is rare in protein-protein recognition site (Janin and Chothia, 1990) causes the complementary action involves close packing of the atoms between two protein molecules. The destabilization of a protein complex with respect to the cavity made by the mutation with β-branched residue is directly proportional to the cavity size. This uneven empty cavity followed by the closed packing from the β-branched residue mutation finally alters the geometric structure of the interface and this effect are described in our residual model. That is why the binding free energies of Ile18I, Val18I, Thr18I and Ser18I are more widely spread in Figure 2.6(b). On the other hand, our model can still account for the major changes of the binding energies due to single mutations for this system besides some effects due to atomic details.

2.5 Concluding Remarks

Three sets of protein-protein binding complexes, BPTI-trypsin, barnase-barstar and OMTKY3-SGPB are studied using our residue level protein-protein interaction model (Song, 2003; Song and Zhao, 2004). These complex sets involve changes in binding affinities of mutations on positively charged, negatively charged and neutral residues on the interfacial surfaces. Using the Poisson-Boltzmann linear integral equation solver implemented with the fast multipole method to calculate the electrostatic and the van der Waals interaction free energy, reasonable agreements with the binding affinities of these complexes from experiments demonstrate the utility of such a coarse-grained model to capture the most important contributions of protein binding.



(a) Leu 18I **(b) Val 18I**

(c) Ser 18I $O^{\gamma}A$ **(d) Ser 18I $O^{\gamma}B$**

Figure 2.8 Structural analysis on the interface of SGPB-OMTKY3 complexes. The molecular surface of SGPB is filled with white. The molecular surface of OMTKY3 inhibiotr residues are depicted as a gold mesh around the atoms(Oxygen, red; Nitrogen, blue; Carbon, gold) (a):SGPB:OMTKY3-Leu18I(wild-type); (b): SGPB:OMTKY3-Val18I; (c): SGPB:OMTKY3-Ser18I $O^{\gamma}A$; (d): SGPB:OMTKY3-Ser18I $O^{\gamma}B$. All figures were taken from Bateman et al. (Bateman et al., 2000)

At the same time additional effects due to atomic details during binding have to be considered to yield accurate binding affinities. For example, for P_1 Asp and P_1 Glu mutants in BPTI-trypsin complexes, the calculated pK_a based on the PROPKA 2.0 (Delphine et al., 2008) to describe the neutral behaviors of acidic residues greatly improve the correlations with experimental data. Considering the limited accuracy of calculations of residual pK_as, there will be a possible improvement of binding free energy calculations by using the experimental pK_as especially for the residues may have large charge changes upon binding.

CHAPTER 3. Calculations of the Second Virial Coefficients of Proteins with the Extended Fast Multipole Method

3.1 Abstract

The osmotic second virial coefficients B_2 are directly related to the solubility of protein molecules in electrolyte solutions and determined by molecular interactions involving both solvent and solute molecules. The calculations of interaction energies account for the electrostatic and the van der Waals interactions with the structural anisotropic properties of protein molecules. The orientational dependence of interaction energies between two proteins is determined by the crystal space group operations and relatively small number of protein-protein pair configurations according to the anisotropic patch model are required to calculate B_2 in this model. With the extended fast multipole methods both with double-tree and single-tree algorithm, the boundary element formulations of interaction energies can be applied with relatively low computational cost to the large protein molecules. B_2 Calculations of the Bovine Pancreatic Trypsin Inhibitor are first performed to validate our model and the results of lysozyme protein under different salts, salt concentrations, pH and temperatures are correlated to the experimental B_2.

3.2 Introduction

In a remarkable observation, George and Wilson found that there is a correlation between slightly negative second virial coefficient of a protein solution and its successful

crystallization condition (George et al., 1997). There is also a correlation between the solubility of a protein in an electrolyte solution and the osmotic second virial coefficient B_2 of the solution (Veesler et al., 1996; Boistelle et al., 1997). These observations have led to numerous studies on the second virial coefficients of protein solutions with the hope to use this property to narrow down the parameter space of protein solutions for the search of optimal crystallization conditions. For example, even for membrane proteins rapid screening of small molecules and detergents as crystallization additives are achieved to improve the crystallization conditions of light harvesting protein complexes (Gabrielsen et al., 2010).

Experimentally, the osmotic second virial coefficients B_2 can be measured by using the Static Light Scattering(SLC) (George et al., 1997; Farnum and Zukoski, 1999; Guo et al., 1999), the Small Angle X-ray Scattering(SAXS) (Bonneté et al., 2004), Small Angle Neutron Scattering(SANS) (Velev et al., 1998) or Self-Interaction Chromatography(SIC) (Tessier et al., 2002). All of these methods, however, are quite demanding due to large amounts of proteins used in the measurements. So far, using the B_2 of protein solutions as a tool to screen the solution conditions is not a routine practice yet in most crystallographers' labs.

To overcome the protein consumption problem in B_2 measurements, one possible alternative is to use computational methods to calculate the second virial coefficients of protein solutions. B_2 is related to molecular interactions in terms of the orientationally averaged potential of mean force(PMF), $W(r_{12})$, where r_{12} is the center-to-center distance,

$$B_2 = -2\pi \int_0^\infty \left(e^{-W(r_{12})/k_BT} - 1\right) r_{12}^2 \mathrm{d}r_{12}, \tag{3.1}$$

where W is the interaction free energy between two proteins, k_B is the Boltzmann constant and T the temperature. Previous efforts to model the interaction free energy between two protein molecules and to compute B_2 have been based on idealized descrip-

tions of proteins. The protein molecules are mostly treated as spheres, although Vilker et al. modeled a protein(bovine serum albumin) as an ellipsoid (Vilker et al., 1981). For spherical model approaches, the interaction is normally divided into two parts: the first part is due to the excluded volume to account for the size of protein molecules and the second part accounts for the solution dependent effective interaction between protein molecules. Due to the spherical shape approximation of protein molecules, the thickness of the hydration layer is often considered as an adjustable parameter for B_2 calculations. The solution dependent contributions to B_2 are modeled using standard colloidal methods (Hunter, 1987). Namely the van der Waals interaction is treated in the Lifshitz-Hamaker framework and the electrostatic interaction (Gallagher and Woodward, 1989; Muschol and Rosenberger, 1997; Kuehner et al., 1997; Vilker et al., 1981) is obtained using the Poisson-Boltzmann approach. For such idealized spherical models, with adjustable parameters such as the Hamaker constant the computed B_2 have been partially successful to capture the trend of experimental data at various solution conditions.

Neal et al. calculated the second virial coefficients by applying orientational dependence protein-protein interaction models (Neal et al., 1998). Electrostatic interactions in their study were obtained by distributing charges to the ionizable residues, thus an orientationally dependent charge distribution but treating the protein as a spherical dielectric cavity. The van der Waals interactions were calculated by a semi-empirical approach. When the intermolecular distance is large enough, the Lifshitz-Hamaker approach (Roth et al., 1996) was implemented with the realistic shape of proteins in mind. At shorter distance, the Optimized Potentials for Liquid Simulations(OPLS) parameter set (Jorgensen and Tirado-Rives, 1988) was used to capture the short-range interaction. Even though the comparison between their calculations and experimental measurements yields large errors for some B_2 calculations, this approach did not use any further adjustable parameters.

The goal of our work is to develop a protein-protein interaction model to account for the realistic shape of proteins and at the same time to capture the effect of solutions without adjustable parameter. To this end, a residue level model of protein-protein interaction had been introduced (Song, 2003; Song and Zhao, 2004).

In this model, each residue of a protein is represented by a sphere located at the geometric center of the residue determined by its native or approximate structure. The diameter of the sphere is determined by the molecular volume of a residue in a solution environment (Zamyatnin, 1984). The molecular surface of our model protein is defined as the Richard-Connolly surface spanned by the union of these residue spheres using MSMS program (Sanner, 1996). Each residue carries a permanent dipole moment located at the center of its sphere and the direction of the dipole is given by the amino acid type from protein's native structure. If a residue is charged the amount of charge is given by the Henderson-Hasselbalch equation using the generic pK_a values of residues, thus the local environmental effects on K_a values are neglected. For each residue there is also a polarizable dipole at the center of the sphere, whose nuclear polarizability had been determined from our recent work (Song, 2002a) and the electronic polarizability is estimated from optical dielectric constant augmented with quantum chemistry calculations (Millefiori et al., 2008). There are three kinds of interactions in this model: the electrostatic interaction due to the electric double layer effect, the van der Waals attraction due to the polarizable dipoles and a short range correction term to account for the short range interactions such as the desolvation energy, hydrophobic interaction and so on. In this report, we only consider the electrostatic interaction which gives the most contribution to the protein-protein interaction (Dong and Zhou, 2006; Brock et al., 2007), the van der Waals interaction and the short range interaction which is accounted for using the excluded volume based upon realistic shape of the protein.

The electrostatic problem in the electrostatic and the van der Waals interaction is solved using the Poisson-Boltzmann equation where the realistic shape of protein

molecules are considered. The boundary element method (BEM) in combination with the fast multipole method (Greengard, 1988; Greengard and Rokhlin, 1997) is implemented to circumvent the extensive memory problem similar to the recent work (Lu et al., 2007). The validity of our model was already tested by binding affinity calculations of several protein complexes (Kim et al., 2010). Direct comparisons between our calculations of B_2 and experimental measurements under various solution conditions were made and reasonable agreements from these comparisons provide the another concrete evidence that our model can be used as a universal model for studies of non-specific protein-protein interactions in aqueous solutions.

3.3 Theoretical development

3.3.1 General formulation for the second virial coefficient calculation using a residue level patch model

The osmotic second virial coefficient(B_2) can be expressed in terms of the interaction energy between two proteins McQuarrie (1976).

$$B_2 = -\frac{1}{8\pi}\int_{\Omega_2}\int_{\Omega_1}\int_0^\infty \left(e^{-W(R,\Omega_1,\Omega_2)/kT} - 1\right) R^2 \mathrm{d}R\mathrm{d}\Omega_1\mathrm{d}\Omega_2 \tag{3.2}$$

where the interaction energy W describes the anisotropic interaction between two proteins depending on the center-to-center distance, R, and the relative orientations of two molecules. Ω_1 and Ω_2 describe the angular position and the direction of both protein molecules. We can only calculate the interaction energy where two proteins are not too close each other so that Eq. (3.2) can be split into two parts, the hard sphere contribution and the intermolecular interaction.

We can only calculate the interaction energy where the two proteins are not too close each other so that the Eq. (3.2) can be split into the two parts, the hard sphere

contribution and the intermolecular interaction.

$$B_2 = \frac{1}{8\pi}\int_{\Omega_2}\int_{\Omega_1}\left\{\frac{1}{3}r_c^3 - \int_0^\infty \left(e^{-W(R,\Omega_1,\Omega_2)/kT} - 1\right) R^2 \mathrm{d}R\right\}\mathrm{d}\Omega_1\mathrm{d}\Omega_2, \quad (3.3)$$

where r_c, a function of Ω_1 and Ω_2, is the hard sphere diameter described by the separation distance between two protein molecules. Arising problem for solving the integral in Eq. (3.3) is to integrate out the solid angle dependence, Ω_1 and Ω_2. It is a hard work to consider all the possible combinations of the position and orientation of two proteins. We use the anisotropic patch model (Vega et al., 2008) to set up the interaction pairs defined by two patches which are the closest surface elements to the inter-particle(center-to-center) vector between two proteins. The patch vector which is from the center-of-mass of a protein to the surface patch defines the orientation of a protein and a pair of two patch indices from two proteins represents the relative orientation of the pair interaction between two proteins. The number of interaction pairs in patch model depends on the number of surface elements N and M on each protein A and B respectively and it is $N \times M$. This number of interaction pairs is still too large to compute all the pair interaction energies. So an assumption is made that the only dominant pairs of orientations based upon the crystal lattice structure whose information is described in the Protein Data Bank(PDB) file are considered to calculate the interaction energy as a function of the center-to-center separation $R(i)$, where i is the index of pair orientations. With this sampling we can only compute the interaction energies within the small number of pair interactions. But the hard sphere contribution should be considered more precisely. Just a simple treatment using a single sphere for the protein separation can cause a significant problem because the interaction within the short range dominates the value of the second virial coefficient B_2 (Rosenbaum and Zukoski, 1996; Neal and Lenhoff, 1995). Calculation of the hard sphere separation can be done if the separation distance is calculated on each surface element of the first protein with its possible interaction

pairs from the second protein as follows,

$$\begin{aligned} B_2 &= 2\pi \sum_{l=1}^{N_{\rm A}} \sum_{m=1}^{N_{\rm B}} \frac{1}{3} r_{c,l,m}^3 \frac{\sigma_l \sigma_m}{\sigma_{\rm A} \sigma_{\rm B}} \\ &- \frac{2\pi}{p} \sum_{i=1}^{p} \int_0^\infty \left(e^{-W(R(i),i)/kT} - 1 \right) R(i)^2 {\rm d}R(i), \end{aligned} \tag{3.4}$$

where $r_{c,l,m}$ is the hard sphere separation of two protein molecules when the surface element l on protein A and the surface m on protein B are upon contact. $N_{\rm A}$ and $N_{\rm B}$ are the number of surface elements, $\sigma_{\rm A}$ and $\sigma_{\rm B}$ are total surface areas of protein A and B and σ_l and σ_m are the surface areas of the surface element l on protein A and the surface element m on protein B respectively. The surface area ratio, $\sigma_{\rm l}(\sigma_{\rm m})$ to $\sigma_{\rm A}(\sigma_{\rm B})$, represents the solid angle dependence, ${\rm d}\Omega_1({\rm d}\Omega_2)$. The pair configurations based on the crystal space group is indicated as i and its total number p. So the center-to-center distance of each pair configuration and the interaction energy have dependence on the pair configuration index i, thus W_i and $R(i)$. In this report, the interaction energy between two protein molecules can be calculated by the sum of the electrostatic interaction energy and the van der Waals interaction energy for each distance separation and pair configuration.

$$W(R(i), i) = \Delta E_{\rm elec}(R(i), i) + \Delta E_{\rm vdw}(R(i), i) \tag{3.5}$$

In the next section, we describe how the pair configurations between two protein structures can be set up and interaction energy of each configuration can be calculated.

3.3.2 General formulation of the electrostatic interaction free energy between two proteins with the Boundary Element Method

We derived the integral equations of the linearized Poisson-Boltzmann equation for two protein model (Song, 2003) following the previous work (Juffer et al., 1991) whose study was based on the single domain problem. Consider the molecular surfaces $\sum_1$ and $\sum_2$ which cover two protein molecules respectively. There are N charges q_i and dipoles

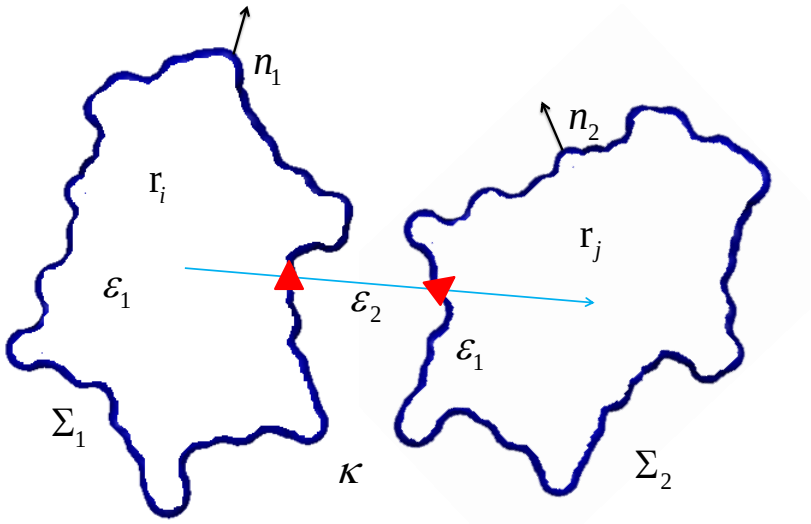


Figure 3.1 Schematic illustration showing the formulation of the electrostatic interaction of two proteins. The orientations of two proteins are defined by two surface patches(red triangles) nearest to the inter-particle vector(blue arrow) whose magnitude is the center-to-center distance R. The molecular surfaces are defined by $\sum_1$ and $\sum_2$ for each protein and the n_1 and n_2 are the outward unit normals on $\sum_1$ and $\sum_2$. ε_1 and ε_2 are the dielectric constants of the inside protein cavity and solution respectively. κ represents the inverse Debye screening length. Charge q_i and dipole μ_i are located on each residue center.

$\vec{\mu_i}$ at position $\mathbf{r}_i$ enclosed by the surface $\sum_1$ and also there are N charges q_j and dipoles $\vec{\mu_j}$ at position $\mathbf{r}_j$ enclosed by the surface $\sum_2$. Inside each dielectric cavity the dielectric constant is ε_1 and the dielectric constant of the solution is given as ε_2(see Figure 3.1). The inverse Debye screening length κ is given by the solution's ionic strength and the temperature, $\kappa = \sqrt{\frac{2IF^2}{4\pi\varepsilon_0\varepsilon RT}} = \sqrt{\frac{I}{T}} \times (1.586115104)\, \dot{A}^{-1}$, where ε_0 is the permittivity of free space, ε is the dielectric constant of water, R is the gas constant, F is the Faraday constant and I is the ionic strength of the electrolyte. The integral equations for the potential $\varphi_1(\mathbf{r})$ and $\varphi_2(\mathbf{r})$ and their gradient $\partial\varphi_1(\mathbf{r})/\partial(n_1)$ and $\partial\varphi_2(\mathbf{r})/\partial(n_2)$ are given by the following integral equations (Song, 2003),

$$\frac{1}{2}\left(1+\frac{\varepsilon_2}{\varepsilon_1}\right)\varphi_1(\mathbf{r_{01}})+\iint_{\sum_1} L_1(\mathbf{r_1},\mathbf{r_{01}})\varphi_1(\mathbf{r_1})\mathrm{d}\mathbf{r_1}$$
$$+\iint_{\sum_1} L_2(\mathbf{r_1},\mathbf{r_{01}})\frac{\partial\varphi_1(\mathbf{r_1})}{\partial n_1}\mathrm{d}\mathbf{r_1}$$
$$-\iint_{\sum_2} L_1(\mathbf{r_2},\mathbf{r_{01}})\varphi_2(\mathbf{r_2})\mathrm{d}\mathbf{r_2}$$
$$+\iint_{\sum_2} L_2(\mathbf{r_2},\mathbf{r_{01}})\frac{\partial\varphi_2(\mathbf{r_2})}{\partial n_2}\mathrm{d}\mathbf{r_2}$$
$$=\sum_{i=1}^{2N}\left\{q_iF(\mathbf{r_i},\mathbf{r_{01}})+\vec{\mu_i}\cdot\nabla F(\mathbf{r_i},\mathbf{r_{01}})\right\}/\varepsilon_1, \tag{3.6}$$

$$\frac{1}{2}\left(1+\frac{\varepsilon_1}{\varepsilon_2}\right)\frac{\partial\varphi_1(\mathbf{r_{01}})}{\partial n_1}+\iint_{\sum_1} L_3(\mathbf{r_1},\mathbf{r_{01}})\varphi_1(\mathbf{r_1})\mathrm{d}\mathbf{r_1}$$
$$+\iint_{\sum_1} L_4(\mathbf{r_1},\mathbf{r_{01}})\frac{\partial\varphi_1(\mathbf{r_1})}{\partial n_1}\mathrm{d}\mathbf{r_1}$$
$$-\iint_{\sum_2} L_3(\mathbf{r_2},\mathbf{r_{01}})\varphi_2(\mathbf{r_2})\mathrm{d}\mathbf{r_2}$$
$$+\iint_{\sum_2} L_4(\mathbf{r_2},\mathbf{r_{01}})\frac{\partial\varphi_2(\mathbf{r_2})}{\partial n_2}\mathrm{d}\mathbf{r_2}$$
$$=\sum_{i=1}^{2N}\left\{q_i\frac{\partial F}{\partial n_{01}}(\mathbf{r_i},\mathbf{r_{01}})+\vec{\mu_i}\cdot\nabla\frac{\partial F}{\partial n_{01}}(\mathbf{r_i},\mathbf{r_{01}})\right\}/\varepsilon_1, \tag{3.7}$$

$$\frac{1}{2}\left(1+\frac{\varepsilon_2}{\varepsilon_1}\right)\varphi_2(\mathbf{r_{02}})-\iint_{\sum_1} L_1(\mathbf{r_1},\mathbf{r_{02}})\varphi_1(\mathbf{r_1})\mathrm{d}\mathbf{r_1}$$
$$+\iint_{\sum_1} L_2(\mathbf{r_1},\mathbf{r_{02}})\frac{\partial\varphi_1(\mathbf{r_1})}{\partial n_1}\mathrm{d}\mathbf{r_1}$$
$$+\iint_{\sum_2} L_1(\mathbf{r_2},\mathbf{r_{02}})\varphi_2(\mathbf{r_2})\mathrm{d}\mathbf{r_2}$$
$$+\iint_{\sum_2} L_2(\mathbf{r_2},\mathbf{r_{02}})\frac{\partial\varphi_2(\mathbf{r_2})}{\partial n_2}\mathrm{d}\mathbf{r_2}$$
$$=\sum_{i=1}^{2N}\left\{q_iF(\mathbf{r_i},\mathbf{r_{02}})+\vec{\mu_i}\cdot\nabla F(\mathbf{r_i},\mathbf{r_{02}})\right\}/\varepsilon_1, \tag{3.8}$$

$$\frac{1}{2}\left(1+\frac{\varepsilon_1}{\varepsilon_2}\right)\frac{\partial\varphi_2(\mathbf{r_{02}})}{\partial n_2}-\iint_{\sum_1}L_3(\mathbf{r_1},\mathbf{r_{02}})\varphi_1(\mathbf{r_1})\mathrm{d}\mathbf{r_1}$$
$$+\iint_{\sum_1}L_4(\mathbf{r_1},\mathbf{r_{02}})\frac{\partial\varphi_1(\mathbf{r_1})}{\partial n_1}\mathrm{d}\mathbf{r_1}$$
$$+\iint_{\sum_2}L_3(\mathbf{r_2},\mathbf{r_{02}})\varphi_2(\mathbf{r_2})\mathrm{d}\mathbf{r_2}$$
$$+\iint_{\sum_2}L_4(\mathbf{r_2},\mathbf{r_{02}})\frac{\partial\varphi_2(\mathbf{r_2})}{\partial n_2}\mathrm{d}\mathbf{r_2}$$
$$=\sum_{i=1}^{2N}\left\{q_i\frac{\partial F}{\partial n_{02}}(\mathbf{r_i},\mathbf{r_{02}})+\vec{\mu_i}\cdot\nabla\frac{\partial F}{\partial n_{02}}(\mathbf{r_i},\mathbf{r_{02}})\right\}/\varepsilon_1, \tag{3.9}$$

where

$$L_1(\mathbf{r},\mathbf{r}_0) = \tfrac{\partial F}{\partial n}(\mathbf{r},\mathbf{r}_0)-\tfrac{\varepsilon_2}{\varepsilon_1}\tfrac{\partial P}{\partial n}(\mathbf{r},\mathbf{r}_0), \tag{3.10}$$

$$L_2(\mathbf{r},\mathbf{r}_0) = P(\mathbf{r},\mathbf{r}_0)-F(\mathbf{r},\mathbf{r}_0), \tag{3.11}$$

$$L_3(\mathbf{r},\mathbf{r}_0) = \tfrac{\partial^2 F}{\partial n_0\partial n}(\mathbf{r},\mathbf{r}_0)-\tfrac{\partial^2 P}{\partial n_0\partial n}(\mathbf{r},\mathbf{r}_0), \tag{3.12}$$

$$L_4(\mathbf{r},\mathbf{r}_0) = -\tfrac{\partial F}{\partial n_0}(\mathbf{r},\mathbf{r}_0)+\tfrac{\partial P}{\partial n_0}(\mathbf{r},\mathbf{r}_0)\tfrac{\varepsilon_1}{\varepsilon_2} \tag{3.13}$$

and

$$F(\mathbf{r},\mathbf{r}_0) = \tfrac{1}{4\pi|\mathbf{r}-\mathbf{r}_0|},$$
$$P(\mathbf{r},\mathbf{r}_0) = \tfrac{e^{-\kappa|r-\mathbf{r}_0|}}{4\pi|\mathbf{r}-\mathbf{r}_0|}. \tag{3.14}$$

