2009

Proteins in silico-modeling and sampling

Parimal Kar
Michigan Technological University

PROTEINS IN SILICO- MODELING AND SAMPLING

By

PARIMAL KAR

A DISSERTATION

Submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

(Physics)

MICHIGAN TECHNOLOGICAL UNIVERSITY

2009

This dissertation, "Proteins in Silico- Modeling and Sampling", is hereby approved in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in the field of Physics.

DEPARTMENT:
Physics

Signatures:

Dissertation Advisor
Dr. Ulrich H.E. Hansmann

Committee
Dr. Max Seel

Dr. Ranjit Pati

Dr. Marta Wloch

Department Chair
Dr. Ravi Pandey

Date

Dedication

To My Maa o Baapi

Contents

List of Figures

List of Tables

Acknowledgments

During my time as a PhD student I have come to realized that a graduate student can not finish his PhD thesis alone. Several individuals play a vital role during his graduate study. I would like to take this opportunity to thank all those individuals who have played a role in my doctoral life. First and foremost, I would like to thank my advisor Prof. Ulrich H. E. Hansmann (Uli). It has been a previleged to be his student. He has helped me learn the subject of protein folding and has extended my knowledge beyond the scope of this subject. When I joined his group I knew nothing about proteins. Now I am writing my PhD thesis on proteins. Without his help and guidance this would never have been possible. He is truly a great researcher and is very enthusiastic in his works. He always has time for his students. Whenever I stopped by his office to ask any silly questions, he listened carefully and answered my questions. He is very organized and is a perfectionist in his research.

I would also like to thank all of my previous and current group members- Dr. Yanjie Wei, Liang Han, Dr. Siegfried Höfinger, Dr. Walter Nadler, Dr. Maksim Kouza, Priya Anand, Nari Kang. They all deserve high gratitude of me. During my PhD study I collaborated with Dr. Höfinger, Dr. Nadler and Dr. Wei. I learnt a lot from their faculties. We had stimulating and interesting discussions. I would like to especially emphasize Dr. Höfinger's name here. I collaborated with him on three projects and learned a lot from him. He taught me programming and scripting. He also provided me with a script for canonical REMD, which I modified for Microcanonical REMD. He is always ready with his helping hands.

Additionally he is a friend, philosopher and mentor to me.

I would also like to convey my sincere gratitude to my PhD committee members- Dr. Max Seel, Dr. Ranjit Pati, and Dr. Marta Wloch for their advice, constant encouragement, and helpful suggestions concerning my research. Their wise suggestions helped me a lot to improve my dissertation.

I want to take this opportunity to express my gratitude to our Physics Department Chair, Dr. Ravi Pandey for his help, support, and constant encouragement during my doctoral life. I am also thankful to Mike, Andrea, Kathy and Marg. In my first year as a PhD student, I was a teaching assistant. I was very scared and nervous to accept this position, but because of Mike's and Wil's help and tips, I had a sweet experience in teaching. Thank you very much Mike and Wil!

I would also like to thanks my batch-mates, office-mates, and department-mates – Madhusudan, Archana, Abhisekh, Pradeep, Neluka, Joy, Dr. Wang, and Chee. I would also like to reserve a special thanks to the Bengali Community in Houghton. Without them, life would be pretty boring in Houghton. They have never let me recognize awarness of the fact that I am several thousands miles away from my home. Thank you very much Siladitya, Banasree, Partha, Colina, Ananyo, Saikat, and Subhasish. Your collective friendship is invaluable to me.

I would like to devote a special thought to my parents for their never-ending support. They are my constant source of inspiration and energy. If I achieve anything in my life, it is all due to them. Thank you very much my dear Maa (mother) and Bapi (father) for ensuring my education and constantly inspiring me to achieve my goals. I Love you so much !

Finally, I would like to thank my sweet wife Jayeeta. All I can say is that it would take another thesis to express my love for you. You have always upheld me with your love, patience, and encouragement. You were always there with me even in those days when I spent more times in front of my laptop than with you. Pursuing my PhD has been a long journey. There have been many ups and downs throughout this journey, but you have smoothened those ripples and made my journey easy. I love you !

Abstract

Proteins are linear chain molecules made out of amino acids. Only when they fold to their native states, they become functional. This dissertation aims to model the solvent (environment) effect and to develop & implement enhanced sampling methods that enable a reliable study of the protein folding problem *in silico*.

We have developed an enhanced solvation model based on the solution to the Poisson-Boltzmann equation in order to describe the solvent effect. Following the quantum mechanical Polarizable Continuum Model (PCM), we decomposed net solvation free energy into three physical terms– Polarization, Dispersion and Cavitation. All the terms were implemented, analyzed and parametrized individually to obtain a high level of accuracy.

In order to describe the thermodynamics of proteins, their conformational space needs to be sampled thoroughly. Simulations of proteins are hampered by slow relaxation due to their rugged free-energy landscape, with the barriers between minima being higher than the thermal energy at physiological temperatures. In order to overcome this problem a number of approaches have been proposed of which replica exchange method (REM) is the most popular. In this dissertation we describe a new variant of canonical replica exchange method in the context of molecular dynamic simulation. The advantage of this new method is the easily tunable high acceptance rate for the replica exchange. We call our method Microcanonical Replica Exchange Molecular Dynamic (MREMD). We have described the theoretical frame work, comment on its actual implementation, and its application to Trp-

cage mini-protein in implicit solvent. We have been able to correctly predict the folding thermodynamics of this protein using our approach.