Although the traditional Boundary Element Method such as Atkinson and his coworkers' (Atkinson and Han, 2009) can be used to solve above integral equations, the memory requirement is too costly even on the newest computers using either a direct linear system solver or iterative solver, such as Generalized minimal residual method(GMRES)for a moderate size protein. In the current work the fast multipole method is implemented and the details will be outlined in chapter 5. Once the above integral equations are

solved the potentials inside the dielectric cavity are,

$$\begin{aligned}
\varphi_1(\mathbf{r_{01}}) &= -\iint_{\sum_1} L_1(\mathbf{r_1},\mathbf{r_{01}})\varphi_1(\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&\quad - \iint_{\sum_1} L_2(\mathbf{r_1},\mathbf{r_{01}})\frac{\partial\varphi_1(\mathbf{r_1})}{\partial n_1}\mathrm{d}\mathbf{r_1}, \qquad (3.15)\\
\varphi_2(\mathbf{r_{02}}) &= -\iint_{\sum_2} L_1(\mathbf{r_2},\mathbf{r_{02}})\varphi_2(\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&\quad - \iint_{\sum_2} L_2(\mathbf{r_2},\mathbf{r_{02}})\frac{\partial\varphi_2(\mathbf{r_2})}{\partial n_2}\mathrm{d}\mathbf{r_2}, \qquad (3.16)\\
\nabla_{01}\varphi_1(\mathbf{r_{01}}) &= -\iint_{\sum_1} \nabla_{01}L_1(\mathbf{r_1},\mathbf{r_{01}})\varphi_1(\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&\quad - \iint_{\sum_1} \nabla_{01}L_2(\mathbf{r_1},\mathbf{r_{01}})\frac{\partial\varphi_1(\mathbf{r_1})}{\partial n_1}\mathrm{d}\mathbf{r_1}, \qquad (3.17)\\
\nabla_{02}\varphi_2(\mathbf{r_{02}}) &= -\iint_{\sum_2} \nabla_{02}L_1(\mathbf{r_2},\mathbf{r_{02}})\varphi_2(\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&\quad - \iint_{\sum_2} \nabla_{02}L_2(\mathbf{r_2},\mathbf{r_{02}})\frac{\partial\varphi_2(\mathbf{r_2})}{\partial n_2}\mathrm{d}\mathbf{r_2}. \qquad (3.18)
\end{aligned}$$

The electrostatic free energy between the protein molecules at a center-to-center distance, R, and relative orientations, Ω_1 and Ω_2, is given by,

$$\begin{aligned}
E_{ele}(R,\Omega_1,\Omega_2) &= \sum_{i=1}^{N}\left\{\frac{q_i}{\epsilon_1}\varphi_1(\mathbf{r_i}) + \frac{1}{\epsilon_1}\vec{\mu_i}\cdot\nabla\varphi_1(\mathbf{r_i})\right\} \\
&+ \sum_{j=1}^{N}\left\{\frac{q_j}{\epsilon_1}\varphi_2(\mathbf{r_j}) + \frac{1}{\epsilon_1}\vec{\mu_j}\cdot\nabla\varphi_2(\mathbf{r_j})\right\}. \qquad (3.19)
\end{aligned}$$

Finally, the effective electrostatic interaction between two proteins is

$$\begin{aligned}
\Delta E_{ele}(R,\Omega_1,\Omega_2) &= E_{ele}(R,\Omega_1,\Omega_2) - E_{ele}(R\to\infty,\Omega_1,\Omega_2) \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\varepsilon_1}\left\{q_iT_{ij}q_j - q_i\sum_{\alpha}T_{ij}^{\alpha}\mu_{j,\alpha}\right. \\
&\left. + \sum_{\alpha}\mu_{i,\alpha}T_{ij}^{\alpha}q_j - \sum_{\alpha}\mu_{i,\alpha}T_{ij}^{\alpha\beta}\mu_{j,\beta}\right\}, \qquad (3.20)
\end{aligned}$$

where the interaction tensors, charge-charge, charge-dipole, dipole-charge and dipole-

dipole, are given by,

$$\begin{aligned} T_{ij} &= \frac{e^{-\kappa R}}{R}, \\ T_{ij}^{\alpha} &= \nabla_\alpha T_{ij} = e^{-\kappa R}\frac{(1+\kappa R)}{R^3} r_{ij,\alpha}, \\ T_{ij}^{\alpha\beta} &= \nabla_\alpha \nabla_\beta T_{ij} = e^{-\kappa R}\left\{\left(\frac{3}{R^5}+\frac{3\kappa}{R^4}+\frac{\kappa^2}{R^3}\right) r_{ij,\alpha} r_{ij,\beta} \right. \\ &\quad \left. -\left(\frac{1}{R^3}+\frac{\kappa}{R^2}\right)\delta_{\alpha\beta}\right\}. \end{aligned} \tag{3.21}$$

Here ∇_α is $\frac{\partial}{\partial r_{ij,\alpha}}$ for each $\alpha = x, y, z$. The last summation terms in Eq. (3.20) are the interaction energies between charges and dipoles in two proteins when the solution is in the same dielectric constant ε_1 such as the inside protein and with the inverse Debye screening length κ.

3.3.3 General formulation of the van der Waals interaction free energy

The van der Waals interaction free energy between two proteins is defined as,

$$\begin{aligned} \Delta E_{\text{vdw}}(R, \Omega_1, \Omega_2) &= E_{\text{vdw}}(R, \Omega_1, \Omega_2) \\ &- E_{\text{vdw}}(R \to \infty, \Omega_1, \Omega_2). \end{aligned} \tag{3.22}$$

Song and Zhao formulated the van der Waals interaction between the protein molecules in an electrolyte solution as the following effective action in Fourier space (Song and Zhao, 2004),

$$\begin{aligned} S\left[\mathbf{m}_{\mathbf{r},n}\right] &= -\frac{\beta}{2}\sum_{\mathbf{r}}\sum_{n=-\infty}^{n=\infty}\frac{1}{\alpha_{\mathbf{r},n}}\mathbf{m}_{\mathbf{r},n}\cdot\mathbf{m}_{\mathbf{r},-n} \\ &+\frac{\beta}{2}\sum_{\mathbf{r}\neq\mathbf{r}'}\sum_{n=-\infty}^{n=\infty}\frac{1}{\alpha_{\mathbf{r},n}}\mathbf{m}_{\mathbf{r},n}\cdot T(\mathbf{r}-\mathbf{r}')\cdot\mathbf{m}_{\mathbf{r},-n} \\ &+\frac{\beta}{2}\sum_{\mathbf{r},\mathbf{r}'}\sum_{n=-\infty}^{n=\infty}\frac{1}{\alpha_{\mathbf{r},n}}\mathbf{m}_{\mathbf{r},n}\cdot R_n(\mathbf{r}-\mathbf{r}')\cdot\mathbf{m}_{\mathbf{r},-n}, \end{aligned} \tag{3.23}$$

where $\alpha_{\mathbf{r},n}$ is the frequency-dependent polarizability of a residue located at positon $\mathbf{r}$. $T(\mathbf{r}-\mathbf{r}')$ is the dipole-dipole interaction tensor between $\mathbf{r}$ and $\mathbf{r}'$. $R_n(\mathbf{r}-\mathbf{r}')$ is the

reaction field tensor at frequency $\omega_n = 2\pi n/\beta\hbar$. If the electrolyte solvent is treated by the Debye-Hückel theory, this reaction field tensor can be calculated by solving the Poisson-Boltzmann equation with the dielectric constant $\varepsilon(i\omega_n)$. The quantum partition function from this effective action of the system is,

$$Q\left(R,\Omega_1,\Omega_2\right) = \prod_n \left[\frac{2\pi}{\beta \det A_n\left(R,\Omega_1,\Omega_2\right)}\right]^{1/2}, \tag{3.24}$$

where R represents the center-to-center distance between two proteins, Ω_1 and Ω_2 is the relative orientations of two proteins and A_n's matrix element is given by,

$$A_n(\mathbf{r},\mathbf{r}') = \frac{1}{\alpha_{\mathbf{r},n}}\delta_{\mathbf{r},\mathbf{r}'} - T(\mathbf{r}-\mathbf{r}') - R_n(\mathbf{r}-\mathbf{r}'), \tag{3.25}$$

where $\mathbf{r}$ and $\mathbf{r}'$ represent residues in each protein. The symbol "det" represents the determinant of the matrix. Finally, the van der Waals interaction free energy is given by,

$$\begin{aligned}\Delta E_{\text{vdw}} = \frac{1}{2}k_B T \sum_{n=-\infty}^{n=\infty} &\left[\ln\left\{\det A_n\left(R,\Omega_1,\Omega_2\right)\right\}\right. \\ &\left. - \ln\left\{\det A_n\left(R\to\infty\right)\right\}\right].\end{aligned} \tag{3.26}$$

In order to evaluate the van der Waals interaction in our model, the reaction field matrix $R_n(\mathbf{r}-\mathbf{r}')$ has to be calculated with the properties of the proteins and the solution. The boundary element formulation which is used to evaluate the electrostatic free energy also can be used to calculate the reaction field matrix. Consider two molecular surfaces $\sum_1$ and $\sum_2$ spanned by two protein molecules. There are N polarizable dipoles $\mathbf{m_r}$ at position $\mathbf{r}$ enclosed by each surface $\sum_1$ and $\sum_2$. Inside this dielectric cavity the dielectric constant is one and the dielectric constant of the solution is $\varepsilon(i\omega_n)$ at the Matsubara frequency ω_n. The inverse Debye screening length κ is given by the solution's ionic strength and the temperature. If we recognize that in order to calculate the potential at the molecular surface a dipole $\mathbf{m}$ at position $\mathbf{r}_0$ can be described by an effective charge density $\rho_{eff}(\mathbf{r}) = -\mathbf{m}\nabla\delta(\mathbf{r}-\mathbf{r}_0)$ (Jackson, 1999), the reaction field matrix involving residues $\mathbf{r}_i$ and $\mathbf{r}_j$ can be given as,

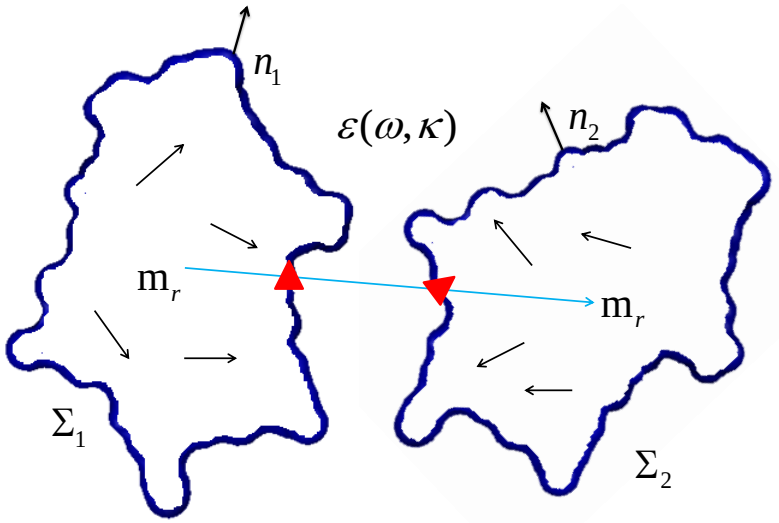


Figure 3.2 Schematic illustration showing the formulation of the van der Waals interaction of two proteins. The molecular surfaces are defined by $\sum_1$ and $\sum_2$ for each protein and the n_1 and n_2 are the outward unit normals on $\sum_1$ and $\sum_2$ and ε is the dielectric constant of the outside solution as a function of the frequency ω and the inverse Debye screening length κ. The orientations of two proteins are defined by two surface patches(red triangles) nearest to the inter-particle vector(blue arrow) whose magnitude is the center-to-center distance R. m_rs stand for the polarizable dipoles located on the residue center.

$$R(\mathbf{r}_i,\mathbf{r}_j) = \iint_{\sum_p} \left[\nabla_i F(\mathbf{r}_i,\mathbf{r}_j) \; -\nabla_i P(\mathbf{r}_i,\mathbf{r}_j)\right] \frac{\partial \varphi_p}{\partial n_p}(\mathbf{r}_j,\mathbf{r}_p) d\mathbf{r}_p$$
$$+ \iint_{\sum_p} \left[-\nabla_i \frac{\partial F}{\partial n_j} F(\mathbf{r}_i,\mathbf{r}_j) \; +\nabla_i \frac{\partial P}{\partial n_j}(\mathbf{r}_i,\mathbf{r}_j)\varepsilon\right] \varphi_p(\mathbf{r}_j,\mathbf{r}_p) d\mathbf{r}_p, \tag{3.27}$$

where F and P are followed by the previous notation in Eq. (3.14) and p which can be 1 or 2 depends upon $\mathbf{r}_j$ in $\sum_1$ or $\sum_2$. φ_p and $\partial\varphi_p$ can be obtained by solving the following linear equations of integral equations (Song, 2003; Juffer et al., 1991).

$$\begin{aligned}
&\frac{1}{2}\left(1+\varepsilon(i\omega_n)\right)\varphi_1(\mathbf{r}_i,\mathbf{r_{01}}) \\
&+ \iint_{\sum_1} L_1(\mathbf{r_1},\mathbf{r_{01}})\varphi_1(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&+ \iint_{\sum_1} L_2(\mathbf{r_1},\mathbf{r_{01}})\frac{\partial\varphi_1}{\partial n_1}(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&- \iint_{\sum_2} L_1(\mathbf{r_2},\mathbf{r_{01}})\varphi_2(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&+ \iint_{\sum_2} L_2(\mathbf{r_2},\mathbf{r_{01}})\frac{\partial\varphi_2}{\partial n_2}(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&= \nabla_i F(\mathbf{r}_i,\mathbf{r_{01}}),
\end{aligned} \tag{3.28}$$

$$\begin{aligned}
&\frac{1}{2}\left(1+\frac{1}{\varepsilon(i\omega_n)}\right)\frac{\partial\varphi_1}{\partial n_1}(\mathbf{r}_i,\mathbf{r_{01}}) \\
&+ \iint_{\sum_1} L_3(\mathbf{r_1},\mathbf{r_{01}})\varphi_1(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&+ \iint_{\sum_1} L_4(\mathbf{r1},\mathbf{r_{01}})\frac{\partial\varphi_1}{\partial n_1}(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&- \iint_{\sum_2} L_3(\mathbf{r_2},\mathbf{r_{01}})\varphi_2(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&+ \iint_{\sum_2} L_4(\mathbf{r2},\mathbf{r_{01}})\frac{\partial\varphi_2}{\partial n_2}(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&= \nabla_i \frac{\partial F}{\partial n_{01}}(\mathbf{r}_i,\mathbf{r_{01}}),
\end{aligned} \tag{3.29}$$

$$
\begin{aligned}
&\frac{1}{2}\left(1+\varepsilon(i\omega_n)\right)\varphi_2(\mathbf{r}_i,\mathbf{r_{02}}) \\
&-\iint_{\sum_1} L_1(\mathbf{r_1},\mathbf{r_{02}})\varphi_1(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&+\iint_{\sum_1} L_2(\mathbf{r_1},\mathbf{r_{02}})\frac{\partial\varphi_1}{\partial n_1}(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&+\iint_{\sum_2} L_1(\mathbf{r_2},\mathbf{r_{02}})\varphi_2(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&+\iint_{\sum_2} L_2(\mathbf{r_2},\mathbf{r_{02}})\frac{\partial\varphi_2}{\partial n_2}(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&=\nabla_i F(\mathbf{r}_i,\mathbf{r_{02}}),
\end{aligned}
\tag{3.30}
$$

$$
\begin{aligned}
&\frac{1}{2}\left(1+\frac{1}{\varepsilon(i\omega_n)}\right)\frac{\partial\varphi_2}{\partial n_2}(\mathbf{r}_i,\mathbf{r_{02}}) \\
&-\iint_{\sum_1} L_3(\mathbf{r_1},\mathbf{r_{02}})\varphi_1(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&+\iint_{\sum_1} L_4(\mathbf{r1},\mathbf{r_{02}})\frac{\partial\varphi_1}{\partial n_1}(\mathbf{r}_i,\mathbf{r_1})\mathrm{d}\mathbf{r_1} \\
&+\iint_{\sum_2} L_3(\mathbf{r_2},\mathbf{r_{02}})\varphi_2(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&+\iint_{\sum_2} L_4(\mathbf{r2},\mathbf{r_{02}})\frac{\partial\varphi_2}{\partial n_2}(\mathbf{r}_i,\mathbf{r_2})\mathrm{d}\mathbf{r_2} \\
&=\nabla_i\frac{\partial F}{\partial n_{02}}(\mathbf{r}_i,\mathbf{r_{02}}),
\end{aligned}
\tag{3.31}
$$

where L_1, L_2, L_3, and L_4 are defined in Eqs. (3.10), (3.11), (3.12) and (3.13). To evaluate the van der Waals interaction free energy in Eq. (3.26), the reaction field matrix should be built corresponding to the dielectric constant $\varepsilon(i\omega_n)$ for each frequency ω_n. And the total polarizability of a residue in a protein can be given by,

$$
\alpha_n=\alpha(i\omega_n)=\frac{\alpha_{nu}}{1+\omega_n/\omega_{rot}}+\frac{\alpha_{el}}{1+\left(\omega_n/\omega_I\right)^2},
\tag{3.32}
$$

where α_{nu} is the static nuclear polarizability of a residue (Song, 2002a) and ω_{rot} is a characteristic frequency of nuclear collective motion from a generalization of the De-

bye model. α_{el} is the static electronic polarizability of a residue and ω_I is the ionization frequency of a residue as in the Drude oscillator model of electronic polarizabilities. $\omega_{rot} = 20cm^{-1}$ for this calculation which is typical rotational frequency of molecules (Israelachvili, 1985). Other properties are listed in Table 2.1 based on the calculated result (Millefiori et al., 2008). An accurate parametrization of the dielectric function $var\epsilon(i\omega)$ of water based on the experimental data is taken from Parsegian's work (Parsegian, 1975).

3.3.4 Solving the linear system: the iterative double-tree Fast Multipole Method

The system of linear equations in Eqs. (3.6), (3.7), (3.8) and (3.9) for the electrostatic interaction energy and Eqs. (3.28), (3.29), (3.30) and (3.31) for the van der Waals interaction energy have the following form,

$$(\mathrm{I} - \mathrm{L})\mathrm{A} = \mathrm{B} \tag{3.33}$$

where A and B are single column vectors with the size of $2N$, the number of surface elements on the protein molecules for the electrostatic energy calculation and also can be $(2N) \times (2N)$ matrix for the reaction field calculation of the van der Waals energy calculation. Rewriting this equation with details makes the following form of the linear system.

$$I \begin{pmatrix} \varphi_{00} \\ \varphi_{11} \\ \varphi_{22} \\ \varphi_{33} \end{pmatrix} - \begin{pmatrix} L_1^{00} & L_2^{01} & L_1^{02} & L_2^{03} \\ L_3^{10} & L_4^{11} & L_3^{12} & L_4^{13} \\ L_1^{20} & L_2^{21} & L_1^{22} & L_2^{23} \\ L_3^{30} & L_4^{31} & L_3^{32} & L_4^{33} \end{pmatrix} \begin{pmatrix} \varphi_{00} \\ \varphi_{11} \\ \varphi_{22} \\ \varphi_{33} \end{pmatrix} = \begin{pmatrix} F_{00} \\ F_{11} \\ F_{22} \\ F_{33} \end{pmatrix} \tag{3.34}$$

where I is the identity matrix with the size of $(2N) \times (2N)$, the matrix element, L_1, L_2, L_3 and L_4 are defined in Eqs. (3.10), (3.11), (3.12) and (3.13), and the first upper indices are the equation indices from Eq. (3.6) to Eq. (3.9) or from Eq. (3.28) to Eq.

(3.31) for the electrostatic interaction and the van der Waals interaction respectively. The second upper indices are the indices of term L's indices in each integral equation. This system of linear equations is the order $O((2N)^2)$, so we need to consider how we can save the computational cost by reducing the size of matrix. If the interaction between two different bodies compared with the self-interaction of each body, the contribution from the matrix elements in indices 02, 03, 12, 13 and 20, 21, 30, 31 to the matrix-vector multiplications compared with other elements is relatively small, so we can define the following subsets of linear system based on the self-body interactions.

$$\begin{pmatrix} \bar{\varphi}_{00} & 0 \\ 0 & \bar{\varphi}_{11} \end{pmatrix} - \begin{pmatrix} L_1^{00} & L_2^{01} \\ L_3^{10} & L_4^{11} \end{pmatrix} \begin{pmatrix} \bar{\varphi}_{00} \\ \bar{\varphi}_{11} \end{pmatrix} = \begin{pmatrix} F_{00} \\ F_{11} \end{pmatrix}, \tag{3.35}$$

$$\begin{pmatrix} \bar{\varphi}_{22} & 0 \\ 0 & \bar{\varphi}_{33} \end{pmatrix} - \begin{pmatrix} L_1^{22} & L_2^{23} \\ L_3^{32} & L_4^{33} \end{pmatrix} \begin{pmatrix} \bar{\varphi}_{22} \\ \bar{\varphi}_{33} \end{pmatrix} = \begin{pmatrix} F_{22} \\ F_{33} \end{pmatrix}, \tag{3.36}$$

where

$$\varphi_{ii} = \bar{\varphi}_{ii} + \delta\varphi_{ii} \tag{3.37}$$

and $i = 0, 1, 2, 3$, that is, the potential φ can be separated by the potential of the self-interaction $\bar{\varphi}$ and the perturbation due to the second body interaction $\delta\varphi$. Inputting Eq. (3.37) to Eq. (3.34) and using the definition from Eq. (3.35) and Eq. (3.36) give new system of linear equations as,

$$I \begin{pmatrix} \delta\varphi_{00} \\ \delta\varphi_{11} \\ \delta\varphi_{22} \\ \delta\varphi_{33} \end{pmatrix} - \begin{pmatrix} L_1^{00} & L_2^{01} & L_1^{02} & L_2^{03} \\ L_3^{10} & L_4^{11} & L_3^{12} & L_4^{13} \\ L_1^{20} & L_2^{21} & L_1^{22} & L_2^{23} \\ L_3^{30} & L_4^{31} & L_3^{32} & L_4^{33} \end{pmatrix} \begin{pmatrix} \delta\varphi_{00} \\ \delta\varphi_{11} \\ \delta\varphi_{22} \\ \delta\varphi_{33} \end{pmatrix} = \begin{pmatrix} L_1^{02}\bar{\varphi}_{22} + L_2^{03}\bar{\varphi}_{33} \\ L_3^{12}\bar{\varphi}_{22} + L_4^{13}\bar{\varphi}_{33} \\ L_1^{20}\bar{\varphi}_{00} + L_2^{21}\bar{\varphi}_{11} \\ L_3^{30}\bar{\varphi}_{00} + L_4^{31}\bar{\varphi}_{11} \end{pmatrix}. \tag{3.38}$$

Using the same assumption to have Eqs. (3.35) and (3.36), this linear system can be reduced to the following two linear systems with the order $O(N^2)$.

$$\begin{pmatrix} \delta\varphi_{00} & 0 \\ 0 & \delta\varphi_{11} \end{pmatrix} - \begin{pmatrix} L_1^{00} & L_2^{01} \\ L_3^{10} & L_4^{11} \end{pmatrix} \begin{pmatrix} \delta\varphi_{00} \\ \delta\varphi_{11} \end{pmatrix} = \begin{pmatrix} L_1^{02}\varphi_{22} + L_2^{03}bar\varphi_{33} \\ L_3^{12}\varphi_{22} + L_4^{13}\varphi_{33} \end{pmatrix}, \tag{3.39}$$

$$\begin{pmatrix} \delta\varphi_{22} & 0 \\ 0 & \delta\varphi_{33} \end{pmatrix} - \begin{pmatrix} L_1^{22} & L_2^{23} \\ L_3^{32} & L_4^{33} \end{pmatrix} \begin{pmatrix} \delta\varphi_{22} \\ \delta\varphi_{33} \end{pmatrix} = \begin{pmatrix} L_1^{20}\varphi_{00} + L_2^{21}\varphi_{11} \\ L_3^{30}\varphi_{00} + L_4^{31}\varphi_{11} \end{pmatrix}. \tag{3.40}$$

To solve the system of linear equations in Eq. (3.33), first we solve the linear systems of self-interactions in Eq. (3.35) and Eq. (3.36). Then the right-hand side vectors in Eq. (3.39) and Eq. (3.40) are obtained by matrix-vector products between the previous solution vectors from the self-interactions and the matrix elements from the two separate bodies. The perturbations $\delta\varphi$ are computed after solving two systems of linear integral equations in Eq. (3.39) and Eq. (3.40). And we can get the new solution φ by the sum of the self-interaction and the perturbation in Eq. (3.37). By solving Eq. (3.39) and Eq. (3.40) with inputting new φ iteratively, the change of potential $\delta\varphi$ can be finally obtained when the change of the solution of linear system is converged to the given tolerance value after each iteration. In this iterative method, we only need one matrix-vector product operation between two separated bodies in each iteration. This iteration is called the "outer" iteration to separate the term with the "inner" iteration which is used to solve the single linear system with the iterative solver, such as GMRES. The "outer" iteration can reduce the size of system from $O(2N \times 2N)$ to $O(N \times N)$ and the "inner" iteration can be accelerated by introducing the fast multipole method(FMM) (Kim et al., 2010). Figure 3.3 shows how the double tree structures are defined to cover one body in one tree and the interactions between two separated bodies are allowed in FMM algorithm to calculate matrix-vector products in Eq. (3.39) and Eq. (3.40) to calculate the right-hand side vectors.

This double-tree FMM with "outer" iterative method has an advantage that can

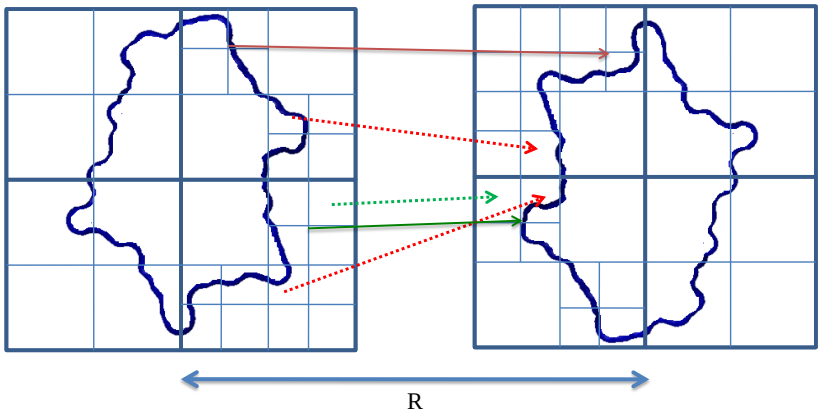


Figure 3.3 Schematic illustration showing the double tree fast multiple method(dt-FMM). Two trees are located with the center-to–center distance(R) separation. On level=2, all the Multipole–to-Local translations(M2L) are computed for far-field interactions. On level=3, the long interaction(solid red) is not allowed in the M2L translation list(the interaction list) but the interaction(solid green) is allowed. On level=4, also long interactions(red dashed line) are not allowed but the interaction within the interaction list(dashed green) is computed.

reduce the computational resource from the traditional direct boundary element method, $O((2N)^2)$ to the one of the single body problem, $O(N)$. But one possible drawback is the closest distance between two bodies. The center-to-center distance of two bodies should be longer than the size of tree structure in any dimension. No overlap of trees is allowed in this double-tree FMM. For example, the closest center-to-center distance between two BPTI proteins in the crystal lattice structure is about the range in 24-28$\mathring{A}$, but it should be more than 33$\mathring{A}$ in double-tree FMM to avoid the tree overlapping. The accuracy of the double-tree FMM is going to be worse if two trees are getting closer(see Figure 3.7). In this case, the number of "outer" iteration is also getting increase. Thus, the overall performance will be slower. In general, the double-tree FMM is useful when the center-to-center distance is about 1.5-2 times longer than the size of tree.

3.3.5 Solving the linear system: the single-tree Fast Multipole Method

In order to calculate the interaction energy when two bodies are too close to be calculated by the double-tree FMM, we introduce the single-tree FMM in Figure 3.4. This method is based on the single body FMM (Kim et al., 2010). The system of linear integral equations in Eqs. (3.6), (3.7), (3.8) and (3.9) for the electrostatic interaction and Eqs. (3.28), (3.29), (3.30) and (3.31) for the van der Waals interaction should be reduced to the equations of a single body. One problem we need to solve is the additional negative signs of L_1^{02}, L_3^{12}, L_1^{20} and L_3^{30} in Eq. (3.34) where the signs of gradients are changed because of the convention for the outside of the cavity. So we need to consider this sign change when the integral is performed by the surface on the second body when the source is in the first body. In the traditional single body FMM, there is no way to deal with this conventional change, but this problem can be solved by transferring the additional information of the ownership of surface elements during the process of Multipole-to-Multipole(M2M) and Local-to-Local(L2L) translations. Figure 3.5 shows the details how the ownership of each surface element in a leaf cell can be transferred to the parent's cell in FMM.

Because the single-tree FMM is based on the single-body FMM, the computational cost follows the order $O(2N)$, that is about twice more than the one of the double-tree FMM algorithm. Even thought the single-tree FMM takes twice more memory than the double-tree FMM, this cost is still highly competitive compared with the traditional direct Boundary Element Method. Figure 3.6 shows that the direct BEM follows the quadratic increase via the number of surface elements and two FMMs follow only the linear increase via order $O(N)$ or $O(2N)$ for the double and single-tree FMM respectively.

To test both FMM methods, we applied them to the electrostatic interaction energy calculation of two identical spherical particles. According to Figure 3.7, both solutions gave correct effective electrostatic interaction energies compared with the analytic so-

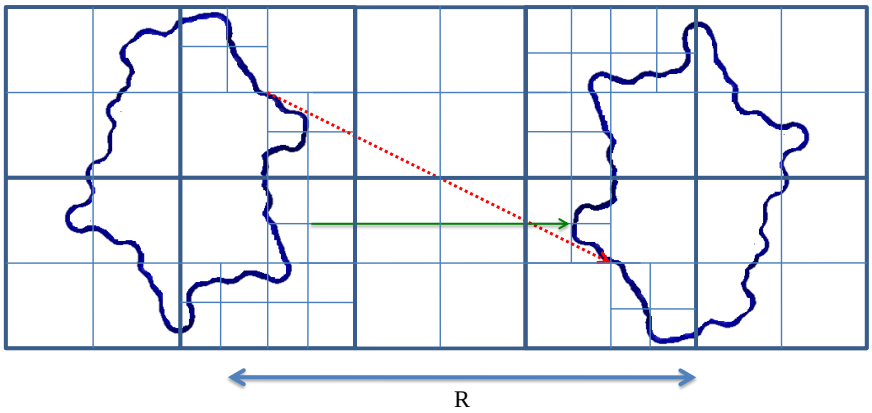


Figure 3.4 Schematic illustration showing the single tree fast multiple method(st-FMM). Only one tree is located to cover two surface sets of proteins with the center-to-center distance(R) separation. On level=2, only the Multipole-to-Local translations(M2L) which are in the interaction list(solid green) are computed but the long interaction(dashed red) is not allowed for the M2L translation.

lution of two identical spheres based on Eq. (A.13) in the appendix. Also we had the consistent results by two FMM methods when the effective electrostatic interaction energies between the two BPTI molecules are computed. Also these results were compared to the result from the direct BEM solver and we found that the single-tree FMM is slightly more accurate when two particles are getting closer and the double-tree FMM is more accurate when two particles are farther than the twice of the size of a particle. So we used both FMM methods to calculate the effective interaction energy between two protein molecules.

3.3.6 Preparation of protein molecules

The Bovine pancreatic trypsin inhibitor(BPTI) is useful to validate our model to calculate the osmotic second virial coefficients of a protein in a solution because it is relatively small protein(the number of residues are 58), the structure is well-known and the experimental B_2 data is presented (Farnum and Zukoski, 1999). We use the anisotropic patch model (Vega et al., 2008) and extend this patch model with treat-

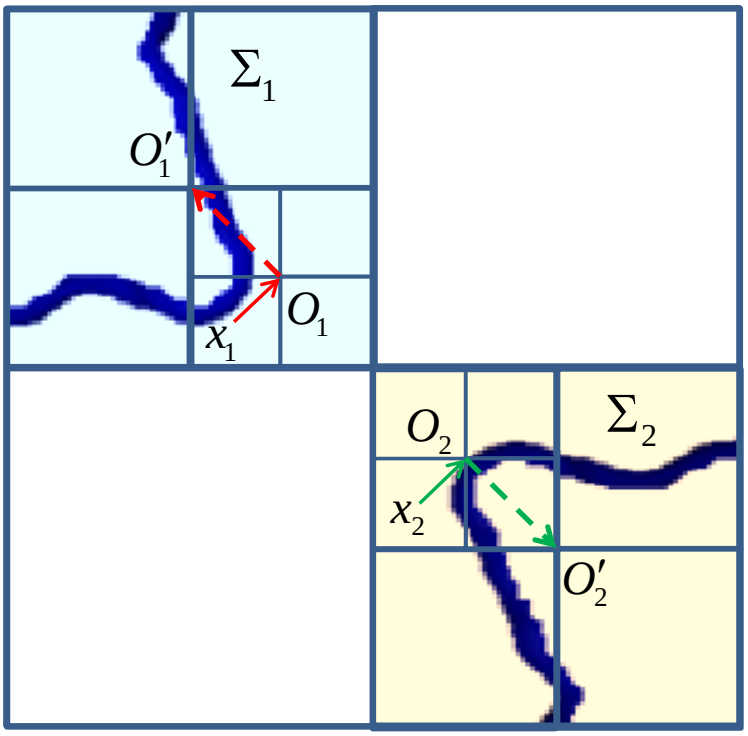


Figure 3.5 Schematic illustration showing the single-tree fast multiple method(st-FMM) in level=2 to level=5. $\sum_1$ and $\sum_2$ are the surfaces of two proteins. All cells with light blue shade belong to the surface $\sum_1$ and cells with light yellow shade belong to the surface $\sum_2$ respectively. From the lowest level, level=5, the surface index(either 1 or 2) is transferred from the level=5 center x_1 or x_2 to the level=4 center O_1 or O_2 by Multipole-to-Multipole(M2M) translations. This index also can be transferred to the upper level's cell. For example, on level=3 the center O_1' or O_2' has the surface index during the process of M2M translations. The red arrows indicate the flow of the surface index 1 and the green arrows for the surface index 2. The dashed arrows represent level=5 to level=4 M2M translations and solid arrows, level=4 to level=3 M2M translations respectively.

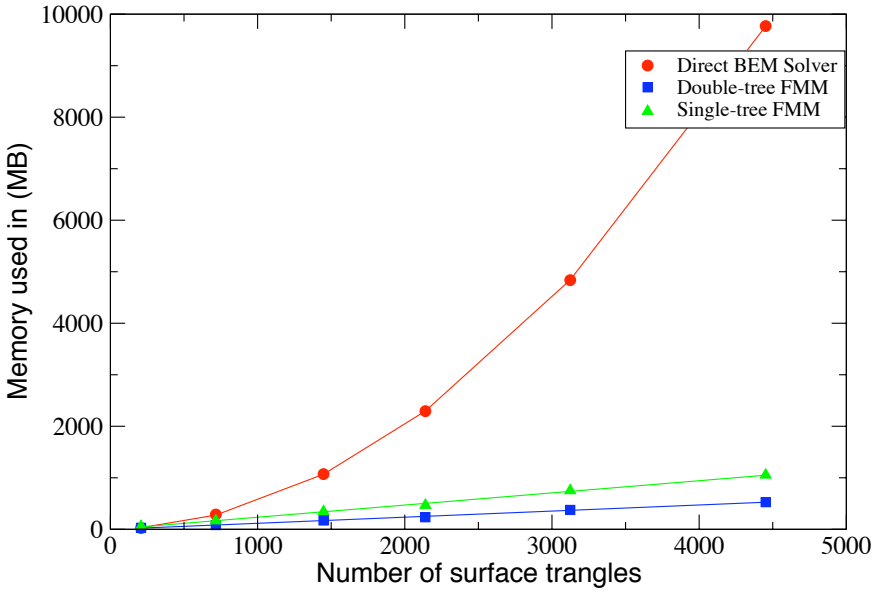


Figure 3.6 Memory cost comparison between the direct Boundary Element Method(BEM) in red circle, the double-tree FMM(blue square) and the single-tree FMM(green upper triangle). The number of surface elements indicated is the number of surface elements from a single protein(N). So the order of each method is $O((2N)^2)$ for the direct BEM, $O(N)$ for the double-tree FMM and $O(2N)$ for the single-tree FMM respectively.

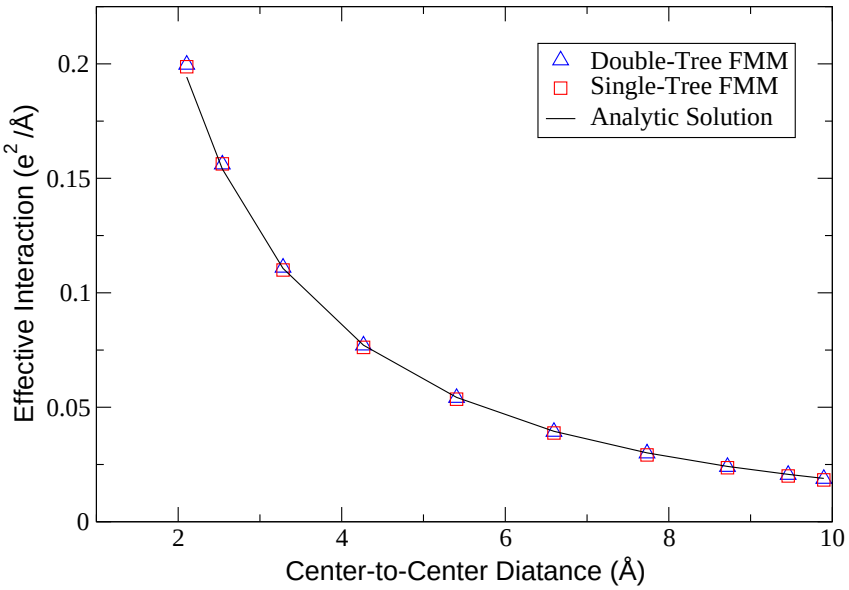


Figure 3.7 Effective electrostatic energy comparison between the analytic solution(solid line) in Eq. (A.13) and the solutions of the double-tree FMM(upper blue triangle) and the single-tree FMM(red square). The radius of both spheres is $1.0\mathring{A}$ and the unit charge is located in each sphere center. The inverse Debye screening length is $0.1\mathring{A}^{-1}$ and the dielectric constant is 1.0 inside the spheres and 10.0 outside spheres.

ing surface elements as patches to define the anisotropic interaction pairs between two protein molecules. Because of the large number of patches on the protein surface, it is really hard work to compute interaction energies of all pairs. To reduce the number of orientation pairs of surface patches between two protein molecules, we should consider the most probable configurations of pair interactions between two BPTI molecules. For this reason the crystal structure(PDB code=6PTI) information is used. According to this information the crystal space group of BPTI is $P2_12_12$. We use the transformation matrix given by this PDB file to apply the linear transformations from the original structure, A to generate other unit cell elements, B, C and D. Figure 3.8 shows how BPTI molecules are located in the unit cell of crystallography. We use the relations between unit cell elements to set up the pair configurations to calculate the interaction energies. The center-of-mass of the element A is located on the origin and all other elements are translated to the origin of A, then the interaction energy can be calculated when the second element, for example, B is translated to the direction of its original center-of-mass, $(\bar{x}, \bar{y}, z)$ as in the space group operation, that is an AB pair configuration, or to the opposite direction $(x, y, \bar{z})$ to have additional AB$'$ pair configuration. From this PDB structural information we have all six pairs of interactions, AB, AC, AD, AB$'$, AC$'$ and AD$'$. Figure 3.9 describes the relative orientations of BPTI elements in a unit cell.

We set up an effective model for the interaction free energy of two proteins based on the residual model. A single sphere represents an amino acid residue in a protein molecule where the radius of each residual sphere is calculated from the volume of amino acid in a solution (Zamyatnin, 1984). CHARMMING web portal (Miller et al., 2008) is used to prepare the topology and coordinate files for each protein element to generate input files for CHARMM force field (Brooks et al., 1983) with which we can calculate the position of the center of each residue in a protein molecule and its dipole moment. Compared with the atom based model, this model has two major advantages; the number of problem to be solved is reduced from the number of atoms in protein to

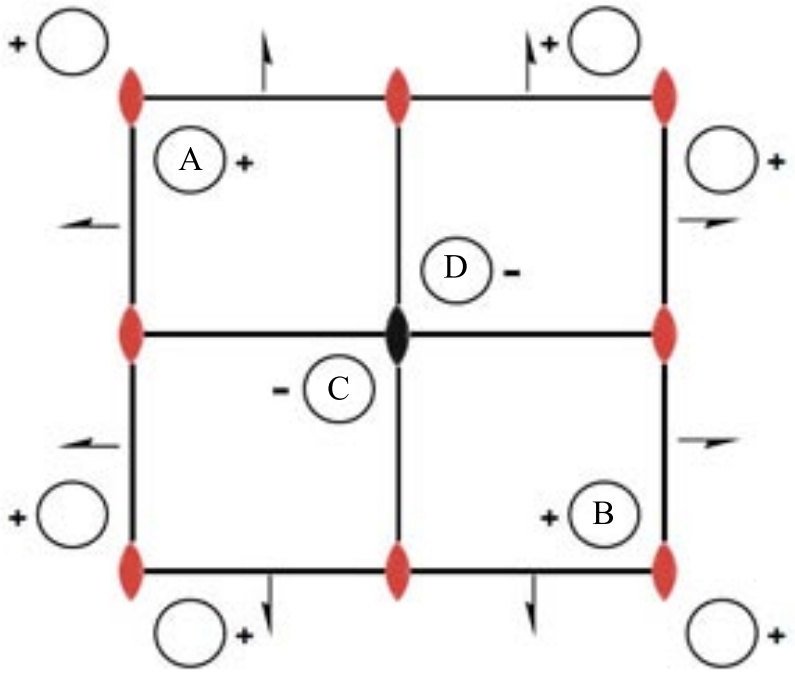


Figure 3.8 2D illustration shows the unit cell of the point group $P2_12_12$. In unit cell, there are four elements indicated by the capital letters: A is on the origin of coordinates and its symmetrical operation is (x, y, z), B can be obtained by the operation $(\bar{x}, \bar{y}, z)$, C can be obtained by the operation $(1/2 + x, 1/2 + \bar{y}, \bar{z})$ and D can be obtained by the operation $(1/2 + \bar{x}, 1/2 + y, \bar{z})$. All the notations follows the Hermann-Mauguin symmetry notation and the style of Wondratschek and Müller (Wondratschek and Müller, 2002). This diagram is taken from Jasinski and Foxman (Jasinski and Foxman, 2007).

the number of residues and the suitable number of triangulization of the protein surface is also decreased (Kim et al., 2010).

The calculations of the osmotic second virial coefficients of the BPTI protein in solutions use the same conditions from the experiment (Farnum and Zukoski, 1999). The temperature of the solution is $20°C$ which is used both in the integral equation of B_2 in Eq. (3.4) and in the inverse Debye screening length. The pH of the solution is also set to 4.9 and is used to calculate the charge of each amino acid residue in a protein in this pH condition by using the Henderson-Hasselbalch equation and the pK_a of a residue calculated by PROPKA 2.0 (Delphine et al., 2008). The generic pK_a values of amino acids are not used because the local pK_a of a residue which is either burred inside the protein or on the surface of the protein may have a shifted pK_a as the P_1 Glu and P_1 Asp mutations of the BPTI-trypsin complexes (Kim et al., 2010) and the PROPKA 2.0 is the most accurate program for the pK_a prediction (Davies et al., 2006). The dependence of B_2 of BPTI molecules on the concentration of the sodium chloride solution and the comparison with the experimental B_2 data will be described in the result section on this paper.