Chapter 1

Introduction

Proteins are essential bio-macromolecules and the building blocks of all cells. Genetic information is encoded into DNA (deoxyribonucleic acid), but must be translated into proteins. To produce a protein, a corresponding gene is first transcribed into mRNA (messenger Ribonucleic Acid) and then translated into a chain of amino acids in the ribosome. This nascent polypeptide folds to its native structure within a very short time frame.

Proteins are cell's work-horse. As enzymes, they catalyze many biochemical reactions, as structural elements they are the founding elements of blood vessels, epidermal keratin etc. As antibodies, they fight with the infection [21]. The mechanism of all these biophysical processes depend on the correct fold of their respective polypeptide chains into 3-dimensional native structure.

1.1 Biochemistry of Proteins

The basic unit of a protein is an amino acid. There are twenty different types of amino acids found in proteins. All of them have a central carbon atom (C_α) and hydrogen atom, amino group (NH_2) and carboxyl groups $(COOH)$ are attached to it. The side chain which is attached to the C_α differentiates various amino acids. The twenty naturally occurring types of amino acids are shown in Fig.1.1. Amino acids are linked together by a peptide bond to form a protein. A peptide bond is formed when the carboxyl group $(COOH)$ of the first amino acid reacts with the amino group of the next releasing water. The formation of a peptide bond is shown in Fig.1.2. The protein chain runs from amino (N) terminus to carboxyl (C) terminus. The formation of a peptide bond generates a "main chain" or "backbone" from which various "side chains" point outwards.

The backbone atoms of a polypeptide are composed of C_α to which the side chain is attached, a NH group bound to C_α, and a carbonyl group $C = O$, where the carbon atom C is attached to C_α. Therefore, the basic repeating unit along the backbone is ($NH - C\alpha$H-CO).

Based upon the chemical structure of the side chains, amino acids are classified into three categories:

† The first class comprises of hydrophobic side chains– Ala (A), Val (V), Leu (L), ILE (I), Phe (F), Pro (P), and Met (M).