In addition to the calculations of the second virial coefficients of Bovine pancreatic trypsin inhibitor(BPTI) as a function of the concentration of the sodium chloride solution, we calculate the osmotic second virial coefficients of the lysozyme protein in various conditions. To set up the most probable configurations of pair interactions between two lysozyme molecules, the crystal structure(PDB code=2ZQ3) information is used. According to this file the crystal space group of lysozyme is $P2_12_12_1$. We use the transformation matrix given by this PDB file to apply linear transformations from the original structure, A, to generate other unit cell elements, B, C and D. Figure 3.10 shows how the lysozyme molecules are located in the unit cell of crystallography and we use the relations between unit cell elements to set up pair configurations to calculate the interaction energies. The center-of-mass of element A is located on the origin and

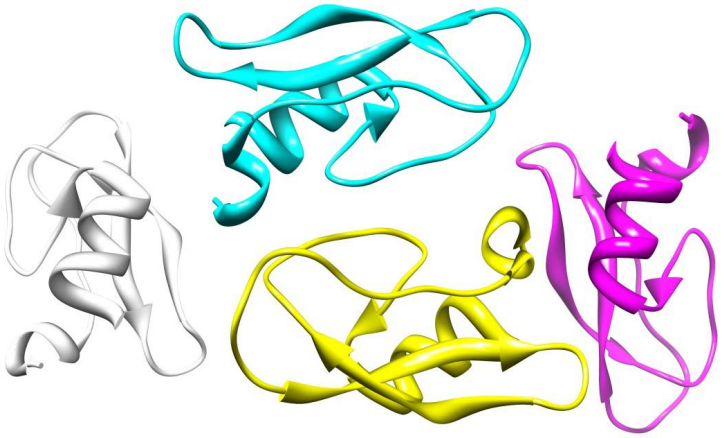


Figure 3.9 3D illustration shows the relative directional orientations of all BPTI elements in a unit cell of $P2_12_12$. White colored ribbon structure indicates the element A, pink, blue and yellow ribbons are the element B, C and D, respectively. UCSF Chimera (Pettersen et al., 2004) was used to draw this figure.

all other elements are translated to the origin of A, then the interaction energies can be calculated when the seconde element, for example, B is translated to the direction of its original center-of-mass, $(1/2 + \bar{x}, \bar{y}, 1/2 + z)$ as in the space group operation, that is an AB pair configuration, or to the opposite direction $(-1/2 + x, y, -1/2 + \bar{z})$ to have additional AB′ pair configuration. From this PDB structural information we have all six pairs of interactions, AB, AC, AD, AB′, AC′ and AD′. Figure 3.11 describes the relative orientations of Lysozyme elements in a unit cell.

The calculations of the osmotic second virial coefficients of the lysozyme in solutions use the same conditions from the Static Light Scattering(SLS) experiment (Guo et al., 1999). According to their experimental data, all second virial coefficients are computed by following conditions. The concentration dependence from 2% to 7% of salt concentration, the pH dependence from pH = 4.0 to pH = 5.4 and the temperature dependence from $25°C$ to $5°C$ are calculated in the sodium chloride solution. The concentration dependence from $0.50M$ to $1.10M$ of the ammonium chloride solution is calculated at

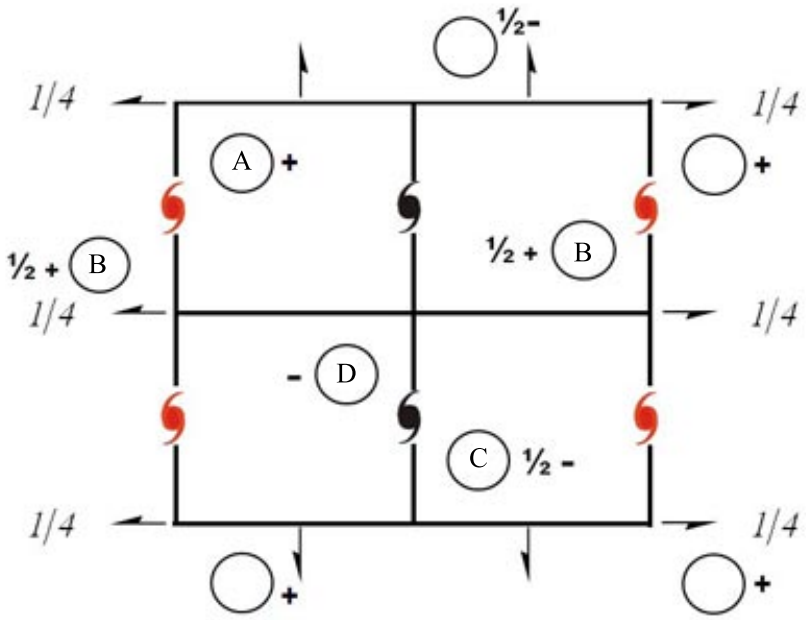


Figure 3.10 2D illustration shows the unit cell of the point group $P2_12_12_1$. In unit cell, there are four elements indicated by the capital letters, A is on the origin of coordinates and its symmetrical operation is (x, y, z), B can be obtained by the operation $(1/2 + \bar{x}, \bar{y}, 1/2 + z)$, C can be obtained by the operation $(\bar{x}, 1/2 + y, 1/2 + \bar{z})$ and D can be obtained by the operation $(1/2 + x, 1/2 + \bar{y}, \bar{z})$. All notations follows the Herman-Mauguin symmetry notation and the style of Wondratschek and Müller (Wondratschek and Müller, 2002). This diagram is taken from Jasinski and Foxman (Jasinski and Foxman, 2007).

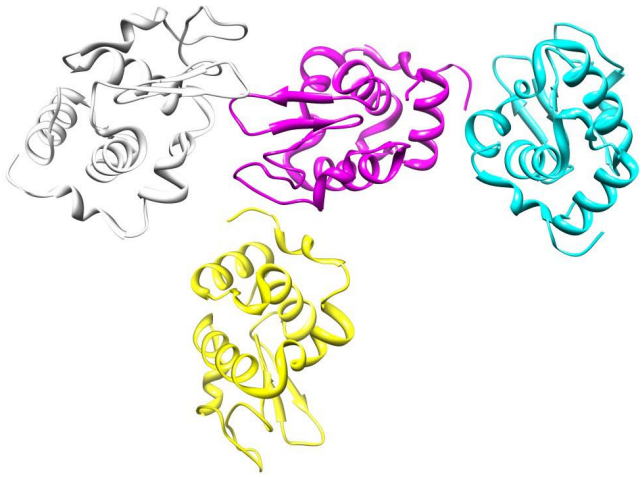


Figure 3.11 3D illustration shows the positions and relative directional orientations of all the lysozyme elements in a unit cell of $P2_12_12_1$. White colored ribbon structure indicates the element A, pink, blue and yellow ribbons are the element B, C and D, respectively. UCSF Chimera (Pettersen et al., 2004) was used to draw this figure.

pH = 4.5 and temperature $18°C$. Finally, the concentration dependence from $0.10M$ to $0.70M$ of the magnesium bromide solution at pH = 7.8 and temperature $23°C$. Comparisons between calculated B_2 and the experimental B_2 are made using the experimental data mostly from the SLS experiment (Guo et al., 1999) and additional data for the magnesium bromide salt condition from the Self-Interaction Chromatography(SIC) experiment (Tessier et al., 2002). The experimental data of the second virial coefficients in the magnesium bromide salt condition and calculated data shows the limitation of our model based on the DLVO theory and the experimental method.

3.4 Results

The electrostatic interaction energies and the van der Waals interaction energies between two BPTI molecules are calculated by the single-tree FMM algorithm when the center-to-center distance between two protein is within the twice of the size of the

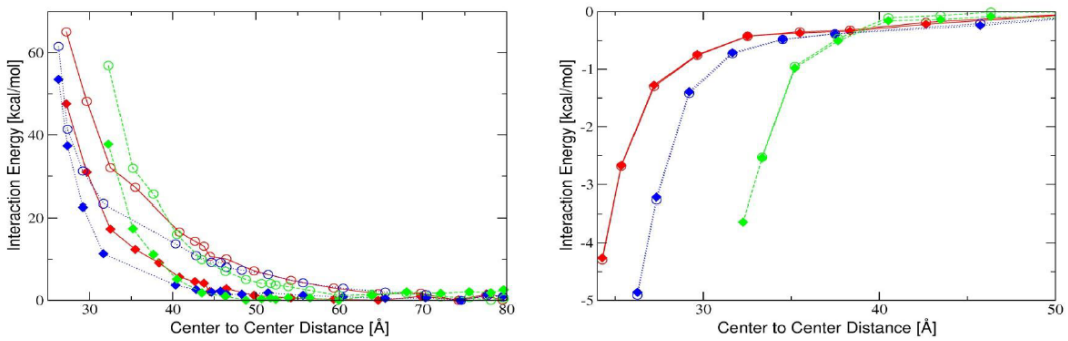


Figure 3.12 The graphs of the electrostatic interaction energies(left) and the van der Waals interaction energies(right) between two BPTI molecules are shown above. Each pair configuration is indicated by red solid line, blue lightly dashed line and green dark dashed line for AB, AC, AD configuration respectively. Two curves for each pair configuration are shown: the open circle indicates the interaction energies from 2% NaCl solution and the color-filled diamond indicates the energy from 7% NaCl solution. Because of the three-dimensional structure of BPTI protein, the starting distance of the single-tree FMM calculation for each pair interaction is different based on the excluded distance.

protein and by the double-tree FMM when the center-to-center distance is far. Figure 3.12 shows the interaction energy changes in the center-to-center distance R between two protein molecules and different pair configurations including the dependence on the inverse debye-Hückel screening length κ which represents the concentration of the NaCl solution. With this calculated interaction energies, we can get calculated B_2 by integrating Eq. (3.3).

In order to validate our assumption that the pair configuration sampling from crystal space group operations can represent the whole angular and directional dependence on Eq. (3.2) and Eq. (3.3) and finally Eq. (3.4) is a good approximation of the Eq. (3.2), the second virial coefficients of the BPTI protein were calculated by solving Eq. (3.4). We calculated the B_2 contributions from each pair interaction with all six pair configurations(in this case $p = 6$) and solved Eq. (3.4) to show our result finally. Figure

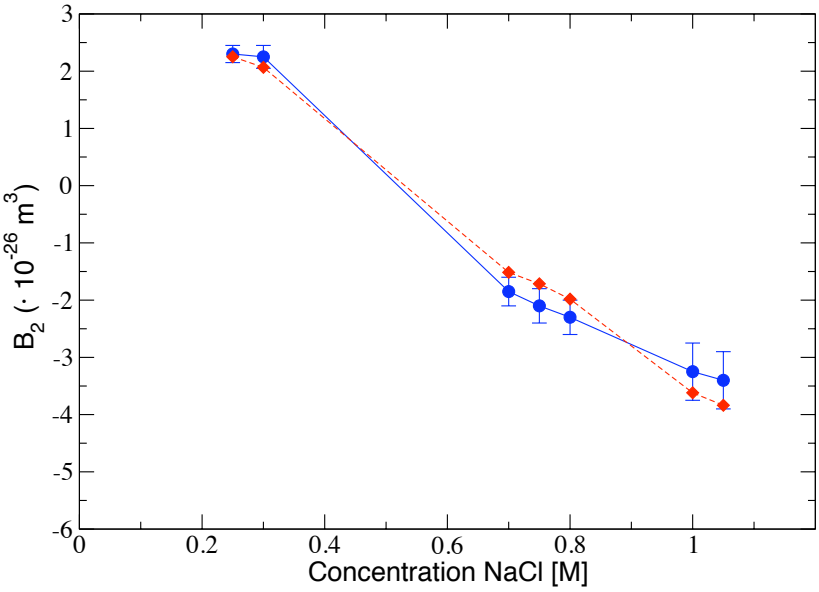


Figure 3.13 The NaCl concentration dependence of the osmotic second virial coefficients of BPTI protein is shown above. The solid blue line with circles indicates the experimental B_2 from Farnum and Zukoski and the dashed red line with diamonds shows our calculated result. The error bars for the experimental data is from the literature (Farnum and Zukoski, 1999).

3.13 shows the NaCl concentration dependence of the osmotic second virial coefficients of the BPTI protein molecule with the experimental data and our calculations. The error bars for the experimental data is from the experimental data (Farnum and Zukoski, 1999). And the result from six pair configurations to the Eq. (3.4) are used to indicate the calculated data. The linear fit correlation coefficient between observed B_2 and calculated B_2 is 0.9552. The variations of the calculated B_2 from observed values are relatively large at high concentrations of NaCl solution. This is because the calculated B_2 data above $1M$ of NaCl concentration is overestimated by our model and causes the linear fit correlation to the experimental data worse. This is an evidence of the limitation of our calculation model that shows the breakdown of the DLVO theory based on the Debye-Hückel theory.

The second virial coefficients of the lysozyme molecule are calculated by the same manner as the calculations of the BPTI protein. The more common form of the second virial coefficient, $A_2(ml \cdot mol/g^2) = B_2(m^3)N_A/{M_w}^2$ is used to report our calculated results and compare with the experimental data (Guo et al., 1999). The averaged B_2 is calculated by using Eq. (3.4) with six different pair configurations based on the crystal space group operations of $P2_22_12_2$. Figure 3.14 shows the experimental data and calculated results of the second virial coefficients from the given solution conditions.

In Figure 3.14(A), the experimental and the calculated B_2 are given as a function of the concentration of the NaCl solution and other conditions remain constant at pH 4.2 and $25°C$. In general the correlation between the experimental and calculated results are obvious, but we also can see the limitation of this model for the high concentration of electrolyte solution, at $7\%(w/v)$ of NaCl solution just as we have the same behavior on the calculation of the BPTI protein. Except the highest concentration result, the linear fit correlation between the observed and calculated data of B_2 increases from 0.8282 to 0.8875.

The B_2 behaviors as a function of the pH of solution in NaCl solution in Figure 3.14(B) shows a reasonable agreement between the experimental and calculated data even though experiments show that slight increase at pH = 5.2. The experiments and calculations are performed at $25°C$ and under 2.0% NaCl concentration. The temperature dependence of B_2 clearly shows that the calculated result has obvious correlation with the experimental data. This dependence also has an exception on the low temperature at $5°C$ about the increment of B_2 in the experiment. However, according to the relation between observed B_2 values and the solubilities of the lysozyme in solutions (Gripon et al., 1997), the solubility of lysozyme shows clearly decrease as the calculated B_2 decreases at this condition. Temperature is the only variable in this relation, so the solution conditions such as pH, 4.2 and the concentration of salt(2.0% NaCl) are remained as constants.

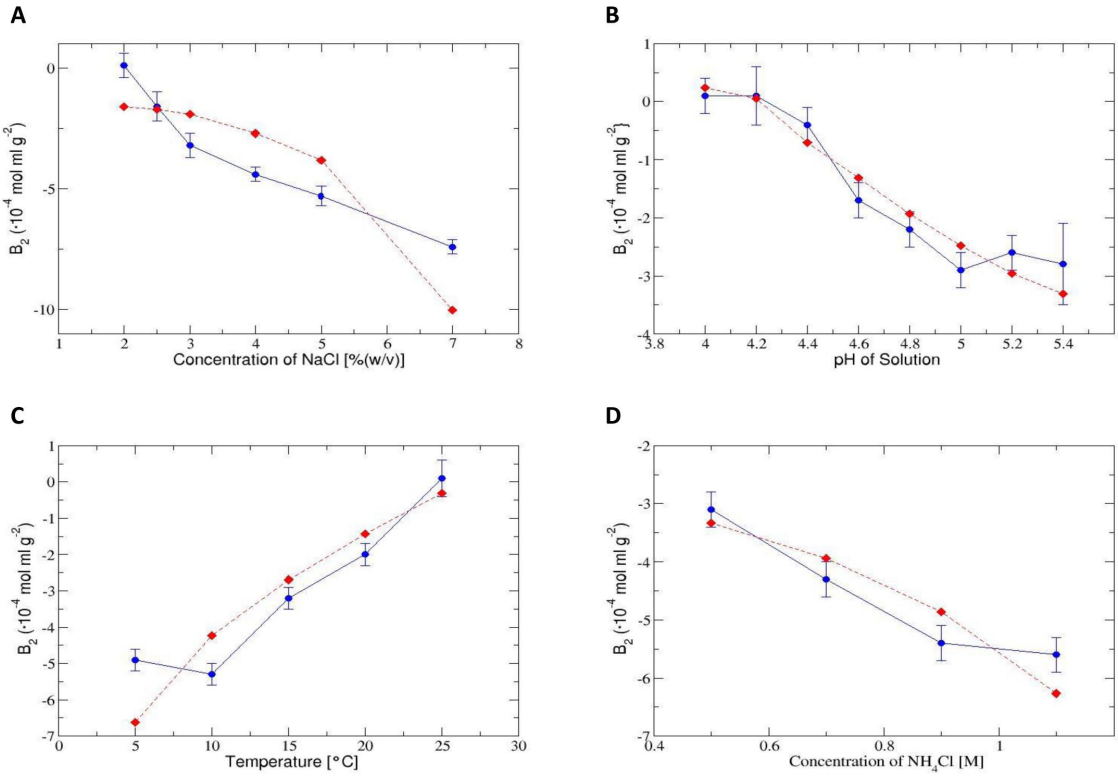


Figure 3.14 The relations between the experimental B_2 (Guo et al., 1999) and the calculated B_2 of the lysozyme protein with the given solution conditions. The dependence of B_2 on NaCl concentration is shown on (A). The pH dependence is on (B). The temperature dependence is on (C). And dependence upon ammonium chloride concentration is shown on (D). The solid blue lines with circles indicate the experimental data and the dashed red lines with diamonds indicate our calculated results. The linear fit correlation coefficients between the experimental data and calculated results are 0.8282, 0.9636, 0.8981 and 0.9016 for (A), (B), (C) and (D) respectively. To draw error bars, we use the experimental data from the literature (Guo et al., 1999).

Finally, in Figure 3.14(D), the experimental and calculated B_2 are given as a function of the concentration of the ammonium chloride solution. We also can see the limitation of this model for the high concentration above $1M$ of NH_4Cl solution. Except the highest concentration result, the linear fit correlation between the observed and calculated data of B_2 increases from 0.9016 to 0.9904. This is done under pH 4.5 and $18°C$ conditions.

3.5 Discussions

3.5.1 Temperature effect on the second virial coefficient of the lysozyme

The temperature of a solution affects the calculations of the second virial coefficients of protein molecules both on the inverse Debye-Hückel screening length κ and the integrand in Eq. (3.2). In general, the decreased temperature raises κ and lower the B_2 and the relation between the temperature change and the B_2 change is almost linear(see Figure 3.14(C)). Another temperature effect is represented by the change of the dielectric constant of water which is the parameter in the dielectric continuum model in our calculations. From $25°C$ to $0°C$, the dielectric constant increases from 80 to 88 (Murrell and Jenkins, 1994) and according to Harvey and Lemmon this increase also gives a decreasing effect on the second virial coefficients under low temperature, $T < 350K$ (Harvey and Lemmon, 2004). The predicted B_2 from our calculations shows the correct correlation to the above statement and observed data (Guo et al., 1999) of lysozyme calculations. But the observed second virial coefficient shows unusual effect by the temperature at $5°C$, it is increased not decreased by decreasing temperature but our calculation still predicts the decrease of B_2.

From the structural study of the lysozyme crystal, the unusual effect of temperature was seen at $280K$ structure (Kurinov and Harrison, 1995). The number of water molecules under $4\mathring{A}$ range, the distance from the lysozyme protein surface in this structure, are smaller than in either the higher temperature($T > 295K$) or the lower temper-

ature ($T < 250K$) structures. The less number of waters can cause the less interactions between water molecules and atoms on the protein surface. This can be a possible reason that the second virial coefficient at $5°C$ is observed as an abnormal behavior considering the overall trend with the temperature changes.

3.5.2 The limitation of model: Debye-Hückel Theory

In Figure 3.13 and Figures 3.14(A) and (D), the calculated B_2 at high concentration of salt of both sodium chloride and ammonium chloride are highly overestimated and causes the linear fit correlations to the experimental values much worse. According to our calculations this over-estimation occurs at the high concentration of an ionic solution whose ionic strength is greater than $1M$ and the inverse Debye-Hückel screening length κ is large (> 0.1). At such higher a concentration, the Debye-Hückel theory fails and the DLVO theory which is based on the Debye-Hückel theory and the base of our model follows. According to Fawcett, one important reason for the failure of this theory is that it considers the only central ion has a size and ignores the size of the other ions in the outside atmosphere (Fawcett, 2004). As a result, the thickness of the outside ionic atmosphere is underestimated at high concentration of solutions. And in constant volume condition the extra work involves when introducing the additional electrolyte ions into the highly concentrated solution, but this work is neglected in this theory. Also the structure of the solvent water is strongly affected by ion-solvent interactions. If the concentration of ions increases the fraction of water molecules which are associated with ions increases and the dielectric permittivity of this solution can decrease. So it should be noted that change of the Debye-Hückel constant with the electrolyte concentration should reflect the corresponding change of the dielectric constant of solvent, ε.

This limitation is more dominant especially when the ionic salt is changed to the divalent ion such as magnesium bromide. Figure 3.15 shows the extreme case of the failure of our model based on the Debye-Hückel theory. The observed second virial

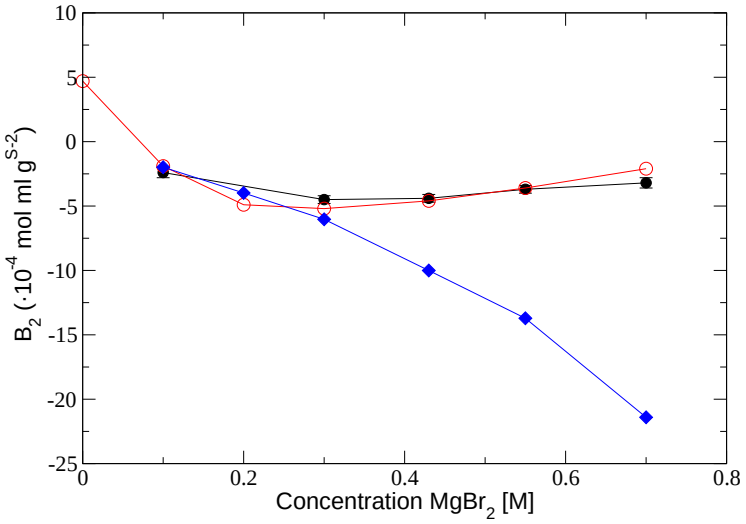


Figure 3.15 The $MgBr_2$ concentration dependence of the osmotic second virial coefficients of the lysozyme protein at pH 7.8 is shown above. The solid black line with black colored circles are measured by the Static Light Scattering(SLS) (Guo et al., 1999), and the solid red line with open circles are from the Self-Interaction Chromatography(SIC) (Tessier et al., 2002). The blue solid lines with diamonds are our calculations. Both observed results of B_2 become more positive at higher ionic strength. But the calculated results do not show the increase of the second virial coefficients at high ionic strength of magnesium bromide solutions.

coefficients of lysozyme show the minimum at the concentration of $MgBr_2 \sim 0.3M$, and start increasing as the ionic strength increases. Both experimental results from the Static Light Scattering(SLS) (Guo et al., 1999) and the Self-Interaction Chromatography(SIC) (Tessier et al., 2002) agree closely. The calculations predict the decrease of B_2 as the concentration increases and agree with the experimental data only to the minimum point from the experiments. But at high concentration of $MgBr_2$, the calculations only predict the second virial coefficients decrease to the large negative value and at this point the inverse Debye-Hückel screening length κ is already greater than 0.1.

It is obvious that applying the extended Debye-Hückel theory for the dielectric continuum model to the high concentration limitation of our model can reduce the overestimation of the second virial coefficients upon calculations and finally interpret the positive increase behavior in the $MgBr_2$ solution. The information of the frequency dependent dielectric function (Song, 2009) can be used to extract the effective Debye screening length and the corresponding effective dielectric function. This frequency dependent dielectric function is already applied to the formulation of the solution to the dynamical Poisson-Boltzmann equation with the system of linear equations of the boundary element integrals which correspond to Eqs. (3.6), (3.7), (3.8) and (3.9) for the electrostatic interaction and Eqs. (3.28), (3.29), (3.30) and (3.31) for the va der Waals interaction.

Another reason can be made that B_2 values become more positive at high concentration of $MgBr_2$ solution. The binding affinity of Mg^{2+} ions to the surface of lysozyme increases as the concentration of $MgCl_2$ increases (Grigsby et al., 2000; Arakawa et al., 1990). The extent of Mg^{2+} ion binding increases as the pH of the solution increases to the isoelectric point of the protein(for lysozyme, 9.2) because the net positive charge on the protein surface approaches zero at this point. The open active site residues of lysozyme are Glutamic acid(E53) and Aspartic acid(D70) and both are negatively charged at this pH condition and the overall net charge of lysozyme decreases from 13.3 at pH = 4.0 to 7.65 at pH = 7.8 under $23°C$ which is the condition used in the experiments and

our calculations. Due to the binding of Mg^{2+} divalent cations to the acidic residues of lysozyme, the repulsive interactions between lysozyme molecules increase then cause more positive second virial coefficients observed in both SLS and SIC experiments.

3.6 Concluding Remarks

The extended fast multipole method for two protein are implemented to solve the system of linear equation of the solution of the linearized Poisson-Boltzmann equation to calculate the effective interaction energy of both electrostatic and van der Waals contributions. The traditional Boundary Element Method (Atkinson and Han, 2009) implementation following Juffer et al. (Juffer et al., 1991) requires the computational cost both memory and time with the order $O((N + M)^2)$ if the number of surface elements from the first protein is N and from the second protein is M respectively. This computational cost is the bottleneck for the comprehensive study on the interactions between such large proteins. The two-body extended FMM algorithm to this problem overcomes this computational cost problem to the order of $O(N)$ if N is grater than M for the double-tree FMM with the additional outer iteration method and the order of $O(N + M)$ for the single-tree FMM. The double-tree FMM is suitable at the relatively large distance cases of interaction energy calculations and the single-tree method is good at shorter distance than about the twice of the size of protein molecule. The accuracy and performance of both methods can be controlled by adjusting the depth of trees, the number of expansion terms and the tolerance factor of iteration (Yoshida, 2001).

The Bovine Pancreatic Trypsin Inhibitor(BPTI) protein is used to validate our model for calculations of the osmotic second virial coefficients B_2. The orientation dependence of the interaction energy in the integral in Eq. (3.2) is simplified by using the pair configuration on the crystal space group operations and the integral is performed by the simple distance dependence with six pair configurations of interaction energies. The calculated

B_2 is generally agreed with observed values with changes in the NaCl concentrations. This model is also applied to calculate the second virial coefficients of lysozyme molecules under various solution conditions. Salt concentration dependence(sodium chloride, ammonium chloride), pH and temperature dependence and salt effects which are all important factor for the solubilities and crystal growing of proteins (George et al., 1997; Rosenbaum and Zukoski, 1996; Veesler et al., 1996) are carried and our calculation results show the evidence that our model is suitable to describe behaviors of the osmotic second virial coefficients of proteins under these solution conditions.

In the present work, this model can break down when calculating the second virial coefficient at high concentration of ionic salts and with multivalent ionic salts such as Mg^{2+} ion. Our results show the overestimation of B_2 when the ionic strength is greater than $0.1M$ in general and do not show the repulsive effect of the magnesium ion upon binding to the negatively charged amino acid residues, which causes the positive increase of B_2 even if the ionic salt concentration increases. This is the evidence that the limitation of the DLVO theory for the interaction of two particle base on the Debye-Hückel theory. Introducing the extended Debye-Hückel theory (Song, 2009) into this model can overcome such a limitation of high ionic strength problem.

CHAPTER 4. The phase diagram calculations of protein

4.1 Introduction

The phase diagram and protein crystallization are related directly through the location of the solubility curve. Crystals can only form in supersaturated solutions. Thus, knowing where the solubility curve is located is important to help to grow crystals for X-ray structure determination. The phase diagram is also important because it can describe the information about the interactions among all the components including the proteins and salts in the solutions. In the presence of liquid-liquid phase separation, the effective interactions between the protein molecules are attractive (Rosenbaum and Zukoski, 1996) and this attraction is a necessary condition for crystallization (George et al., 1997). The enthalpies and entropies of the proteins in the liquid and solid phases can be determined from the liquid-liquid coexistence curve and the solubility curve respectively (Petsev et al., 2003). Also the study of a phase diagram can help to find the possibilities to predict the suitable conditions under which proteins can be crystallized or to reduce the number of possible trials of finding the optimal solution conditions.

The effective interaction energies between two protein molecules as we discussed in chapter 3 are useful to predict the phase diagram with the given solution conditions. As a first step of the determination of a phase diagram, the anisotropic interaction potentials between two protein molecules should be calculated to obtain the free energy differences between two phases (Vega et al., 2008). In this chapter, we will present a model to calculate the pair interaction potential which is based on our anisotropic patch

model (Kim and Song, 2010) and how to reduce the computational effort to calculate it.

4.2 The anisotropic patch model

In the canonical ensemble the Helmholtz free energy A is given by the following equation (McQuarrie, 1976),

$$A = -k_B T \ln(Q(N,V,T)) = -k_B T \ln\left(\frac{q^N}{N!} \times \int \exp\left[-\beta U(\mathrm{r}_1,\ldots,\mathrm{r}_n,\Omega_1,\ldots,\Omega_n)\mathrm{d}1\ldots\mathrm{d}N\right]\right), \tag{4.1}$$

where $\beta = 1/(k_B T)$, U is the intermolecular energy of the whole system, q is the molecular partition function and $\mathrm{d}i$ is the abbreviated form of $\mathrm{d}\mathbf{r}_i\mathrm{d}\omega_i$ where $\mathrm{d}\mathbf{r}_i = \mathrm{d}x_i\mathrm{d}y_i\mathrm{d}z_i$ (Vega et al., 2008). The angle, Ω_i defines the orientation of the molecule i and the location of molecule i is given by the Cartesian coordinates x_i, y_i, z_i for each molecule located at $\mathbf{r}_i$.

The total potential in Eq. (4.1) can be written as a sum of the two-body interaction terms that depend on the center-to-center distance r_{ij} between two particles and their relative direction of orientations Ω_i (Noya et al., 2007).

$$U(\mathbf{r}_1,\ldots,\mathbf{r}_n,\Omega_1,\ldots,\Omega_n) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} V(r_{ij},\Omega_i,\Omega_j). \tag{4.2}$$

The interaction between two particles is described by a potential with an isotropic repulsive potential core and an anisotropic angular dependent potential,

$$V(r_{ij},\Omega_i,\Omega_j) = \begin{cases} V_{LJ}(r_{ij}) & r_{ij} < \sigma_{LJ} \\ V_{LJ}(r_{ij})V_{ang}(r_{ij}\Omega_i,\Omega_j) & r_{ij} \geq \sigma_{LJ} \end{cases}, \tag{4.3}$$

where $V_{LJ}(r)$ is the Lennard-Jones potential,

$$V_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma_{LJ}}{r}\right)^{12} - \left(\frac{\sigma_{LJ}}{r}\right)^{6}\right], \tag{4.4}$$

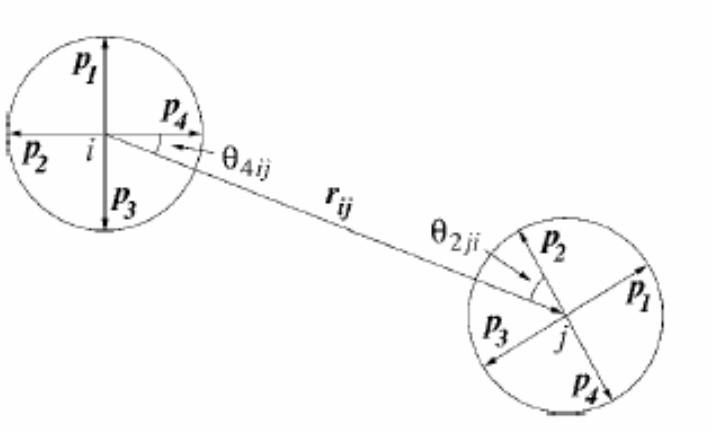


Figure 4.1 Schematic illustration shows the geometry of the interaction between two particles. Each particle has four patches in 2D space arranged evenly with the directions indicated by patch vectors, p_i. The inter-particle vector, $\mathbf{r}_{ij}$ defines the index of the patch in each particle used in the Eq. (4.5) based on the closest distance to this vector (Noya et al., 2007).

and ϵ is the depth of the pair potential well. According to Noya et al, its angular dependence of attractive potential is modulated by a product of Gaussian functions whose center is at the position of each patch (Noya et al., 2007),

$$V_{ang}(r_{ij}, \Omega_i, \Omega_j) = \exp\left(-\frac{\theta^2_{k_{min,ij}}}{2\sigma^2}\right) \exp\left(-\frac{\theta^2_{l_{min,ij}}}{2\sigma^2}\right), \tag{4.5}$$

where σ is the standard deviation of the Gaussian, $\theta_{k_{min,ij}}$ is the angle between patch k on molecule i and the inter-particle vector $\mathbf{r}_{ij}$ and k_{min} is the patch which minimizes the magnitude of this angle. The magnitude of the interaction depends on the magnitude of the patch angle. In Figure 4.1, a schematic illustration of this patch model(four patches on each spherical surface in two dimensional space) is shown for the interaction of two particles (Noya et al., 2007).

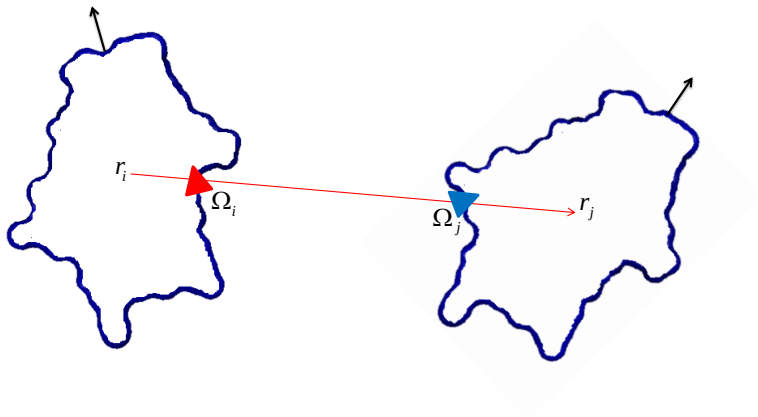


Figure 4.2 Schematic illustration shows the geometry of the interaction between two protein molecules with our patch model. Each protein molecule is discretized by a number of surface elements and rotated to the direction of a patch $\Omega_{i(j)}$ on protein 1(2). The inter-particle vector, $\mathbf{r}_i$ to $\mathbf{r}_j$ lies on two patch vectors indicate the center-to-center displacement of two protein molecules. The pair of patch indices, (i, j), finally represent the relative orientations of two protein molecules.

4.3 A residue level patch model for proteins

The simple patch model based on the spherical particle is not a good approximation for proteins to define the interactions between two protein molecules because of the complexity of the geometry of the protein molecules. One possible way to capture the realistic geometry of the protein surface is to construct a patch model based upon the boundary elements used in our residue level protein-protein interaction model. Discussed in chapter 3, the electrostatic and the van der Waals interaction energies can be calculated with different center-to-center separation distances and different orientations. Thus, we can calculate the pair interaction potential by arranging the protein molecule with the patch vector which is the vector from the center of a protein molecule to the designated boundary surface element along with the inter-particle vector whose magnitude is the center-to-center distance between two proteins(see Figure 4.2).

To represent realistic anisotropic interactions between two proteins, we have to consider a large number of patches, thus a high cost. To overcome this problem, we should think about two things. First, defining a surface of a protein molecule should be done by using a small number of triangles of the surface tessellation. Second, a method to reduce the number of calculations, both the electrostatic pair interactions and the van der Waals pair interactions with a patch model should be considered to reduce the overall costs. We propose an interpolation strategy to reduce the number of evaluations between various patches of two proteins.

To discretize the molecular surface of a protein molecule, we use the MSMS (Sanner, 1996). To achieve the first improvement to reduce the number of pathes, we need to discretize the molecular surface of a protein molecule using as small number of triangles as possible but maintaining the anisotropy of molecule. This can be done by controlling the parameters in MSMS. Using the minimum density of the surface tessellation and maximum size of the probe radius to define the molecular surface, we can get the smallest number of surface patches. For the calculations, 248 patches can be obtained to calculate the pair interaction potentials between two BPTI proteins and 780 patches are used to capture the pair interaction potentials between two lysozyme proteins.

Even though a small number of surface elements on the protein surface are generated, the number of pair interactions between two protein molecules is still very large. For example, the number we need to compute for the BPTI is at least 248×248. It is too time consuming to compute all the pair interactions. So the method designed to reduce the number of pair interaction calculations is to interpolate the functions of the pair interaction potentials based on the patch vectors which represent the orientations of the protein molecules. This can be done using a similar strategy as presented in chapter 3. The second virial coefficients are calculated using a small number of pair configurations of the two proteins from crystal space group operations. Calculations of only six pair orientations were enough to represent the good correlations with the

experimental values of the second virial coefficients from the small BPTI protein to the relatively large lysozyme protein even though the calculation of the second virial coefficient in Eq. (3.2) requires the contributions of all the possible pair interaction energies.

Starting number of pair interaction potentials in this patch model is six in BPTI protein calculations and their orientations are from the crystal space group, $P2_12_12$ operation(see Figures 3.8 and 3.9) and is also six in lysozyme protein calculations and their orientations are from the crystal space group, $P2_12_12_1$ operation(see Figures 3.10 and 3.11). The indices of the patches from the first method are assigned to the six pair interactions from the crystal group operations. In 780 patches of the lysozyme surface, the six computed pair interactions have the pair indices (i, j) as $(604, 229)$, $(218, 401)$ and $(62, 175)$ for AB, AC and AD pairs respectively and $(76, 125)$, $(503, 288)$ and $(722, 373)$ for BA=AB′, CA=AC′ and DA=AD′ respectively, where the symbol "′" means the negative direction of the inter-particle vector.

To get an interpolated function of the pair interaction potential as a function of the patch index i, j and the center-to-center distance R, $F(i, j; R)$, calculated data sets of the pair interaction potentials should be fitted as function forms. We use the non-linear least squares method (Kelley, 1999) to fit data sets of the pair interactions separated into the electrostatic interaction potentials and van der Waals interaction potentials. For the non-linear least square fitting, we use the trust region algorithm (Byrd et al., 1987) and the power law (Weisstein, 2010) to get a function form as $F = a \cdot R^b + c$. Table 4.1 shows the coefficients of these functions of the van der Waals interaction potentials between two lysozyme proteins after the non-linear least square fitting with trust region method.

To get pair interaction functions for each electrostatic and van der Waals interaction potential with all the orientations (i, j), the pair potentials $F(i, j; R)$ are interpolated from the six fitted functions. The non-linear squares fitting (Kelley, 1999) with trust

Table 4.1 Coefficients of the fitted function from the six computed pair interactions of the van der Wassls interaction potentials between two lysozyme proteins. The pairs from crystal operations are converted to the patch index (i, j). And a, b and c are the coefficients of $F = a \cdot R^b + c$ where R is the center-to-center distance between two proteins. R^2 values of fitting are also shown.

i	j	a	b	c	R^2
604	229	-4.79E+14	-9.002	-1.376	0.9732
218	401	-4.79E+14	-9.002	-1.376	0.9732
62	175	-4.88E+14	-8.831	-1.261	0.9821
76	125	-4.88E+14	-8.831	-1.261	0.9821
503	288	-4.62E+14	-9.020	-1.683	0.9888
722	373	-4.62E+14	-9.020	-1.683	0.9888

region method (Byrd et al., 1987) on 2D surface can be used to interpolate the coefficients a, b and c. But after fitting the predicted coefficients b can make the fitted functions worse to represent the pair interactions because the predicted coefficients of power b may have wide ranges of values whereas the coefficients b from the six computed pairs have relatively constant values. So we can assume that the power coefficient of the fitted function, b, is remained as an averaged constant. After using the constant power $b = -8.951$, we have the following fitted coefficients from six computed pairs of the van der Waasl interaction potential(Table 4.2). Considering small changes of the fit correlation coefficients, R^2, our assumption of the constant power coefficient b is valid.

After applying the 2D surface fitting with the non-linear squares fitting and trust region method, we have the following form of the coefficients a and c for the pair interaction potentials $F(i, j : R) = a \cdot R^{-8.951} + c$ for the van der Waals interactions between two lysozyme proteins,

$$f(i, j) = p_{00} + p_{10} \times i + p_{01} \times j + p_{20} \times i^2 + p_{11} \times i \times j + p_{02} \times j^2 \qquad (4.6)$$

where $p_{kl}(k, l = 1, 2)$ are the fitted parameters on the quadratic polynomial fitting. This 2D surface fitting can be done to all the coefficients a and c for both electrostatic and van der Waals interaction potentials on the patch model. The fitted surfaces are drawn in

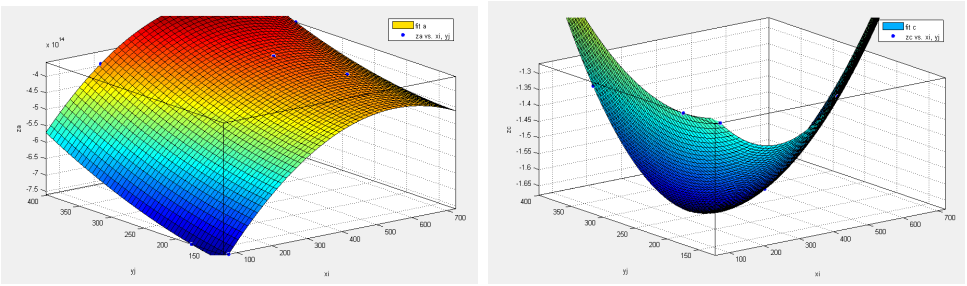
Table 4.2 Coefficients of the fitted functions from six computed pair interactions of the van der Wassls interaction potentials with the constant power coefficient b. The pairs from crystal operations are converted to the patch index (i, j). And a, $b = -8.951$ and c are the coefficients of $F = a \cdot R^b + c$ where R is the center-to-center distance between two proteins. R^2 values of fitting are also shown.

i	j	a	c	R^2
604	229	-3.97E+14	-1.373	0.9730
218	401	-3.97E+14	-1.373	0.9730
62	175	-7.61E+14	-1.270	0.9830
76	125	-7.61E+14	-1.270	0.9830
503	288	-3.59E+14	-1.679	0.9886
722	373	-3.59E+14	-1.679	0.9886

Figures 4.3(a) and 4.3(b) for coefficients a and c respectively. The obtained parameters a and c in the function for both electrostatic and van der Waals interaction potentials between two lysozyme molecules are listed in Table 4.3.

4.4 Calculation of the phase diagram of a protein

Once the pair potential of the anisotropic interaction between two protein molecules are computed, the total potential in Eq. (4.2) can be obtained. And this can be used in Monte Carlo(MC) simulations to calculate the equation of state for the solid and the liquid phases. And the thermodynamic integration is followed to obtain free energy differences between two phases (Frenkel and Smit, 2002). The coexistence line can be obtained using Gibbs-Duhem integration method introduced by Kofke (Kofke, 1993a,b). This coexistence line is obtained by integrating the Clausius-Clapeyron equation. As a first step of this procedure, we developed a method to calculate the anisotropic pair interaction potentials between two protein molecules of both the electrostatic contributions and the van der Waals contributions using a small number of different orientation pairs. In our future work we will apply our results to the calculation of the lysozyme



(a) Coefficient a (b) Coefficient c

Figure 4.3 2D surface non-linear least square fitting for the coefficients a and c of val der Waals pair interaction potentials between two lysozyme protein molecules. Axis x and y stand for the patch index i and j.

Table 4.3 Fitted parameters after applying the 2D non-linear square fitting on the coefficients of fitted functions from six computed pair interactions of electrostatic and van der Waals interaction potentials between two lysozyme proteins with the constant power coefficient b. For electrostatic potentials, the constant $b = -4.105$ is used to fit the other coefficients of $F = a \cdot R^b + c$ where R is the center-to-center distance between two proteins. The term "elec" means electrostatic interaction and "vdw", van der Waals interaction.

parameter	a(elec)	c(elec)	a(vdw)	c(vdw)
p_{00}	3.66E+08	-6.681	-9.00E+14	-1.003
p_{10}	-7.30E+05	0.02366	1.73E+12	-0.0007
p_{01}	-1.37E+06	0.01843	-2.43E+10	-0.00218
p_{20}	534.1	-4.63E-05	-1.52E+09	3.85E-06
p_{11}	561.2	4.83E-05	-4.44E+08	-9.15E-06
p_{02}	3102	-7.03E-05	1.54E+09	7.92E-06

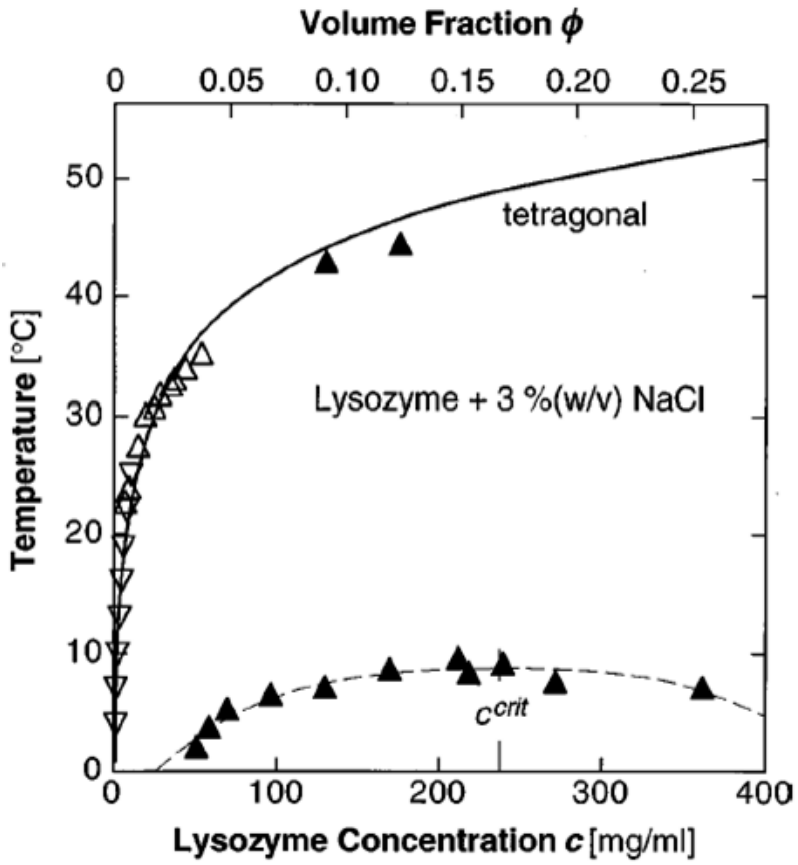


Figure 4.4 Phase diagram for lysozyme protein with 3% NaCl at pH= 4.5 in 0.1M NaAc buffer (Muschol and Rosenberger, 1997).

phase diagrams which are known experimentally(Figure 4.4) (Muschol and Rosenberger, 1997).