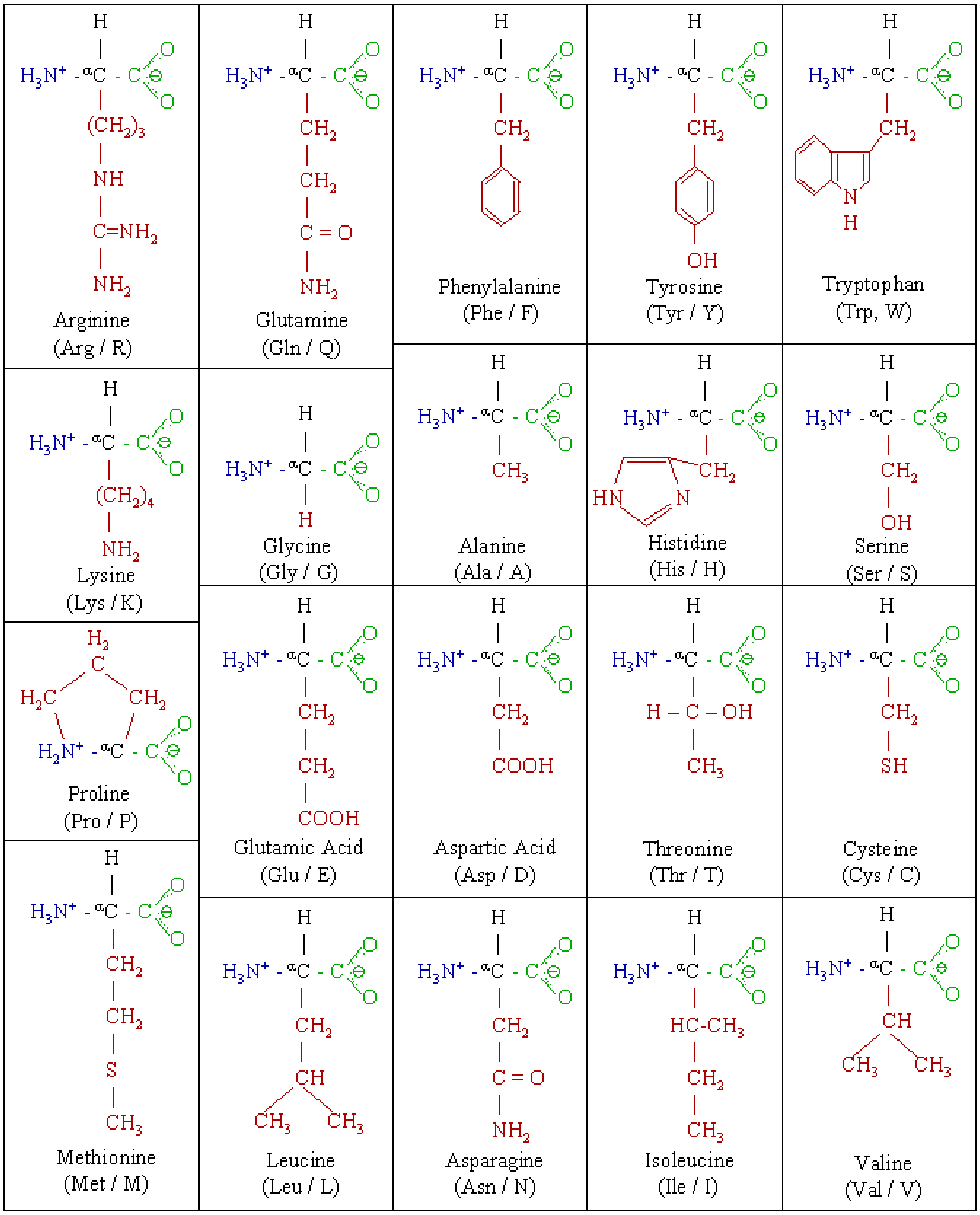


Figure 1.1: Structure of twenty different amino acids with their 3-letter and single letter codes [1]. Copyright notice can be found in Appendix D.

† The second class is made of charged residues—Asp (D), Glu (E), Lys (K), and Arg (R).

† The third class is made of those with polar residues– Ser (S), Thr (T), Cys (C), Asn

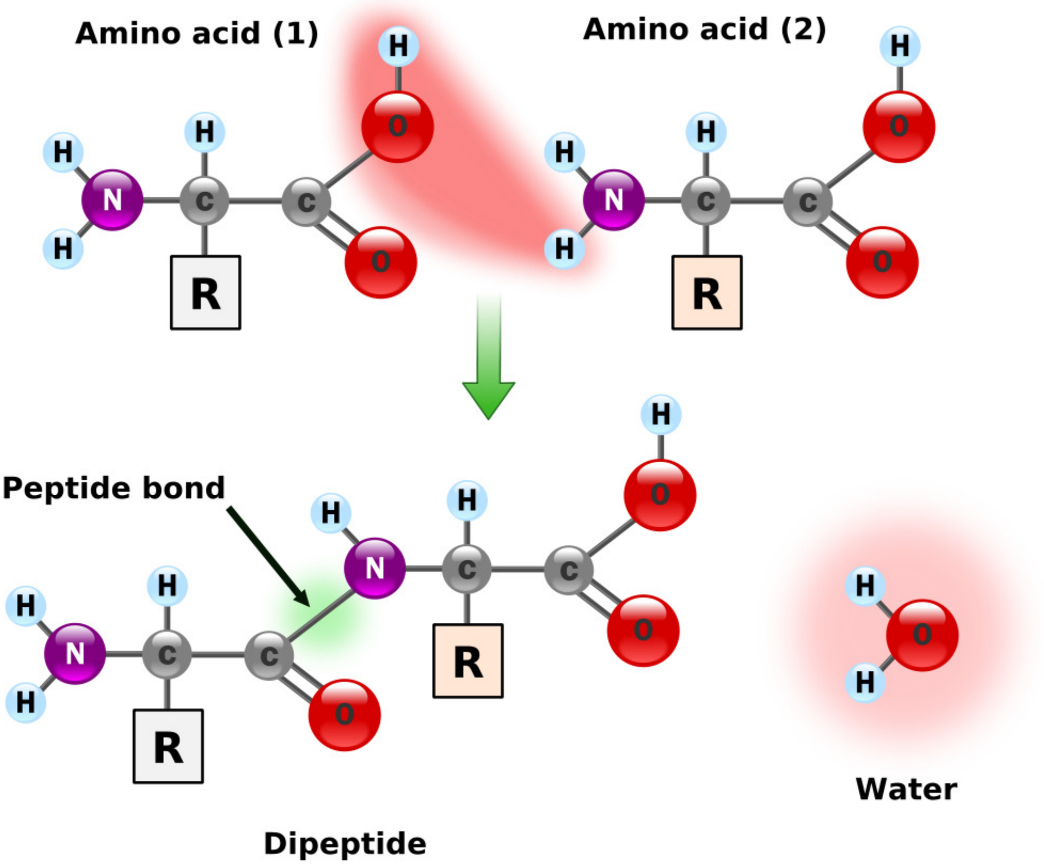


Figure 1.2: Condensation of two amino acids to form peptide bond [2]. Copyright notice can be found in Appendix D.

(N), Gln (Q), His (H), Tyr (Y), Trp (W).

The amino acid Glycine (G) is the simplest amino acid among all the twenty naturally occurring amino acids, since it has only a hydrogen atom as the side chain. The amino acid Proline (P) differs from the others as both ends of the side chain are covalently bound to the main chain forming a ring structure.

Apart from glycine, all amino acids are chiral molecules. They can exist in two different forms with different hands, known as L or D. During the protein synthesis process, only

L-forms are found. The general structure of an α-amino acid is shown in Fig 1.3.

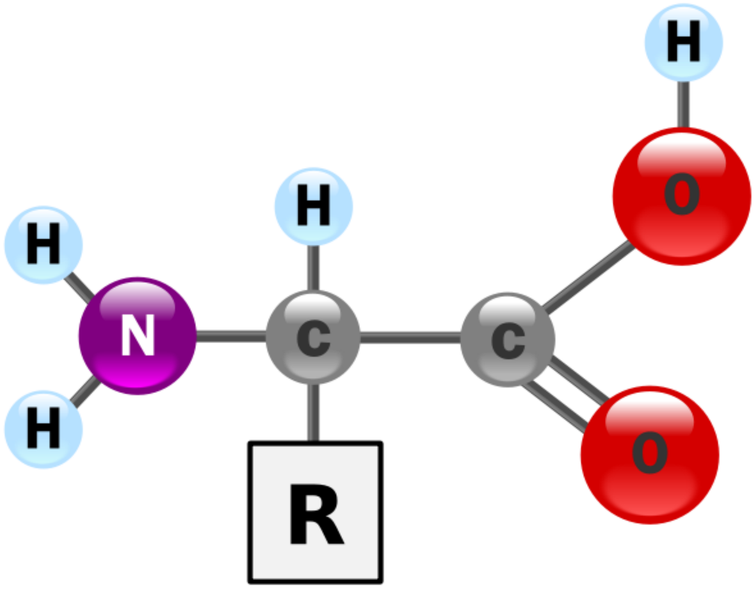


Figure 1.3: The general structure of an α-amino acid with amino group on the left and carboxyl group on the right [3]. Copyright notice can be found in Appendix D.

1.1.1 Protein Organization Level

Proteins are made up in combination of twenty different types of amino acids. There are four different structural hierarchy present in proteins. These are shown in Fig.1.4. **Primary Structure**: The sequence of amino acids is called the primary structure. It starts from the amino-terminal (N-terminal) end to the carboxyl-terminal (C-terminal) end.

Secondary Structure: Secondary structure is the local arrangements of amino acids in proteins and occurs due to the hydrogen bonding interactions between adjacent amino acids.

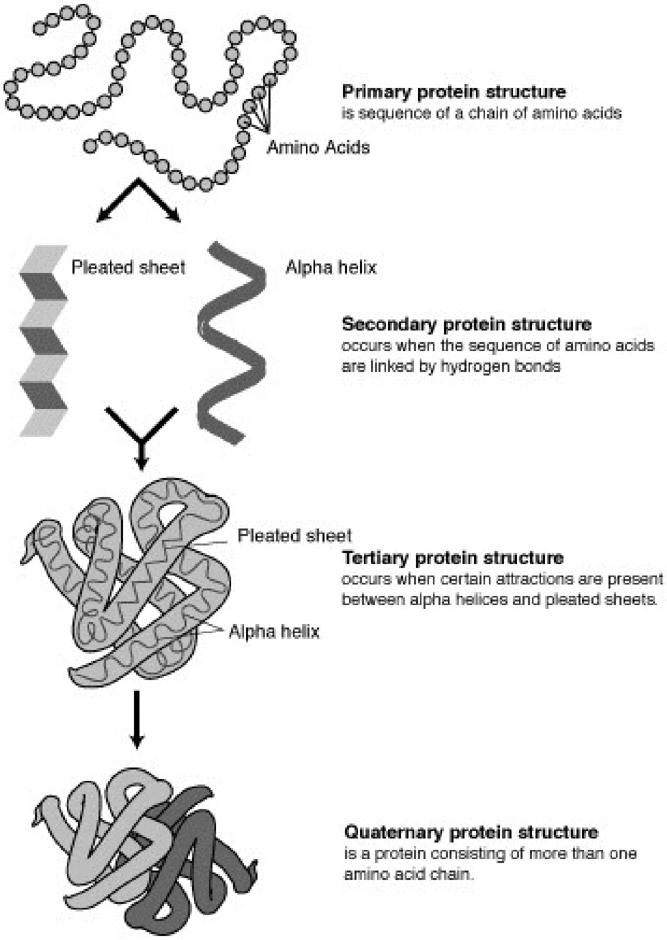

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Figure 1.4: Structural levels of proteins [4]. © National Human Genome Research Institute, the arm of NIH, USA.

The hydrogen bonds in proteins form between the backbone carboxyl oxygens and amide hydrogens. The patterns of backbone hydrogen bonds define the secondary structures– α-helices, β-sheets and turns and loops [22, 23, 24].

α-helix: α-helices are spring-like structures. The inner part of the helix is formed by the coiled backbone and the side chains project outwards in a helical array. The structure is stabilized by hydrogen bonds between NH and CO groups of the backbone four residues earlier. Each residue is 0.15 nm long along the helix axis and a rotation of 100°. This gives 3.6 amino acid residues per turn of helix, in a clockwise direction resulting in a pitch of 0.54 nm. The helix is about 0.6 nm in diameter with all of the side chains sticking outwards [25]. Helix can be right handed or left handed. Since there is a less steric clash between the side chains and the main chain, right handed helices are energetically more favorable, and all α-helices found in proteins are right handed (except glycine-based helix).

β-sheet: A β-sheet is formed by linking two or more β-strands by hydrogen bonds. In a β-sheet, a β strand is almost fully extended rather than being coiled as in α-helices. The distance between adjacent amino acids along β-strand is $\sim$ 0.35 nm (3.5 Å) whereas a distance of 0.15 nm (1.5 Å) is observed along α-helix. Beta sheets can be parallel (adjacent chains run in the same direction) or anti-parallel (adjacent chains run in opposite direction). In parallel arrangement, for each amino acid, the NH group is hydrogen bonded to the CO group of one amino acid on the adjacent strand, whereas the CO group is hydrogen bonded to the NH group on the amino acid two residues further along the chain. In anti-parallel arrangement, the NH group and the CO group of each amino acid are bonded to the CO and NH group of a partner on the adjacent chain. β-sheets are formed by many strands with minimum being two (e.g., β-hairpin) and maximum being ten (e.g., β-barrel).