CHAPTER 5. Implementation of the Fast Multipole Method

5.1 Introduction

The electrostatic and van der Waals interactions between protein molecules in chapters 2, 3 and 4 are estimated from the linearized Poisson-Boltzmann (PB) equation where the realistic shape of protein molecules are considered. The most popular method for solving this PB equation is the finite difference method (FDM). But it is obviously difficult since a huge grid space is needed for FDM to get a reasonable accuracy in the dielectric media with proteins as it used to discretize a three dimensional volume. In contrast to the FDM method, boundary element method (BEM) discretizes only the boundary of the protein surface which is two-dimensional. Because of this reduction of dimensionality, BEM can be expected to be advantageous in large-scale problems. The application of this method has been limited to the relatively small proteins, because building up the coefficient matrix in BEM takes the full matrix space (usually not sparse) which is $O(N^2)$, where N is the number of unknowns which is the number of surface elements in BEM. The operation number to solve the linear system requires even more operations up to the order of $O(N^3)$ using the conventional direct solver such as the Gaussian elimination. In the study of protein molecules, to get a reasonable accuracy of the solution of PB equation comparing with the analytic solution, such as Eq. (A.13), N should be more than tens of thousands for the average size of proteins and even the small protein such as BPTI which has only 58 amino acid residues requires several thousands of discretized surface elements and the memory cost already

exceeds the order of Giga-Bites (GB) to more than ten GBs. However, the appearance of fast multipole method (FMM) dramatically changed the limitations for BEM. The use of FMM together with iterative solvers such as conjugate gradient (CG) method and generalized minimal residual (GMRES) method has been shown to reduce the memory requirement to $O(N)$ and the operation count to $O(N)$. Thus, FMM finally enables us to apply BEM to the large-scale problems such as protein-protein interactions. FMM in BEM was first introduced by Rokhlin (Rokhlin, 1985) for integral equations of two-dimensional Laplace equations and then developed by Greengard (Greengard, 1988) for pairwise force calculation in many body problems with Coulombic potential. It is further developed to achieve the order of $O(N)$ (Greengard and Rokhlin, 1997). In this chapter, we will show how this FMM algorithm can be applied to the solution of PB equation in protein-protein interaction energy calculations.

5.2 Formulation of the Fast Multipole Method

The fast multipole method (FMM) can speed up the matrix-vector multiplication on the following particular type of problem,

$$s(x_j) = \sum_{i=1}^{N} \alpha_i \phi(x_j - x_i), \ \{s_j\} = [\Phi_{ji}] \{a_i\} . \tag{5.1}$$

The solution on target x_j is evaluated by summing up the matrix (Φ_{ji} whose elements are defined by the potential function $\phi(x_j - x_i)$ between source x_i and its target x_j) and vector (a_i) products. Above sum of matrix-vector products requires $O(MN)$ operations where the total number of the surface index j is M. Applying the FMM in Eq. (5.1) can reduce this evaluation to $O(M + N)$ operations. The basic idea how this algorithm reduces the computational cost is described here.

The matrix element in the form of Eq. (5.1) is the evaluation of a function between two nodes, so called source and target points which typically represent the discretized

surface elements of a protein in our problem. The simplest example of the functions to satisfy this condition is the solution of the Laplace equation, its Green's function. The Green's function on the Cartesian coordinates can be expanded to the form of Eq. (5.1) using the Spherical Harmonics (Abramowitz and Stegun, 1964). By using the equality of Green's function, following identity can be obtained by shifting the geometric center from O to O' (see Figure 5.1),

$$\frac{1}{|x-y|} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} R_{n,m}(\overrightarrow{Oy})\overline{S_{n,m}}(\overrightarrow{Ox}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} R_{n,m}(\overrightarrow{O'y})\overline{S_{n,m}}(\overrightarrow{O'x}), \tag{5.2}$$

where $R_{n,m}$ and $S_{n,m}$ are the solid harmonics defined as,

$$\begin{aligned} R_{n,m}(\overrightarrow{Oy}) &= \frac{1}{(n+m)!}P_n^m(\cos\alpha)e^{im\beta}\rho^n, \\ \overline{S_{n,m}}(\overrightarrow{Ox}) &= (n-m)!P_n^m(\cos\theta)e^{-im\phi}\frac{1}{r^{n+1}}. \end{aligned} \tag{5.3}$$

If the shifted origin O' goes closer to the source point y, the number of summation in Eq. (5.2) to make this solution converged is reduced from more than 100s to less than 10s with the designated accuracy with about 4 to 6 significant figures (Rokhlin, 1985). Reducing the number of summation terms is the starting point of speed up during matrix-vector product operations in FMM.

This identity also can be applied to the gradient form of the function like,

$$\nabla_{n_y}\frac{1}{|x-y|} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \nabla_{n_y}R_{n,m}(\overrightarrow{Oy})\overline{S_{n,m}}(\overrightarrow{Ox}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \nabla_{n_y}R_{n,m}(\overrightarrow{O'y})\overline{S_{n,m}}(\overrightarrow{O'x}), \tag{5.4}$$

where $R_{n,m}$ and $S_{n,m}$ are the solid harmonics defined in Eq. (5.3) and the normal gradient on the target point x also has the same identity. This additional identity shows that the FMM can be applied to many aspects of physical problems involving simple solution like Eq. (5.2) but also like Eq. (5.4). If any solution of the problem can be expanded to the sum of series expansions and can have the identity between origin shifts, the FMM can be applied and speeds up the matrix-vector product operations.

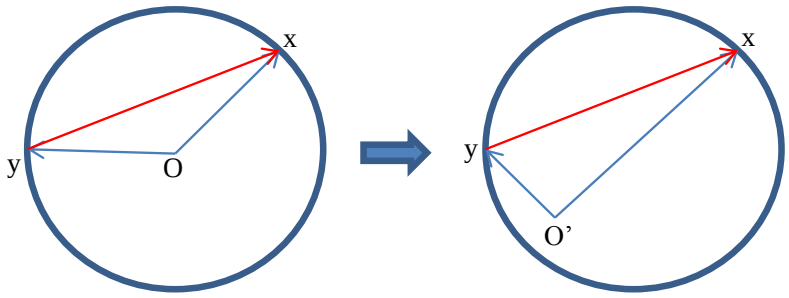


Figure 5.1 Schematic illustration showing the starting idea of the fast multipole method (FMM). Evaluation of the given function is described the red arrow between source point y and target point x. The translation of geometric center from O to O' is defined.

Above example is the minimization of computational cost only between two evaluation points. To keep minimum number of summation terms, minimization of the distance between source point and shifted origin is required. But one minimization of distance for one source point can be maximization for another point on the protein surface. So we need to introduce multiple origins and consider only the interactions between source points and a single origin point within the designated distance. To reduce computational cost and to maintain accuracy, multiple layers of expansion origins are required. And each origin also carries the information of the sum of coefficients in expansions from its source points, we can call them "children", and transfer this information to the nearest origin, "parent" (Figure 5.2). Detail formulations for series expansions of the functions and translations of coefficients will be described in next section about the application of FMM to the PB equation.

There is a practical problem how we can define multiple layers of expansion origins and the correlations between children expansion centers and parent centers and assign them to the nearest surface elements either as source or target points. In two-dimensional space, the surface domain is positioned inside the biggest rectangular box (called $level = 0$ box) and this rectangular box is divided to 4 smaller boxes with the same size ($level =$

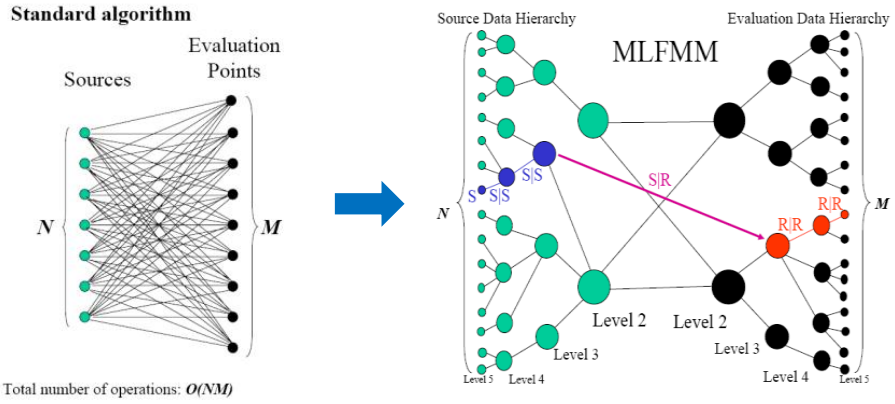


Figure 5.2 Schematic illustration showing the conversion of the general BEM to the multi-level FMM (MLFMM). The order $O(N+M)$ method in BEM is re-designed to the multi-level FMM by introducing the multiple layers of expansion origins. The translations of expansion coefficients from source points are indicated by $S|S$. $S|R$ translations is the evaluation of the function from source to target points to evaluate the matrix-vector product. The translations of expansion coefficients between target points are also indicated by $R|R$. This figure was taken and re-assembled from Gumerov and Duraiswami (Gumerov and Duraiswami, 2005).

1) and this procedure is repeated to small enough box size so that individual discretized surface elements are inside the box. The number of levels which decides the sizes of the boxes at each level depends on the balance between the expected accuracy and computational cost because the smaller size of the boxes at the finest level and the larger number of expansion coefficients (the number of summation terms) can guarantee the accuracy but with more memory and operation costs. The relations between rectangular boxes among the levels are described by the tree structure and in 2D we can call this tree as a "quad-tree" because each box always has four children. In three dimensional space, each cubic cell has eight children and therefore the "oct-tree" structure should be built to store all the information about the cell-to-cell relations and the translations of expansion coefficients. Figure 5.3 shows how the indices of the surface elements can be stored in the tree structure in two-dimension.

5.3 Application of the Fast Multipole Method

Eqs. (2.4), (2.5) (the electrostatic interaction) and Eqs. (2.21), (2.22) (the van der Waals interaction) in chapter 2 and Eqs. (3.6), (3.7), (3.8), (3.9) (the electrostatic interaction) and Eqs. (3.28), (3.29), (3.30), (3.31) (the van der Waals interaction) in chapter 3 can be described as systems of linear equations. For example, Eqs. (2.4) and (2.5) in chapter 2 are

$$\begin{aligned}
&\frac{1}{2}\left(1+\frac{\varepsilon_2}{\varepsilon_1}\right)\varphi(\mathbf{r}_0)+\iint_{\sum} L_1(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r})\mathrm{d}\mathbf{r}+\iint_{\sum} L_2(\mathbf{r},\mathbf{r}_0)\frac{\partial\varphi(\mathbf{r})}{\partial n}d\mathbf{r}\\
&=\sum_{i=1}^{N}\left\{q_i F(\mathbf{r_i},\mathbf{r}_0)+\vec{\mu_i}\cdot\nabla F(\mathbf{r_i},\mathbf{r}_0)\right\}/\varepsilon_1, \qquad (5.5)
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{2}\left(1+\frac{\varepsilon_1}{\varepsilon_2}\right)\frac{\partial\varphi(\mathbf{r}_0)}{\partial n}+\iint_{\sum} L_3(\mathbf{r},\mathbf{r}_0)\varphi(\mathbf{r})\mathrm{d}\mathbf{r}+\iint_{\sum} L_4(\mathbf{r},\mathbf{r}_0)\frac{\partial\varphi(\mathbf{r})}{\partial n}d\mathbf{r}\\
&=\sum_{i=1}^{N}\left\{q_i\frac{\partial F}{\partial n_0}(\mathbf{r_i},\mathbf{r}_0)+\vec{\mu_i}\cdot\nabla\frac{\partial F}{\partial n_0}(\mathbf{r_i},\mathbf{r}_0)\right\}/\varepsilon_1, \qquad (5.6)
\end{aligned}$$

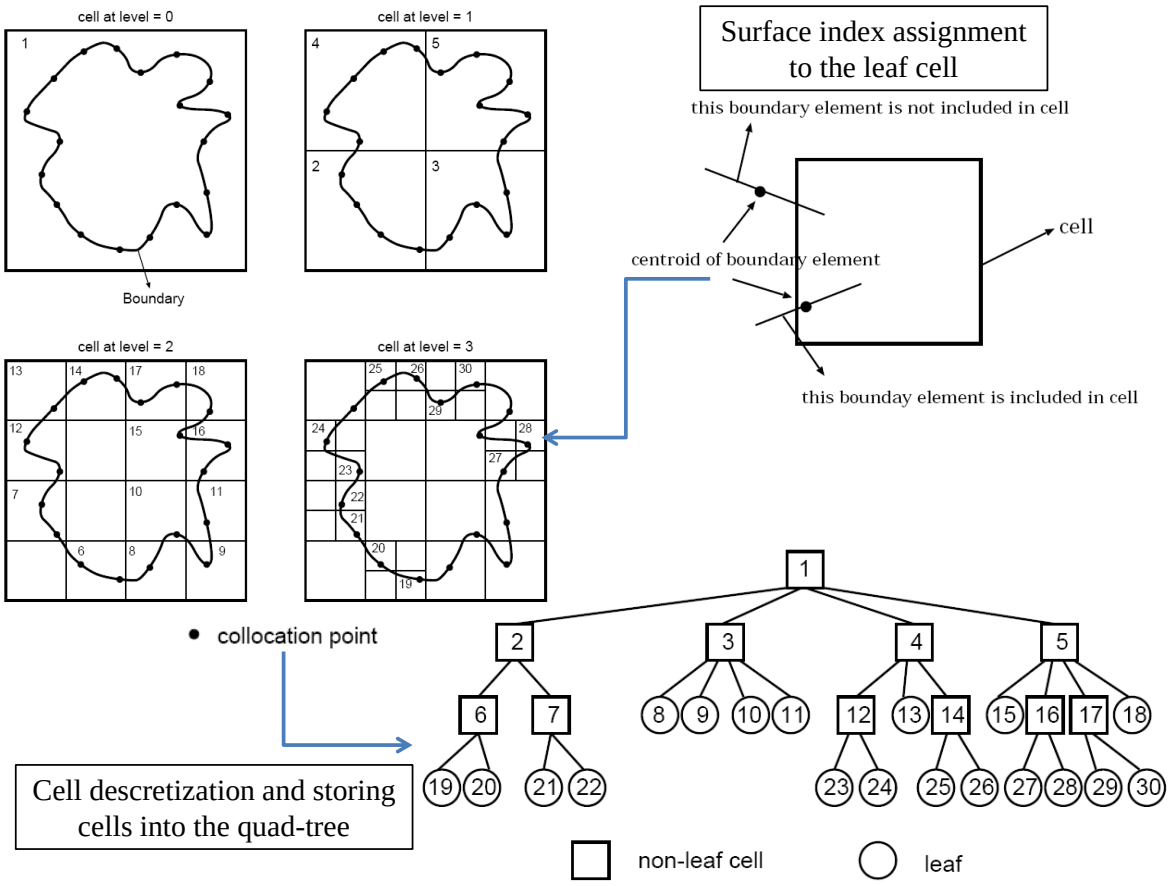


Figure 5.3 Schematic illustration shows how the boundary surface elements are assigned to the nearest leaf cell (leaf cell is the finest cell and has no children) and how whole domain is discretized to the multi-level rectangular cells. The indices of surface elements are stored into the leaf cells and relations between cells in different levels such as parent-children relations are stored into the quad-tree in two-dimensional space. This figure was taken and re-assembled from Yoshida (Yoshida, 2001).

where L_1, L_2, L_3 and L_4 are defined in Eqs. (2.6), (2.7), (2.8) and (2.9). All Ls are functions between source $\mathbf{r}_0$ and target $\mathbf{r}$ located on the surface whose number is N. Thus, Ls are $N \times N$ matrices. Discretizing the functions $\varphi(\mathbf{r}_0)$, $\frac{\partial\varphi(\mathbf{r}_0)}{\partial n}$ and the right hand side of the above equations yields the following linear system with simplifying notations $\varphi(\mathbf{r}_0)$ to φ_0 and $\frac{\partial\varphi(\mathbf{r}_0)}{\partial n}$ to φ_1 and the right-hand side F and its normal gradient to F_0 and F_1,

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \varphi_0 \\ \varphi_1 \end{pmatrix} - \begin{pmatrix} L_1 & L_2 \\ L_3 & L_4 \end{pmatrix} \begin{pmatrix} \varphi_0 \\ \varphi_1 \end{pmatrix} = \begin{pmatrix} F_0 \\ F_1 \end{pmatrix}, \tag{5.7}$$

where the size of vectors is N and the size of matrices is $N \times N$.

$$(I - L)A = B, \tag{5.8}$$

where I is the identity matrix with the size of N^2 and A, B are single column vectors with the size N. This form also can be applied to protein interaction calculations with the size $2N$ in chapter 3, the number of surface elements to discretize the protein surface of amino acid residues. A and B also can be $N \times N$ ($2N \times 2N$ in chapter 3) matrices for the reaction field of the van der Waals energy calculation.

All matrix elements consist of the sum of the solutions of Coulombic interactions and Screened Coulombic interactions. For example, the matrix element, L_2, are given by,

$$\int_{S_y} \left(\frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_y} - \frac{\varepsilon_2}{\varepsilon_1} \frac{\partial P(\mathbf{x}-\mathbf{y})}{\partial n_y} \right) \phi(\mathbf{y}) \mathrm{d}S_y, \tag{5.9}$$

where F and P are defined in Eqs. (2.10) and (2.11) in chapter 2.

Thus, we need to solve the following matrix $\left(\int_{S_y} \frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_y} dS_y \right)$ and vector $(\phi(\mathbf{y}))$ for example. The key point for solving this linear system with the efficiency is to accelerate the matrix-vector multiplications during iterations in the iterative linear equation solver, such as GMRES. This can be done by introducing the fast multipole method, FMM. Implementation of FMM is described by followings.

The multipole moment expansions for the Coulombic interaction ($\mathbf{y}$ to $\mathbf{O}$ in Figure 5.4) are given by (Yoshida, 2001),

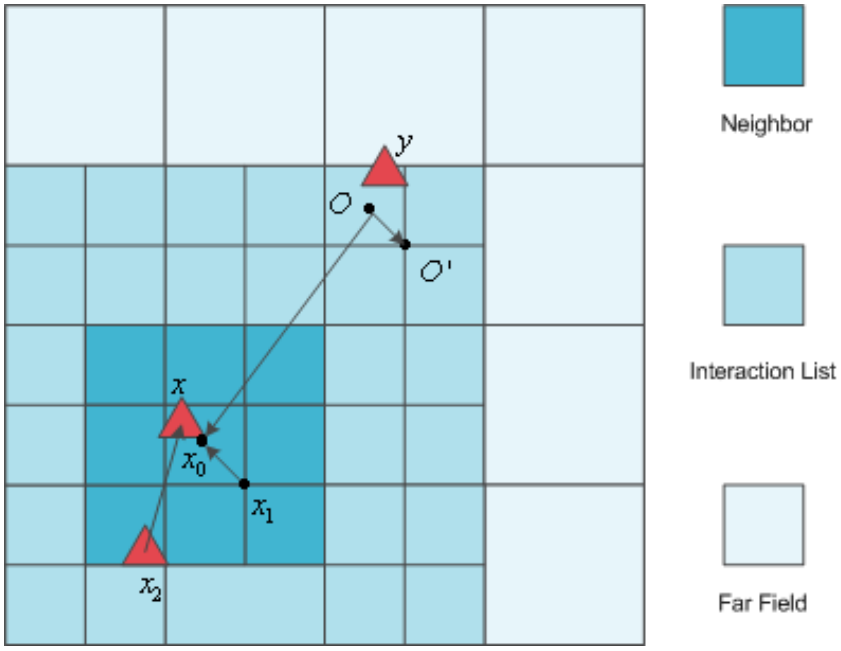


Figure 5.4 Schematic illustration showing the hierarchical rectangular boxes of the fast multipole method in two dimensional space for convenience. The largest box represents the highest level, level-zero, and the smallest boxes are in the finest level, level-three in this picture. The lightly shaded level-two boxes are in the far field list from the target point $\mathbf{x}$. The light blue boxes are in the interaction list (up to 189 boxes in three dimension) which translates the multipole expansion to the local expansion (M2L translation, $\mathbf{x}_0$ to $\mathbf{x}_0$ arrow). The arrows O to O' and $\mathbf{x}_1$ to $\mathbf{x}_0$ indicate multipole to multipole (M2M) and local to local translation (L2L) respectively. Finally the dark blue boxes (up to 27 boxes in three dimension) are the neighbor boxes. The interaction between neighbors including the self interactions can be calculated by the direct BEM solver (Lu et al., 2007).

$$\int_{S_y} \frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_y}\phi(\mathbf{y})dS_y = \frac{1}{4\pi}\sum_{n=0}^{p}\sum_{m=-n}^{n} \overline{S_{n,m}}(\overrightarrow{O\mathbf{x}})M_{n,m}(O), \tag{5.10}$$

where the Green's function $F(\mathbf{x}-\mathbf{y}) = \frac{1}{|\mathbf{x}-\mathbf{y}|}$ and the multipole moment coefficients are

$$M_{n,m}(O) = \int_{S_y} \frac{\partial R_{n,m}(\overrightarrow{O\mathbf{y}})}{\partial n_y}\phi(\mathbf{y})dS_y \tag{5.11}$$

and $R_{n,m}$ and $S_{n,m}$ are the solid harmonics defined as:,

$$\overline{S_{n,m}}(\overrightarrow{O\mathbf{x}}) = (n-m)!P_n^m(\cos\theta)e^{-im\phi}\frac{1}{r^{n+1}}, \tag{5.12}$$

$$R_{n,m}(\overrightarrow{O\mathbf{y}}) = \frac{1}{(n+m)!}P_n^m(\cos\alpha)e^{im\beta}\rho^n. \tag{5.13}$$

The multipole moment expansions for the screened Coulombic interaction can be written as,

$$\begin{aligned}\int_{S_y} \frac{e^{-\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|}\phi(\mathbf{y})dS_y &= \frac{2\kappa}{\pi}\sum_{n=0}^{p}(2n+1)k_n(\kappa r) \\ &\times \sum_{m=-n}^{n}\overline{S_{n,m}}(\theta,\phi)M_{n,m}(\kappa,O),\end{aligned} \tag{5.14}$$

where the multipole coefficients,

$$M_{n,m}(\kappa,O) = \int_{S_y} i_n(\kappa\rho)R_{n,m}(\alpha,\beta)\phi(y)dS_y \tag{5.15}$$

and $i_n(\kappa\rho)$ and $k_n(\kappa r)$ are modified spherical Bessel and modified spherical Hankel functions are defined in terms of Bessel function (Abramowitz and Stegun, 1964).

$$I_\nu(r) = i^{-\nu}J_\nu(ir), \tag{5.16}$$

$$K_\nu(r) = \frac{\pi}{2\sin\nu\pi}\left[I_{-\nu}(r) - I_\nu(r)\right], \tag{5.17}$$

$$i_n(r) = \sqrt{\frac{\pi}{2r}}I_{n+1/2}(r), \tag{5.18}$$

$$k_n(r) = \sqrt{\frac{\pi}{2r}}K_{n+1/2}(r). \tag{5.19}$$

and $R_{n,m}(\alpha,\beta)$ and $\overline{S_{n,m}}(\theta,\phi)$ are the spherical harmonics are defined as,

$$S_{n,m}(\theta,\phi) = R_{n,m}(\theta,\phi) = \sqrt{\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi}, \tag{5.20}$$

where the upper bar represents the complex conjugate of the harmonics. The integrals in Eq. (5.10) and Eq. (5.14) can be evaluated with the local expansion coefficients as follows,

$$\int_{S_y} \frac{\partial F(\mathbf{x}-\mathbf{y})}{\partial n_y} \phi(\mathbf{y}) dS_y = \frac{1}{4\pi} \sum_{n=0}^{p} \sum_{m=-n}^{n} R_{n,m}(\overrightarrow{\mathbf{x}_0\mathbf{x}}) L_{n,m}(\mathbf{x}_0), \tag{5.21}$$

$$\int_{S_y} \frac{e^{-\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|} \phi(\mathbf{y}) dS_y = \frac{2\kappa}{\pi} \sum_{n=0}^{p} (2n+1) i_n(\kappa r) \sum_{m=-n}^{n} \overline{S_{n,m}}(\theta,\phi) L_{n,m}(\kappa, \mathbf{x}_0). \tag{5.22}$$