Tertiary Structure: Tertiary structure is the compact three dimensional structure of a single polypeptide at which they are functional. This structure is formed by assembly of secondary structural elements along with turns and loops into a 3-dimensional arrangement. Tertiary structures are stabilized by weak interactions such as hydrophobic interactions, hydrogen bonding, ionic interactions and by a covalent bond called the disulfide bond. This structure is very compact due to the efficient packing of amino acid side chains [26, 23, 22]. This structure often consists of a hydrophobic core with charged residues on the surface of the protein. The charged residues on the surface gives the protein its biological activity, thus making it biologically functional.

Quaternary Structure: Sometimes more than one tertiary structures of independent folded chains self assemble themselves under physiological conditions to perform specific functions. These structures are known as Quaternary structures (e.g., hemoglobin). This is the fourth level structural organization present in a protein. Non-covalent interactions, hydrophobic interactions, disulfide bonds are responsible for the stabilization of the quaternary structure [23, 22].

1.2 Protein Functions

Proteins are essential macromolecules in all living organisms. They perform virtually all the works in a cell. A specific protein performs a specific function. The function of a protein depends on its structure. Some proteins act as enzymes while others either fight with the infection or provide structural support. Several types of proteins and their functions are described below.

Enzymes are the largest class of proteins. All the biochemical reactions are controlled by the enzymes. Enzymes help to speed up the biochemical reactions significantly by lowering the activation energy of the reactions. Hence they are called biological catalysts. In presence of an enzyme in a cell, a biochemical reaction can be 10^{17} times faster than the same reaction in absence of that particular enzyme [27]. They are very specific to the biochemical reactions. A specific enzyme can only perform to its corresponding substrate. The functions of enzymes are influenced by their environmental factors, such as temperature and pH.

Structural proteins are fibrous and stringy in nature and they provide strength and support to cell and tissues. They are insoluble in water. Examples include keratin, elastin, and collagen. Keratins are found in the form of hair, nail, wool, feather, horn etc. while collagens and elastins provide support for connective tissues such as tendons and ligaments.

Storage proteins act as a reservoir for some essential nutrients. Ferritin is a kind of storage protein. It stores iron and controls the iron level in the body.

Transports proteins are carrier proteins which move particles (ions, proteins etc.) across intracellular compartments and membranes. Examples include hemoglobin, myoglobin, and cytochromes. Hemoglobin and myoglobin are responsible for transportation of oxygen molecules through blood. Cytochromes act as a electron carrier proteins.

Antibodies are specialized proteins which fight with the foreign invaders (antigen) into our bodies. They help our immune system to fight against the bacteria and viruses. They are found in blood or other bodily fluids [25].

Hormonal proteins are regulatory proteins which regulate the function of other proteins under physiological conditions and help to coordinate certain bodily activities. Examples of some regulatory proteins are insulin, oxytocin, and somatotropin. Insulin regulates glucose metabolism in our body by controlling blood-sugar concentration. Oxytocin stimulates contractions in female during child-birth while somatotropin is a growth hormone that stimulates protein production in muscle cells. The gene expression also needs to be regulated by regulatory proteins such as repressors which block gene transcription [24, 23, 22, 25].

Contractile proteins are responsible for movement. Actin and myosin are two contractile proteins. Both are involved in muscle contraction and movement.

1.3 Protein Folding Problem

Proteins are linear chain molecules of amino acids. To perform their own work, proteins need to adopt their correct three dimensional structure, known as native structure. This process of self-assembly is known as **'protein folding'**. Proteins fold themselves into their 3-dimensional functional form within a very short time frame ranging from milliseconds (ms) to microseconds (μs) [26, 28]. But this time frame is very large with respect to computer time (cpu time) which makes it a grand challenge to study folding of proteins *in silico*. So far, the detailed knowledge of the folding mechanism is missing [29]. But in last few years, this field has seen tremendous progress and a general picture of protein folding mechanism is appearing [30].

Protein folding is a rapid and unique process. According to Christian Anfinsen, for any protein's native structure (final 3-dimensional structure) is determined solely by its amino acid sequence [31]. This is known as Anfinsen's dogma. Anfinsen's dogma suggests that at physiological conditions (pressure, temperature, solvent etc.), at which protein folding takes place, the final native configuration is a **unique, stable and kinetically accessible minimum of the free energy**.

How the proteins adopt their native structure from amino acid sequence in a reasonable time frame is a central question in the protein folding problem. For a 100 residues protein, there are nearly 10^{18} conformations available. Even if we have access to the world's fastest search algorithm, still it will take 10^{30} years to find to structure corresponding to the lowest energy. This apparent contradiction is known as Levinthal Paradox [32, 33]. This makes protein folding problem a computationally difficult to study on the computer.

To overcome these problems, many folding mechanisms have been proposed: the framework model [26], the hydrophobic collapse model [34], the diffusion-collision model [5], and the funnel theory [35]. The funnel theory (see Fig. 1.5) is the most popular theory to describe the protein folding process. According to this theory, proteins have very rugged free-energy landscape with multi local-minima (corresponds to unfolded, random state) and a single global minimum, which corresponds to the folded native structure. There are many protein folding pathways.

Protein folding can be studied either by experiment or simulation methods. Different experimental techniques are used to study the folding of a protein. Protein could be unfolded in high concentrations of a chemical denaturant (e.g., urea) and then could be refolded by

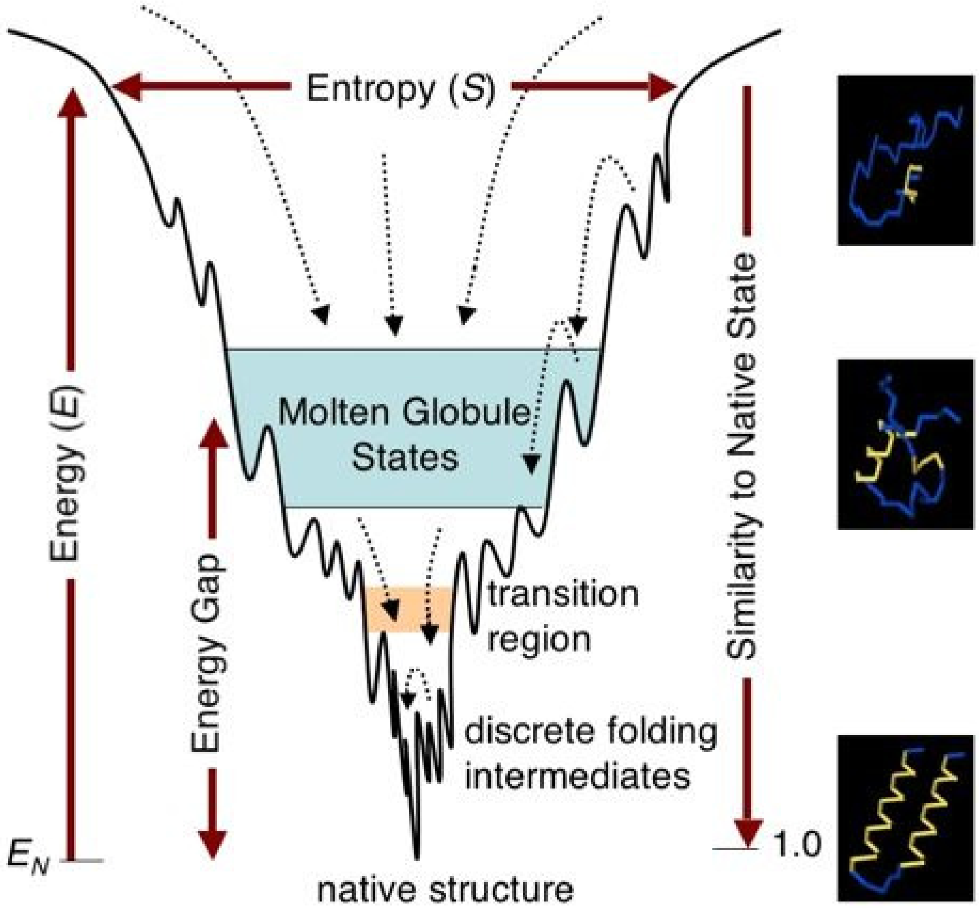


Figure 1.5: Funnel shaped free energy landscape of proteins [5]. Reproduced with permission (See appendix D, Fig D.4).

diluting the solution. During the refolding process, many experimental techniques are used to study the structural changes of the protein. NMR is an experimental technique which provides high time resolution and spatial resolution [36]. Protein engineering methods can be used to probe the role of individual role of residues during the folding and unfolding processes [25]. This way we can study the mutated proteins and the effects of mutation on the stability of proteins [37].

Structure determination is a very important research topic, since structure determines the function of a protein. Much efforts have been put in this area which can be seen from the creation of the protein data bank (PDB) [38] maintained by Brookheaven National Laboratory (BNL). Approximately 50,000 structures have been deposited so far. These structures are determined either by NMR method or by X-ray crystallography [39] or other methods. It is very hard to determine the structures of membrane proteins. NMR has a protein size limit that allows it to be studied [25, 21]. We can also study the protein folding problem *in silico* [40]. Two most important simulation methods are molecular dynamic (MD) simulation [41] and Monte Carlo (MC) method [42]. In MD simulation, we solve the Newton's equation and get the trajectory of the system. While in MC, we sample the configuration space of the protein randomly according to designated criteria [43]. Computer simulations are limited by insufficient computational resources, inefficient algorithms, and several approximations in the force-fields.

1.4 My Contributions

Protein folding is a mysterious process. The mechanism of folding is not yet unraveled. Although a general picture of folding is becoming clearer and a tremendous progress in this area has been made in last decades. Still many questions are remained unsolved. Examples include the role of environment (solvent) in folding, folding/misfolding and related diseases, protein-protein interaction etc. During the folding process, proteins may not fold to their correct 3-dimensional structure. This is called protein misfolding. Protein misfolding is associated with many diseases, such as Alzheimer's disease, madcow disease and many other prion related diseases. This kind of study can lead us to design rational drugs to fight with these diseases.