The expression of the local expansion coefficients ($\mathbf{x}_0$ to $\mathbf{x}$ in Figure 5.4) for the Coulombic interaction can be written as following (Yoshida, 2001),

$$L_{n,m}(\mathbf{x}_0) = \sum_{n'=0}^{p} \sum_{m'=-n'}^{n'} (-1)^{n'} \overline{S_{n+n',m+m'}}(\overrightarrow{O\mathbf{x}_0}) \times M_{n',m'}(O). \tag{5.23}$$

The above procedure is called, "the multiple to local translation (simply M2L translation)" ($\mathbf{O}$ to $\mathbf{x}_0$ in Figure 5.4). The equation for the M2L translation of screened Coulombic interaction can be derived by using the properties (Yoshida, 2001) of the translational equalities in the spherical Bessel and Hankel functions (Epton and Dembart, 1995) and applying them to the modified spherical Bessel and Hankel functions. The final expression of the M2L translation is given by,

$$\begin{aligned} L_n^m(\kappa, \mathbf{x}_0) = & \sum_{n'=0}^{p} \sum_{m'=-n'}^{n'} \sum_{\substack{l=|n-n'| \\ n'+n-l:even}}^{n+n'} (2n'+1) \\ & \times W_{n',n,m',m,l} k_l(\kappa, \overrightarrow{O\mathbf{x}_0}) \times \overline{S_{l,-m-m'}}(\overrightarrow{O\mathbf{x}_0}) M_{n'}, m'(\kappa, O), \end{aligned} \tag{5.24}$$

where $W_{n',n,m',m,l}$ is written by the following equation with the Wigner-3j symbol (Messiah, 1962),

$$W_{n,n',m,m',l} = (2l+1) i^{n'-n+l} \times \begin{pmatrix} n & n' & l \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} n & n' & l \\ m & m' & -m-m' \end{pmatrix}. \tag{5.25}$$

The oct-tree structure source code developed by Song (Chew et al., 2001) is used to define the 'interaction list', which has a key role to connect multipole expansion coefficients to local expansion coefficients, the M2L translation. At the finest level, the interaction between elements in the nearest neighbor, called the near field interaction, can be calculated by the direct boundary element solver with the collocation method from Atkinson and coworkers (Atkinson and Han, 2009) and the interactions from the far field elements, the multipole moment expansion coefficients are translated to the higher level expansions, called "the multipole to multipole translation (M2M)" (**O** to **O**′ in Figure 5.4). Once M2L translations are computed in the higher level of tree structure, they should be translated to local expansions in the lower level, finally to local expansions in the finest level in order to evaluate the integrals and the matrix-vector multiplications. This process is called "local to local translation (L2L)" ($\mathbf{x}_1$ to $\mathbf{x}_0$ in Figure 5.4). The equations for M2M and L2L translations for Coulombic and screened Coulombic interactions are given below,

$$M_{n,m}(O') = \sum_{n'=0}^{n} \sum_{m'=-n'}^{n'} R_{n',m'}(\overrightarrow{O'O}) \times M_{n-n',m-m'}(O), \tag{5.26}$$

$$M_n^m(\kappa, O') = \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} \sum_{\substack{l=|n-n'| \\ n'+n-l:even}}^{n+n'} (2n'+1) \times (-1)^{m'} W_{n,n',m,m',l} i_l \left(\kappa, \overrightarrow{O'O}\right) \times S_{l,-m-m'} \left(\kappa, \overrightarrow{O'O}\right) M_{n',-m'}(\kappa, O), \tag{5.27}$$

$$L_{n,m}(\mathbf{x}_0) = \sum_{n'=n}^{\infty} \sum_{m=-n'}^{n'} R_{n'-n,m'-m} \left(\overrightarrow{\mathbf{x}_1\mathbf{x}_0}\right) \times L_{n',m'}(\mathbf{x}_1), \tag{5.28}$$

$$L_n^m(\kappa, \mathbf{x}_0) = \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} \sum_{\substack{l=|n-n'| \\ n'+n-l:even}}^{n+n'} (2n'+1) \times (-1)^{m} W_{n',n,m',-m,l} i_l \left(\kappa, \overrightarrow{\mathbf{x}_1\mathbf{x}_0}\right) \times S_{l,m-m'} \left(\overrightarrow{\mathbf{x}_1\mathbf{x}_0}\right) L_{n',m'}(\kappa, \mathbf{x}_1). \tag{5.29}$$

The restarted generalized minimal residual method (Barrett et al., 1994) is used to solve the linear equations and we modified the code computing matrix-vector products to make an interface to FMM. The comparison between the direct BEM solver and our FMM solver in computational cost is shown on Figure 5.5 and Figure 3.6 for single protein solver and double protein solver respectively. According to these figures our FMM code has a linear dependence on the number of elements N with the maximum number of levels $\log N$, finally it follows $O(N \ln N)$ algorithm.

5.4 Algorithm and estimation of computational cost

We describe the general algorithm of the fast multipole method in two-dimensional space and predict the computational cost based on this system for the simplicity of understanding.

Step 1 Discretization:

Discretize the given surface domain to the boundary elements, S_y, as in the traditional BEM solver. To compare the computational cost between traditional BEM and FMM-BEM, we use the number of elements as $N = 1000$, for example.

Step 2 Construct the tree structure (quad-tree in 2D, oct-tree in 3D):

Define a rectangular box which holds the surface elements in it and this is the level 0 cell. And divide this cell into 4 equal rectangular boxes and call them the cells of level 1 and children of level 0 cell. Repeat this to the maximum level which is chosen to get the designated accuracy and cost. We need to consider the following parameters. The number of unknowns is N, the maximum number of surface elements in a leaf cell is M and the maximum number of terms in the summation of expansions is p. So the total number of leaves is N/M and the level of this quad-tree is $\log_4(N/M)$. See Figure 5.3

Step 3 The Multipole moment expansions:

Compute the multipole moment expansions following Eqs. (5.10) and (5.14) for Coulombic interaction and screened Coulombic interaction in the lowest level of cells (leaves). This procedure takes N/M(the number of leaves)
$\times M$(the number of elements in one leaf) $\times p$(the summaation terms) $= O(pN) =$ 5000, for example.

Step 4 The Multipole to Multipole (M2M) translations:

From the cells of the lowest level, translate the multipole moment expansion coefficients to the center of the parent's cell by using Eq. (5.26) and Eq. (5.27) for Coulombic interaction and screened Coulombic interaction respectively. Repeat this to the $level = 3$. This procedure takes p^2(the operation cost)
$\times 4N/3M$(the number of cells of translations in all levels)
$= O(4p^2N/3M) = O(3333)$ from the given numbers of parameters.

Step 5 The Multipole to Local (M2L) translations:

If the two cells in the same level are in the "interaction list" according to the tree structure (see Figure 5.4), the multipole moment coefficients can be translated to the target cell to evaluate the given integral to compute matrix-vector products by using Eq. (5.3) and Eq. (5.24) for Coulombic interaction and screened Coulombic interaction respectively. This procedure is repeated to the $level = 2$ to translate all the possible interactions from source elements to their targets. The cost of M2L translations is
p^2(the operation cost) $\times$ 27(the number of the cells in the interaction list)
$\times 4N/3M$(the number of cells in translations) $= O(36p^2N/M) = O(90000)$ from the given numbers of parameters.

Step 6 The Local to Local (L2L) translations:

If the M2L translations are computed on the non-leaf cells, the translated coeffi-

cients are translated to their children. From the cells in the highest level (starting from $level = 2$), translate the local expansion coefficients to the center of children cells by using Eq. (5.28) and Eq. (5.29) for Coulombic interaction and screened Coulombic interaction respectively. Repeat this to the parent cells of cells in the lowest level. This procedure takes p^2(the operation cost)
$\times\, 4N/3M$(the number of cells of translations in all levels)
$= O(4p^2N/3M) = O(3333)$ from the given numbers of parameters.

Step 7 The local expansions (integral evaluations):

Compute the local expansions following Eq. (5.21) and Eq. (5.22) for Coulombic interaction and screened Coulombic interaction respectively from the cells in the lowest level (leaves) to evaluate the integrals to compute matrix-vector products. This procedure takes
N/M(the number of leaves) $\times\, M$(the number of element in one leaf)
$\times\, p$(the summation terms) $= O(pN) = 5000$, for example.

Step 8 The direct computations (integral evaluations):

If the source and target elements are assigned to adjacent cells at all leaf cells, the given integral should be computed by using the traditional direct solver of BEM. The cost to evaluate the integrals to compute matrix-vector products is
N/M(the number of leaves) $\times\, 9$(the numeber of adjacent cells)
$\times M^2$(the cost of direct computation) $= O(9MN) = 90000$ with using the given parameters.

So the total cost from the FMM on the two-dimensional boundary element problem is

$$2O(pN) + 2O(p^2 4N/3M) + O(27p^2 4N/3M) + O(9MN) \sim O(N), \tag{5.30}$$

with the given parameters as an example,

$$2 \times O(5000) + 2 \times O(3333) + O(90000) + O(90000) \sim O(2 \times 10^5). \tag{5.31}$$

This cost is approximately only 20% of the cost from direct BEM ($O(N^2 = 10^6)$). If the number of boundary elements increases to $N = 10^4$, then the cost from FMM is only $O(2 \times 10^6)$ and it is only 2% of the cost from the direct BEM solver ($O(N^2 = 10^8)$). Considering the number of the iterations to solve the system of linear equations, the cost saving can be dramatic both on memory and time cost for computing matrix-vector products. This is the advantage of the fast multipole method. The above ratio can be ideal. To compare the cost with the realistic problem, we made the comparison with the single body protein calculation solver to calculate the binding affinities of mutant protein complexes. Figure 5.5 shows the comparison of memory demand between BEM direct solver and FMM-BEM solver for the calculations of electrostatic energy contribution to the binding affinity of BPTI-trypsin complex. With the implementation of FMM in our model, our electrostatic free energy solver only takes about 1GB of memory, significantly smaller than BEM direct solver which demands more than 7GB using about $15,000$ surface elements. The BEM solver follows $O(N^2)$ algorithm whereas the FMM solver only does $O(N \log N)$ algorithm. This trend also can be applied to expect the time consuming. Because of the reduced number of problems by the residual model and the reduced computational cost by the implementation of FMM algorithm, our model is able to calculate the binding free energy of the mutations on three protein-protein complexes efficiently. The comparison with the double body calculation solvers to compute effective interaction energy calculations of the two proteins in chapter 3 shows that the implementation of FMM to two body BEM reduces a lot more computational cost from $O((2N)^2)$ with direct BEM solver to $O(N)$ and $O(2N)$ with double-tree and single-tree FMM respectively (Figure 3.6).

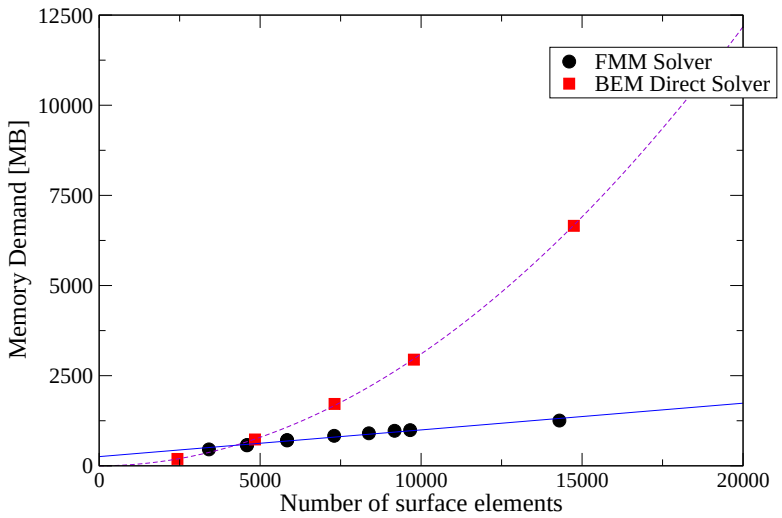


Figure 5.5 The comparison of the memory demand between BEM direct solver and FMM solver for calculations of the electrostatic energy contribution to the binding of BPTI-trypsin complex. The red square boxes represent memory demands from the direct BEM solver and the dashed line is its curve fitting whose power is 1.973. The black spheres indicate the FMM solver data and the blue solid line is its curve fitting with power 1.00 with maximum level of tree = 6.

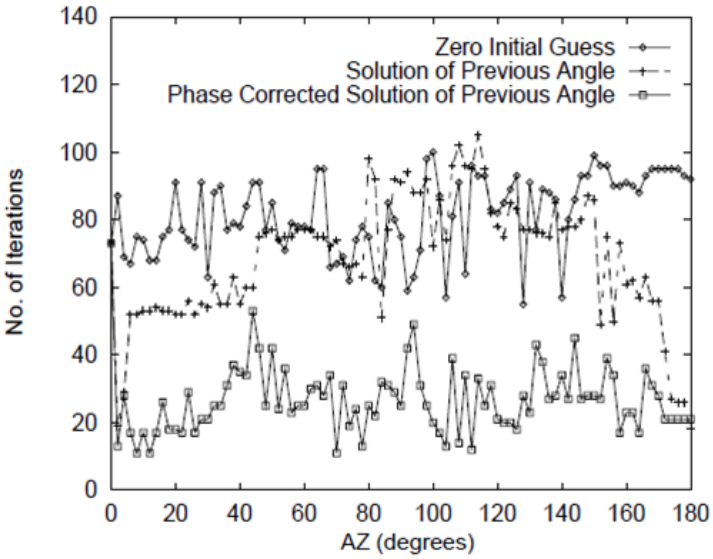


Figure 5.6 The changes of the number of iterations to solve the system of linear equations after applying the Initial Guess method. After applying the previous solution and angle correction the number of iterations to get a solution within the designated tolerance is significantly reduced (Song and Chew, 1998).

5.5 The Initial Guess improvement

For the iterative solution of the electrostatic interaction energy calculations, the number of systems of linear integral equations is only one. But for the iterative solutions of the van der Waals interaction energy calculations, the number of linear systems depends on the number of amino acid residues in proteins. Also in each residue we need to solve the linear system for each x, y, z coordinate to build up the reaction field matrix. For example, the number of linear systems for BPTI protein is $58(\text{Residues}) \times 3(x, y, z) = 174$, that means we need to consider more improvement of our iterative solver with FMM algorithm because of the large number of linear system problems even though this FMM reduces the computational cost and time consuming.

According to Song and Chew, for their iterative solutions, a small change in the incident angle corresponds to a small change in the current solution (Song and Chew,

1998). Using the current solution from the previous angle as an initial guess for the next solution from the next angle can improve the performance of the iterative solution by reducing the number of iterations. Two steps of initial guess were applied in their study. Firstly, they use the previous solution as an initial guess for the next solution.

$$\mathrm{J}_2(\mathbf{r}) = \mathrm{J}_1(\mathbf{r}) \tag{5.32}$$

And the phase correction based on the direction of incident angles from the previous angle k_1 to the next angle k_2 is applied with the general observation that the phase corrected solution, $\widetilde{\mathrm{J}}(\mathbf{r})$, changes more slowly than $\mathrm{J}(\mathbf{r})$ when k_i changes.

$$\widetilde{\mathrm{J}}(\mathbf{r}) = \mathrm{J}(\mathbf{r})e^{-ik_i\cdot\mathbf{r}} \tag{5.33}$$

So $\mathrm{J}_1(\mathbf{r})e^{-ik_1\cdot\mathbf{r}}e^{-ik_2\cdot\mathbf{r}}$ can be an initial guess for the next solution $\mathrm{J}_2(\mathbf{r})$. This technique significantly reduces the number of iterations. In their study, applying the first initial guess from the previous solution only shows the minor improvement of the number of iterations, but the second method with the phase correction significantly reduces the number of iterations (see Figure 5.6).

This initial guess method is also applied to our iterative solver. The first approach in Eq. (5.32) is used to improve the number of iterations in the electrostatic interaction calculations. Even though this linear system yields only one solution, the initial guess based on the right-hand-side (RHS) vector reduces the number of iteration from 52 to 39 in BPTI protein calculations, for example.

For the iterative solutions of van der Waals interaction energy calculations, the second method, the phase correction, is used to improve the number of iterations in each linear system in addition to the RHS vector initial guess method as following,

$$\mathrm{J}_2(\mathbf{r}_2) = \mathrm{J}_1(\mathbf{r}_1)\nabla_2\frac{1}{|\mathbf{r}_2 - \mathbf{r}_1|}. \tag{5.34}$$

And the number of iterations to calculate the van des Waals interactions between two

BPTI proteins is reduced from 52 to 39 after applying the RHS initial guess) and finally to 36 after applying Eq. (5.34).

CHAPTER 6. Final remarks

In this thesis, the author has investigated applications of the Fast Multipole Method (FMM) to the Boundary Element Method (BEM) for calculations of interaction energies between protein molecules in three dimension with both electrostatic energy contribution and van der Waals energy contribution. The results obtained in this thesis are proven to make the conclusion that the first step is gradually taken toward to more practical applications of protein-protein interactions.

The results obtained in this thesis are summarized as follows,

1. In chapter 2, implementation of the FMM to the BEM has been successfully applied to solve the linearized Poisson-Boltzmann equation that is the basis to calculate the electrostatic energy contribution and the van der Waals energy contribution to investigate binding affinity calculations of protein complexes. We built a residual model with the FMM-BEM of a single protein to describe a protein at a residue level. The procedure to generate suitable structures of protein complexes with single mutations on the binding site are indicated and validated with existing experimental PDB structures.

2. In chapter 3, the residual model with the FMM-BEM has been extended to compute effective interaction energies between two protein molecules with the single-tree and the double-tree FMM-BEM strategies. An anisotropic patch model based on a number of surface elements on two proteins was introduced to represent relative orientations between two protein molecules. With the successful applications

of the FMM-BEM to the effective interactions between two proteins and the reduced number of pair interactions from the anisotropic patch model, we are able to calculate the second virial coefficients of proteins in various solution conditions.

3. In chapter 4, the model based on the anisotropic patch model has been applied to calculations of pair interaction potentials between two protein as a first step to calculate a phase diagram of a protein. The pair potentials from many different orientations of two proteins are interpolated from pair potentials of six pairs chosen from surface patches.

4. In chapter 5, implementation of the FMM to the BEM to build solutions for PB based problems is described and proven to show how much computational cost it can save by introducing the FMM algorithm to the realistic problem based on the BEM to compute interaction energies between protein molecules.

The author has plans to do followings as the future work,

1. Apply the recent FMM algorithm to our FMM-BEM solvers to reduce further computational cost. There are still possibilities to save the cost especially the time consumption to solve the system of linear equations. We can save a lot more cost for the van der Waals interaction energy calculation because we need to solve a number of linear systems based on the number of residues in proteins with the residual model.

2. Design a model to reduce the number of pair interactions with the anisotropic patch model based on the surface elements. Reduced number of unknowns enables us to investigate a phase behavior of a protein and to guide the optimal crystallization condition to study the structure of a protein.

3. Apply the FMM-BEM to large-scale problems of interactions between large size proteins in which the number of surface elements easily exceed the order of $10^6 \sim$

10^8 using various new techniques of the FMM, libraries for the parallel implementation such as the Message Passing Interface(MPI) or the Parallel Virtual Machine(PVM).

By implementation of the FMM to the BEM we are able to build efficient models to compute interaction energies between particles in a protein level for both electrostatic interaction energy and van der Waals interaction energy. The author hopes that the FMM-BEM solver for protein-protein interactions become a practical solver to investigate many aspects of protein-protein interactions and finally to be useful in search of optimal solution conditions of the protein crystallization.

APPENDIX A. The analytic expression of the electrostatic interaction free energy

A.1 Electrostatic interaction free energy between two charged spherical particles

In order to validate our boundary element solvers either based on the direct solver or the fast multipole method, we derived the analytic solution for the basic model, two charged identical spheres in electrolyte solution. We follow the approach (Carnie and Chan, 1993) with linearized Poisson-Boltzmann model by adding a charge on the center of each sphere. In the linearized Poisson-Boltzmann model, the electrostatic potential ψ satisfies the following equations.

$$\begin{aligned} \nabla^2\psi = & \quad \kappa^2\psi & \text{outside the sphere} \\ = & \quad -\frac{q_i\delta(r-r_i)}{\varepsilon_1} & \text{inside the sphere,} \end{aligned} \tag{A.1}$$

where κ is the inverse Debye screening length of the electrolyte solution and q_i is the charge located in the center of each sphere i and ε_1 is the dielectric constant inside the sphere. And the solution of Eq. refpb1 in the electrolyte solution(outside of the spheres) (Glendinning and Russel, 1982) can be given as the following form and the coordinate system of the two sphere particles are shown on Figure A.1.