Proteins can work only when they adopt their specific 3-dimensional structures. So to understand the function, we need to study the structure. If we unravel this mystery of sequence-structure relationship, then we can design protein of our desired function which will be extremely useful for medical purposes and in nano-biotechnology industries. Protein adopts its native structure in its native environment. This native environment influences the folding process. It is essential to model this environment (solvent) to understand the folding. This thesis mainly deals with modeling this environment (solvent).

Roughness in free energy landscape (see Fig. 1.5) makes it hard to study via computer simulations. Conventional simulation methods (e.g., molecular dynamics and Monte Carlo) are not good enough to study the protein folding problem. So we need to design efficient algorithms to study protein folding problem *in silico*. In my doctoral studies, I have worked on development and implementation of enhanced sampling methods and applied it to study the folding thermodynamics of a Trp-cage mini-protein in an implicit solvent.

There are different factors that govern the protein folding process, such as (i) mechanical factors– temperature, pressure, etc. (ii) biological factors– molecular chaperons which assist in protein folding, and (iii) chemical factors– pH, salt effect, solvent etc.

Most of the proteins can achieve their 3-dimensional form in their native environment. Solvent plays a vital role in the folding and dynamical process. In my doctoral work I studied extensively the design, implementation and parameterization of an enhanced implicit solvation model. Initial work of my dissertation concentrated on modeling the solvent effect reliably and accurately.

Solvent effects could be incorporated in simulations in two ways. We can treat the solvent in their full atomic details or we can represent the solvent as a structureless dielectric continuum medium. First approach is known as explicit solvent model while the later is called implicit solvent model. Explicit models are much more accurate but computationally costly. On the other hand, implicit models are relatively less accurate but computational cost is smaller than the explicit model. This model can enable us to study protein of relatively larger size. This dissertation deals with an enhanced implicit solvent.

In the implicit solvent model, the solute of interest is represented in their full atomic detail whereas the surrounding medium (solvent) is characterized by structureless continuum, interacting primarily via polarization, dispersion, repulsion, and cavitation effects [44, 45, 46, 47, 48]. The polarization term is obtained by solving the Poisson-Boltzmann (PB) equations [49, 50, 51, 52, 53, 54, 55]. The Poisson-Boltzmann (PB) equation is solved either by finite difference method (FDPB) [49, 50, 51, 52] or by boundary element method (PB/BEM) [54, 55]. Our model is based upon the solution of the Poisson-Boltzmann equation within boundary element framework. This method reduces a 3-dimensional volume integral to a two-dimensional surface integral, which helps us to save computational time. Our model follows quantum mechanical model of implicit solvent, known as Polarizable

Continuum Model (PCM) [46]. Following PCM, we also decompose the total solvation free energy into three terms:

$$\Delta G_{solv} = \Delta G_{pol} + \Delta G_{disp} + \Delta G_{cav,rep} \tag{1.1}$$

Each term is treated separately and parameterized to obtain the highest level of accuracy. This dissertation work is dealt with the first two terms- polarization and dispersion. The third term, Cavitation is extensively studied by Mahajan et al. [56]. ΔG_{pol} is obtained by solving the Poisson-Boltzmann equation using boundary element method. For the dispersion term, we use the Caillet-Claverie [57, 58] approach in the context of boundary element formalism. We also have implemented popular molecular dynamics package AMBER-style [18] of representation to incorporate the dispersion effect. The cavitation term is expressed via the revised Pierotti approximation (rPA) [11, 59, 56], which is based on the Scaled Particle Theory (SPT) [60, 61]. Once this model is parameterized, we apply our method successfully to estimate the electro static potential (ESP) of an anti-fungal protein. ESP maps are very useful in structural biology. We parameterized our model for various solvents–water, methanol, ethanol, cyclohexane etc. We compare our results to quantum mechanical results as well as experimental results.

The second part of my dissertation is concerned with the sampling algorithm to explore the configurations space of proteins. The most popular enhanced sampling method is Replica Exchange [62]. We have implemented a variant of the replica exchange method, which we call Microcanonical Replica Exchange Molecular Dynamic (MREMD) method.

This dissertation is organized in the following way. In chapter 2, I describe different dominant forces acting on proteins and how to model these interactions. I will also describe the different popular forcefields and simulation methods often used in biomolecular simulations. In particular I will explain different terms in AMBER, CHARMM, OPLS forcefields and briefly describe simulation methods such as molecular dynamics (MD), Monte Carlo (MC), simulated annealing (SA), and Replica Exchange method (REMD).

In chapter 3, I give a brief introduction to different solvation models. First, a brief description of explicit solvent models is provided. Then I describe the formalism of implicit solvent model from potential of mean force standpoint. I will also discuss different popular implicit models used in biomolecular simulations. We comment on their shortcomings and limitations.