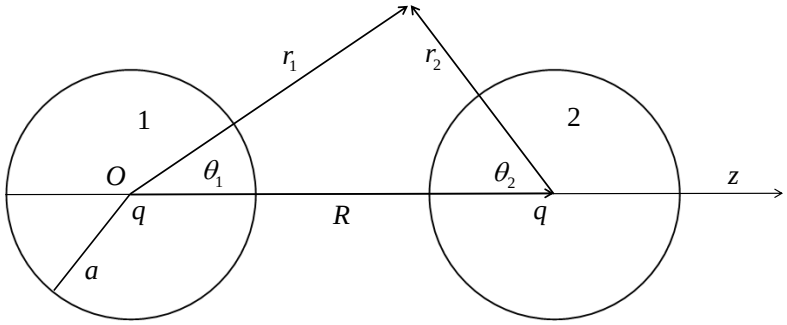


Figure A.1 Schematic diagram of the coordinate system of two sphere problem. a is the radius of sphere, R is the center-to-center distance and r_1, θ_1, r_1 and θ_1 are the coordinate system from sphere 1 and 2 respectively (Carnie and Chan, 1993). The charge q is located on the center of each sphere.

$$\begin{aligned}\psi &= \sum_{n=0}^{\infty} a_n\{k_n(\kappa r_1)P_n(\cos\theta_1) \\ &+ \sum_{m=0}^{\infty}(2m+1)B_{nm}i_m(\kappa r_1)P_n(\cos\theta_1)\},\end{aligned} \tag{A.2}$$

where

$$B_{nm} = \sum_{\nu=0}^{\infty} A_{nm}^{\nu} k_{n+m-2\nu}(\kappa R) \tag{A.3}$$

$$A_{nm}^{\nu} = \frac{\begin{array}{c}\Gamma(n-\nu+1/2)\Gamma(m-\nu+1/2)\Gamma(\nu+1/2) \\ \times(n+m-\nu)!(n+m-2\nu+1/2)\end{array}}{\pi\Gamma(m+n-\nu+3/2)(n-\nu)!(m-\nu)!\nu!}, \tag{A.4}$$

and $i_n(x)$, $k_n(x)$ are the modified spherical Bessel functions of the first and third kind respectively (Abramowitz and Stegun, 1964), R is the center-to-center distace between the two spheres and $\Gamma(z)$ is the gamma function. Also the solution of Eq. (A.1) inside the spheres has the general form,

$$\varphi = \sum_{n=0}^{\infty} b_n r^n P_n(\cos\theta) + \frac{q}{r}, \tag{A.5}$$

where $r = r_1$ or r_2 and $\theta = \theta_1$ or θ_2 by symmetry and $q = q_1$ or q_2 and $\varphi = \varphi_1$ or $\varphi = \varphi_2$ either inside of each sphere. The unknown coefficients a_n and b_n can be determined by applying the boundary conditions of the potential functions given by Eq. (A.2) and Eq. (A.5) on across the surface of sphere at $r_1 = a$,

$$\begin{aligned} \psi|_{r_1=a} &= \varphi|_{r_1=a} \\ \varepsilon_2 \frac{\partial \psi}{\partial r}\bigg|_{r_1=a} &= \varepsilon_1 \frac{\partial \varphi}{\partial r}\bigg|_{r_1=a}, \end{aligned} \tag{A.6}$$

where ε_1 is the dielectric constant inside the sphere and ε_2 is the dielectric constant of the solution, and $\varepsilon = \varepsilon_2/\varepsilon_1$ will be used for further derivation. By applying the boundary conditions Eq. (A.6) on Eq. (A.2) and Eq. (A.5) we evaluate the coefficient b_n and the potential function inside the sphere is,

$$\begin{aligned} \varphi &= \sum_{n=0}^{\infty} \Big[\Big(\frac{r}{a}\Big)^n a_n \{k_n(\kappa a) \\ &\quad + \sum_{m=0}^{\infty} (2m+1) B_{nm} i_m(\kappa a)\} P_n(cos\theta) - \frac{q}{a}\Big(\frac{r}{a}\Big)^n \Big] + \frac{q}{r}. \end{aligned} \tag{A.7}$$

In order to evaluate the electrostatic solvation energy on the charge position, $r \to 0$ limit makes only $n = 0$ term survived and the self-energy term also goes out.

$$\varphi(r \to 0) = a_0 \{k_0\{\kappa a) + \sum_{m=0}^{\infty} (2m+1) B_{0m} i_m(\kappa a)\} - \frac{q}{a}. \tag{A.8}$$

To find another unknown coefficient a_0, we only need the $m = 0$ term after applying the boundary condition Eq. (A.6) with $n = 0$ for solvation energy calculation.

$$a_0 = -\frac{q}{a} \frac{1}{\varepsilon \kappa a} \frac{1}{k'_0(\kappa a) + B_{00} i'_0(\kappa a)} \tag{A.9}$$

So the potential on the charge center is written as,

$$\varphi(r \to 0) = -\frac{q}{a} \frac{1}{\varepsilon \kappa a} \frac{k_0(\kappa a) + B_{00} i_0(\kappa a)}{k'_0(\kappa a) + B_{00} i'_0(\kappa a)} - \frac{q}{a}, \tag{A.10}$$

where $B_{00} = \sum_{\nu=0}^{\infty} A_{00}^{\nu} k_{-2\nu}(\kappa R) = k_0(\kappa R)$.

We validate this result comparing with the exact analytic expression of solvation energy of a single sphere in which the charge is located on the origin of sphere. This is done by using the infinite separation of the two spheres($R \to \infty$). If $R \to \infty$, then

$$B_{00}(R \to \infty) \left\{ = k_0(\kappa R) = \frac{\pi}{2} \frac{e^{-\kappa R}}{\kappa R} \right\} \to 0. \tag{A.11}$$

So the solvation energy W is given as,

$$\begin{aligned} W(R \to \infty) &= \frac{1}{2} q\varphi \\ &= \frac{1}{2} \left\{ -\frac{q^2}{a} \frac{1}{\varepsilon \kappa a} \frac{k_0(\kappa a)}{k'_0(\kappa a)} - \frac{q^2}{a} \right\} \\ &= \frac{1}{2} \frac{q^2}{a} \frac{1 - (1+\kappa a)\varepsilon}{(1+\kappa a)\varepsilon}. \end{aligned} \tag{A.12}$$

This result is exactly equal to the analytic solution for the solvation of the single charged sphere. To calculate the electrostatic interaction free energy of the two identical spheres, we need to subtract the interaction potential of the infinitely separated spheres from the potential between two spheres with finite distance, that is, $\varphi_{AB} = \varphi(R) - \varphi(R \to \infty)$ where A and B represent two spheres.

$$\begin{aligned} \varphi_{AB} &= \varphi(R) - \varphi(R \to \infty) \\ &= -\frac{q}{a} \frac{1}{\varepsilon \kappa a} \left\{ \frac{k_0(\kappa a) + k_0(\kappa R) i_0(\kappa a)}{k'_0(\kappa a) + k_0(\kappa R) i'_0(\kappa a)} + \varepsilon \kappa a \right\}. \end{aligned} \tag{A.13}$$

We used this expression to validate our solution based on the fast multipole method.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition.

Anderson, M. J., Hansen, C. L., and Quake, S. R. (2006). Phase knowledge enables rational screens for protein crystallization. *Proceedings of the National Academy of Sciences of the United States of America*, 103(45):16746–16751.

Arakawa, T., Bhat, R., and Timasheff, S. N. (1990). Why preferential hydration does not always stabilize the native structure of globular proteins. *Biochemistry*, 29(7):1924–1931.

Atkinson, K. and Han, W. (2009). *Numerical Solution of Fredholm Integral Equations of the Second Kind*. Texts Applied in Mathematics. Springer New York, 3 edition.

Bahadur, R. and Zacharias, M. (2008). The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. *Cellular and Molecular Life Sciences*, 65(7):1059–1072.

Barrett, R., Berry, M., Chan, T., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., and van der Vorst, H. (1994). *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM.

Bateman, K. S., James, M. N. G., Anderson, S., Lu, W., Qasim, M. A., and Michael, L. J. (2000). Deleterious effects of beta-branched residues in the s_1 specificity pocket of

Streptomyces griseus proteinase b (sgpb): Crystal structures of the turkey ovomucoid third domain variants ile18i, val18i, thr18i, and ser18i in complex with sgpb. *Protein Science*, 9(1):83–94.

Boistelle, R., Lafont, S., Veesler, S., and Astier, J. (1997). Comparison of solubilities and molecular interactions of bpti molecules giving different polymorphs. *Journal of Crystal Growth*, 173:132–140.

Bonneté, F., Ferté, N., Astier, J., and Veesler, S. (2004). Protein crystallization: Contribution of small angle x-ray scattering (saxs). *J. Phys. IV France*, 118:3–13.

Brandsdal, B. O., Smals, A. O., and Aqvist, J. (2006). Free energy calculations show that acidic p1 variants undergo large pk_a shifts upon binding to trypsin. *Proteins: Structure, Function, and Bioinformatics*, 64(3):740–748.

Brock, K., Talley, K., Coley, K., Kundrotas, P., and Alexov, E. (2007). Optimization of electrostatic interactions in protein-protein complexes. *Biophysical Journal*, 93(10):3340–3352.

Brooks, B., Bruccoleri, R., Olafson, D., States, D., Swaminathan, S., and Karplus, M. (1983). Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217.

Byrd, R. H., Schnabel, R. B., and Shultz, G. A. (1987). A trust region algorithm for nonlinearly constrained optimization. *SIAM Journal on Numerical Analysis*, 24(5):1152–1170.

Carnie, S. L. and Chan, D. Y. (1993). Interaction free energy between identical spherical colloidal particles: The linearized poisson-boltmann theory. *Journal of Colloid and Interface Science*, 155:297–312.

Carugo, O. and Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Science*, 6:2261.

Chayen, N. E. and Saridakis, E. (2008). Protein crystallization: from purified protein to diffraction-quality crystal. *Nat Meth*, 5(2):147–153.

Chew, W. C., Jin, J.-M. M., E., and Song, J. (2001). *Fast and Efficient Algorithms in Computational Electromagnetics*. Artech House, Boston, MA.

Crosio, M., Janin, J., and Jullien, M. (1992). Crystal packing in six crystal forms of pancreatic ribonuelease. *Journal of Molecular Biology*, 228:243.

Dasgupta, S., Iyer, G., Bryant, S., Lawrence, C., and Bell, J. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins: Structure, Function, and Genetics*, 28:494.

Davies, M., Toseland, C., Moss, D., and Flower, D. (2006). Benchmarking pka prediction. *BMC Biochemistry*, 7(1):18.

Delphine, C. B., David, M. R., and Jan, H. J. (2008). Very fast prediction and rationalization of pk_a values for protein-ligand complexes. *Proteins: Structure, Function, and Bioinformatics*, 73(3):765–783.

Dong, F. and Zhou, H.-X. (2006). Electrostatic contribution to the binding stability of protein-protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 65(1):87–102.

Elcock, A., Sept, D., and McCammon, J. A. (2001). Computer simulation of protein-protein interactions. *Journal of Physical Chemistry B*, 105:1504.

Epton, M. A. and Dembart, B. (1995). Multipole translation theory for the three-dimensional laplace and helmholtz equations. *SIAM Journal on Scientific Computing*, 16(4):865–897.

Farnum, M. and Zukoski, C. (1999). Effect of glycerol on the interactions and solubility of bovine pancreatic trypsin inhibitor. *Boiphysical Journal*, 76(5):2716–2726.

Fawcett, W. R. (2004). *Liquid, Solution, and Interfaces: From Classical Macroscopic Descriptions to Modern Microscopic Details*. Oxford Universiy Press, New York.

Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation – From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Press.

Gabrielsen, M., Nagy, L. A., DeLucas, L. J., and Cogdell, R. J. (2010). Self-interaction chromatography as a tool for optimizing conditions for membrane protein crystallization. *Acta Crystallographica Section D*, 66(1):44–50.

Gallagher, W. H. and Woodward, C. K. (1989). The concentration dependence of the diffusion coefficient for bovine pancreatic trypsin inhibitor: A dynamic light scattering study of a small protein. *Biopolymers*, 28(11):2001–2024.

George, A., Chiang, Y., Guo, B., Arabshahi, A., Cai, Z., and Wilson, W. W. (1997). [6] second virial coefficient as predictor in protein crystal growth. In Charles W. Carter, J., editor, *Methods in Enzymology*, volume Volume 276, pages 100–110. Academic Press.

Gerdts, C. J., Tereshko, V., Yadav, M. K., Dementieva, I., Collart, F., Joachimiak, A., Stevens, R. C., Kuhn, P., Kossiakoff, A., and Ismagilov, R. F. (2006). Time-controlled microfluidic seeding in nl-volume droplets to separate nucleation and growth stages of protein crystallization. *Angewandte Chemie, International Edition*, 45(48):8156–8160, S8156/1–S8156/14.

Gilson, M. K., Given, J. A., Bush, B. L., and McCammon, J. A. (1997). The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*, 72(3):1047–1069. 0006-3495.

Glendinning, A. B. and Russel, B. W. (1982). The electrostatic repulsion between charged spheres from exact solutions to the linearized poisson-boltmann equation. *Journal of Colloid and Interface Science*, 93(1):95–104.

Greengard, L. (1988). *The rapid evaluation of potential fields in particle systems*. ACM distinguished dissertations. MIT Press, Cambridge, Mass.

Greengard, L. and Rokhlin, V. (1997). A fast algorithm for particle simulations. *J. Comput. Phys.*, 135(2):280–292.

Grigsby, J. J., Blanch, H. W., and Prausnitz, J. M. (2000). Diffusivities of lysozyme in aqueous mgcl2 solutions from dynamic light-scattering data: Effect of protein and salt concentrations. *The Journal of Physical Chemistry B*, 104(15):3645–3650.

Gripon, C., Legrand, L., Rosenman, I., Vidal, O., Robert, M. C., and Boue, F. (1997). Lysozyme-lysozyme interactions in under- and super-saturated solutions: a simple relation between the second virial coefficients in h2o and d2o. *Journal of Crystal Growth*, 178(4):575–584.

Guex, N. and Peitsch, M. C. (1997). Swiss-model and the swiss-pdb viewer: An environment for comparative protein modeling.

Gumerov, N. A. and Duraiswami, R. (2005). *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions (The Elsevier Electromagnetism Series)*. Elsevier Science, Amsterdam.

Guo, B., Kao, S., McDonald, H., Asanov, A., Combs, L. L., and William Wilson, W. (1999). Correlation of second virial coefficients and solubilities useful in protein crystal growth. *Journal of Crystal Growth*, 196(2-4):424–433.

Harvey, A. H. and Lemmon, E. W. (2004). Correlation for the second virial coefficient of water. *Journal of Physical and Chemical Reference Data*, 33(1):369–376.

Helland, R., Otlewski, J., Sundheim, O., Dadlez, M., and Smal, A. O. (1999). The crystal structures of the complexes between bovine [beta]-trypsin and ten p1 variants of bpti. *Journal of Molecular Biology*, 287(5):923–942.

Hunter, R. J. (1987). *Foundations of Colloid Science*. Oxford University Press, Oxford.

Israelachvili, J. (1985). *Intermolecular and Surface Forces*. Academic Press.

Jackson, J. D. (1999). *Classical Electrodynamics*. John Wiley and Sons, New York, 3 edition.

Janin, J. and Chothia, C. (1990). The structure of protein-protein recognition sites. *Journal of Biological Chemistry*, 265(27):16027–16030.

Janin, J. and Rodier, F. (1995). Protein-protein interactions at crystal contacts. *Proteins: Structure, Function, and Genetics*, 23:580.

Jasinski, J. P. and Foxman, B. M. (2007). http://people.brandeis.edu/~foxman1/teaching/indexpr.html.

Jorgensen, W. L. and Tirado-Rives, J. (1988). The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666.

Juffer, A. J., Botta, E. F. F., van Keulen, B. A. M., van der Ploeg, A., and Berendsen, H. J. C. (1991). The electric potential of a macromolecule in a solvent: A fundamental approach. *J. Comput. Phys.*, 97(1):144–171.

Kelley, C. T. (1999). *Iterative methods for optimization*. Frontiers in applied mathematics.

Kim, B., Song, J., and Song, X. (2010). Calculations of the binding affinities of protein-protein complexes with fast multipole method. *To be published*.

Kim, B. and Song, X. (2010). Calculations of the second virial coefficients of proteins with extended fast multipole method. *To be published*.

Kofke, D. A. (1993a). Direct evaluation of phase coexistence by molecular simulation via integration along the saturation line. *The Journal of Chemical Physics*, 98(5):4149–4162.

Kofke, D. A. (1993b). Gibbs-duhem integration: a new method for direct evaluation of phase coexistence by molecular simulation. *Molecular Physics: An International Journal at the Interface Between Chemistry and Physics*, 78(6):1331 – 1336.

Krowarsch, D., Dadlez, M., Buczek, O., Krokoszynska, I., Smalas, A. O., and Otlewski, J. (1999). Interscaffolding additivity: binding of p1 variants of bovine pancreatic trypsin inhibitor to four serine proteases. *Journal of Molecular Biology*, 289(1):175–186.

Kuehner, D., Heyer, C., Ramsch, C., Fornefeld, U., Blanch, H., and Prausnitz, J. (1997). Interactions of lysozyme in concentrated electrolyte solutions from dynamic light-scattering measurements. *Biophysical Journal*, 73(6):3211 – 3224.

Kui, H., Michael, N. G. J., Wuyuan, L., Michael, Laskowski, J., and Stephen, A. (1995). Water molecules participate in proteinase-inhibitor interactions: Crystal structures of leu18, ala18, and gly18 variants of turkey ovomucoid inhibitor third domain complexed with streptomyces griseus proteinase b. *Protein Science*, 4(10):1985–1997.

Kurinov, I. V. and Harrison, R. W. (1995). The influence of temperature on lysozyme crystals. structure and dynamics of protein and water. *Acta Crystallographica Section D*, 51(1):98–109.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285(5):2177–2198.

Lu, B., Cheng, X., and Andrew McCammon, J. (2007). "new-version-fast-multipole-method" accelerated electrostatic calculations in biomolecular systems. *Journal of Computational Physics*, 226(2):1348–1366.

Lu, W., Apostol, I., Qasim, M. A., Warne, N., Wynn, R., Zhang, W. L., Anderson, S., Chiang, Y. W., Ogin, E., Rothberg, I., Ryan, K., and Laskowski, M. (1997). Binding of amino acid side-chains to s1 cavities of serine proteinases. *Journal of Molecular Biology*, 266(2):441–461.

McQuarrie, D. A. (1976). *Statistical Mechanics*. Harper Collins, New York.

Messiah, A. (1962). *Quantum Mechanics*, volume 2. North Holland, Amsterdam, Netherlands.

Millefiori, S., Alparone, A., Millefiori, A., and Vanella, A. (2008). Electronic and vibrational polarizabilities of the twenty naturally occurring amino acids. *Biophysical Chemistry*, 132(2-3):139–147.

Miller, B. T., Singh, R. P., Klauda, J. B., Hodoscek, M., Brooks, B. R., and Woodcock, H. L. (2008). Charmming: A new, flexible web portal for charmm. *Journal of Chemical Information and Modeling*, 48(9):1920–1929.

Murrell, J. N. and Jenkins, A. D. (1994). *Properties of Liquids and solutions*. John Wiley and Sons, Chichester, England, 2nd edition.

Muschol, M. and Rosenberger, F. (1997). Liquid–liquid phase separation in supersaturated lysozyme solutions and associated precipitate formation/crystallization. *The Journal of Chemical Physics*, 107(6):1953–1962.

Neal, B. L., Asthagiri, D., and Lenhoff, A. M. (1998). Molecular origins of osmotic second virial coefficients of proteins. *Biophysical Journal*, 75(5):2469–2477.

Neal, B. L. and Lenhoff, A. M. (1995). Excluded volume contribution to the osmotic second virial coefficient for proteins. *AIChE Journal*, 41(4):1010–1014.

Noya, E. G., Vega, C., Doye, J. P. K., and Louis, A. A. (2007). Phase diagram of model anisotropic particles with octahedral symmetry. *The Journal of Chemical Physics*, 127(5):054501.

Parsegian, V. (1975). *Physical Chemistry:Enriching Topic from Colloid and Surface Science*. Theorex, La Jolla, Calif.

Petsev, D. N., Wu, X., Galkin, O., and Vekilov, P. G. (2003). Thermodynamic functions of concentrated protein solutions from phase equilibria. *The Journal of Physical Chemistry B*, 107(16):3921–3926.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). Ucsf chimera?-?a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

Qasim, M. A., Ranjbar, M. R., Wynn, R., Anderson, S., and Laskowski, M. (1995). Ionizable p residues in serine proteinase inhibitors undergo large pk shifts on complex formation. *Journal of Biological Chemistry*, 270(46):27419–27422.

Read, R., Fujinaga, M., Sielecki, A., and James, M. (1983). Structure of the complex of streptomyces griseus protease b and the third domain of the turkey ovomucoid inhibitor at 1.8-a resolution. *Biochemistry*, 22(19):4420–33.

Rokhlin, V. (1985). Rapid solution of integral equations of classical potential theory. *Journal of Computational Physics*, 60(2):187–207.

Rosenbaum, D. F. and Zukoski, C. F. (1996). Protein interactions and crystallization. *Journal of Crystal Growth*, 169(4):752–758.

Roth, C., Neal, B., and Lenhoff, A. (1996). Van der waals interactions involving proteins. *Biophysical Journal*, 70(2):977 – 987.

Sanner, M. (1996). http://www.scripps.edu/~sanner/html/msms_home.html.

Schreiber, G., Frisch, C., and Fersht, A. R. (1997). The role of glu73 of barnase in catalysis and the binding of barstar. *Journal of Molecular Biology*, 270(1):111–122.

Service, R. F. (2005). Structural genomics, round 2. *Science*, 307:1554.

Song, J. and Chew, W. C. (1998). The fast illinois solver code: Requirements and scaling properties. *IEEE Comput. Sci. Eng.*, 5(3):19–23.

Song, X. (2002a). An inhomogeneous model of protein dielectric properties: Intrinsic polarizabilities of amino acids. *Journal of Chemical Physics*, 116(21):9359–9363.

Song, X. (2002b). The protein data bank entries are 6lyt,1lzt,1lkr,0lzt,1lys and the crystallization conditions are from wwwbmcd.nist.gov:8080/bmcd. unplublished results.

Song, X. (2003). The extent of anisotropic interactions between protein molecules in electrolyte solutions. *Molecular Simulation*, 29(10):643 – 647.

Song, X. (2009). Solvation dynamics in ionic fluids: An extended debye–h[u-umlaut]ckel dielectric continuum model. *The Journal of Chemical Physics*, 131(4):044503–8.

Song, X. and Zhao, X. (2004). The van der waals interaction between protein molecules in an electrolyte solution. *The Journal of Chemical Physics*, 120(4):2005–2009.

TargetDB (2010). http://targetdb.pdb.org/statistics/index.html.

Tessier, P. M., Lenhoff, A. M., and Sandler, S. I. (2002). Rapid measurement of protein osmotic second virial coefficients by self-interaction chromatography. *Biophysical Journal*, 82(3):1620–1631.

Vaughan, C. K., Buckle, A. M., and Fersht, A. R. (1999). Structural response to mutation at a protein-protein interface. *Journal of Molecular Biology*, 286(5):1487–1506.

Veesler, S., Lafont, S., Marcq, S., Astier, J., and Boistelle, R. (1996). Prenucleation, crystal growth and polymorphism of some proteins. *Journal of Crystal Growth*, 168:124–129.

Vega, C., Sanz, E., Abascal, J. L. F., and Noya, E. G. (2008). Determination of phase diagrams via computer simulation: methodology and applications to water, electrolytes and proteins. *Journal of Physics: Condensed Matter*, 20(15):153101.

Velev, O. D., Kaler, E. W., and Lenhoff, A. M. (1998). Protein interactions in solution characterized by light and neutron scattering: Comparison of lysozyme and chymotrypsinogen. *Biophysical Journal*, 75(6):2682–2697.

Vilker, V. L., Colton, C. K., and Smith, K. A. (1981). The osmotic pressure of concentrated protein solutions: Effect of concentration and ph in saline solutions of bovine serum albumin. *Journal of Colloid and Interface Science*, 79(2):548 – 566.

Weisstein, E. W. (2010). http://mathworld.wolfram.com/LeastSquaresFittingPowerLaw.html.

Wondratschek, H. and Müller, U. (2002). *International Tables for Crystallography, Volume A: Space Group Symmetry*. Springer, New York, 5 edition.

Yoon, B. and Lenhoff, A. (1990). A boundary element method for molecular electrostatics with electrolyte effects. *Journal of Computational Chemistry*, 11:1080.

Yoshida, K.-i. (2001). Applications of fast multipole method to boundary integral equation method. *PhD Thesis, Department of Global Environment Engineering, Kyoto University*.

Zamyatnin, A. A. (1984). Amino acid, peptide, and protein volume in solution. *Annual Review of Biophysics and Bioengineering*, 13(1):145–165.