In chapter 4, we investigate the influence of boundary elements on the outcome of polarization free energy. We have used two popular surface computation programs– SIMS [10] and Connolly's MSROLL program [63] for surface discretization. We have found that SIMS is faster than Connolly's program, since we need less number of boundary elements in case of SIMS to reach to the same level of accuracy. We describe a three-stage procedure to analyze the dependence of Poisson-Boltzmann calculations on the shape, size, and geometry of the boundary between the solute and solvent. Our study is carried out within the boundary element formalism, but our results are also of interest to finite difference techniques of Poisson-Boltzmann calculations. At first, we identify the critical size of the geometrical elements for discretizing the boundary, and thus the necessary resolution required to establish numerical convergence. In the following two steps we perform reference calculations on a set of dipeptides in different conformations using the Polarizable Continuum Model (PCM) and a high-level Density Functional as well as a high-quality basis set. Afterwards, we propose a mechanism for defining appropriate boundary geometries. Finally, we compare the classic Poisson-Boltzmann (PB) description with the Quantum Chemical description, and aim at finding appropriate fitting parameters to get a close match to the reference data. Surprisingly, when using default AMBER partial charges and the rigorous geometric parameters derived in the initial two stages, no scaling of the partial charges is necessary and the best fit against the reference set is obtained automatically.

In chapter 5, implementation and parameterization of the dispersion term is described. We implement a well-established concept to consider dispersion effects within a Poisson-Boltzmann approach of continuum solvation of proteins. We consider Caillet-Claverie [57, 58] approach for our purpose. The theoretical framework is particularly suited for boundary element methods. Free parameters are determined by comparison to experimental data as well as high level Quantum Mechanical reference calculations. The method is general and can be easily extended in several directions. We have tested our model on various chemical substances and found to yield good quality estimates of the solvation free energy without obvious indication of any introduced bias. Once optimized, we applied our model to a series of proteins and then we studied factors, such as protein size or partial charge assignments.

Further optimization and application of our model to estimate the electrostatic potential (ESP) for various charge assignments is discussed in chapter 6.

We have also developed an enhanced sampling method which is a variant of canonical replica exchange molecular dynamic simulation. Instead of temperature ladder (for canonical REMD), an energy ladder is used in our approach. We call our method as Microcanical Replica Exchange Molecular Dynamic simulation. We describe in chapter 7, the theoretical framework of our model and its application to a Trp-cage protein in an implicit solvent. We also have studied the folding thermodynamics of this protein using our method.

Chapter 8 summarizes my dissertation and sketches future directions.

Chapter 2

Forcefields and Simulations

2.1 Forcefields

Proteins are the most chemically, structurally and functionally diverse biological macromolecules. Proteins can perform their function only if they attain their compact 3-dimensional structure. This structure is called the native structure. To study this problem *in silico*, we need two ingredients:

- † mathematical models that can describe the free energy force fields accurately and
- † reliable computational algorithms that will enable us to explore the protein configurations space efficiently and quickly.

The success of the simulation depends on how accurate our forcefield is. For smaller systems, quantum mechanical calculations (e.g, DFT, *ab initio*) can be performed in the gas phase. But proteins are large macromolecules. They have thousands of atoms plus the solvent atoms. So quantum mechanical calculations are not feasible here. In this case, force field simulations have to be done. In atomistic models, atoms are the smallest particles in the system rather than the electrons and nuclei in quantum mechanical models. Empirical energy function includes relatively simple terms to describe the physical interactions that dictate the structure and dynamics of proteins. These empirical forcefields allow us to study proteins for a longer time. A forcefield refers not only to the functional form, but also the parameter sets associated with this function. These parameter sets are generally obtained from experimental results or from high-level quantum mechanical calculations. Forcefield could be all-atom, united-atom or coarse-grained. In all-atom forcefields, the parameters for all the atoms in the system are assigned while in united-atom, the hydrogen and carbon atoms in methyl and methylene groups are treated as single interaction group. Coarse-grained forcefields use an even more reduced presentation of the system and this is used for a very long time simulation of biomolecules.

The basic functional form of a forcefield consists of two terms: bonded term related to atoms linked by covalent bond and nonbonded terms describing the long-range electrostatic interactions and short-range van der Waals force. So we can write

$$V_{Tot} = V_{bonded} + V_{nonbonded} \tag{2.1}$$

We can further decompose both terms into the following terms:

$$V_{bonded} = V_{bond} + V_{angle} + V_{dihedral} \tag{2.2}$$

$$V_{nonbonded} = V_{electrostatic} + V_{vanderWaals} \tag{2.3}$$

Some forcefields include out-of-plane dis-torsions (improper torsion) and cross-terms such as stretch-stretch, stretch-bend, etc.

2.2 Bonded Interactions

Bond Stretching: In a classical forcefield, we treat both the bond and angle terms as a harmonic oscillator. But the bond breaking is not allowed. The mathematical form of the potential energy associated with the bond stretching is given by the following equation

$$V_{bond} = K_l \left(l - l_0\right)^2 \tag{2.4}$$

where K_l is the force constant, l_0 is the equilibrium bond length. This model is valid when l does not deviate much from l_0. If l deviates much from l_0 and if we want to calculate the molecular structures and vibrational frequencies more accurately, then we should go beyond harmonic approximations and higher terms should also be included in such situations. A more realistic and accurate way of treating covalent bond at higher stretching is to incorporate Morse Potential [64]. This is a much more expensive model. The functional form of Morse potential is

$$V(r) = D_e\left(1 - e^{-a(r-r_e)}\right)^2 \tag{2.5}$$

where r is the distance between atoms, r_e is the equilibrium bond distance and D_e is the well-depth and $a = \left(\frac{k_e}{2D_e}\right)^{1/2}$. The two functions are shown in Fig.2.1. Typically molecular dynamics and Monte Carlo simulations are performed at room temperature. It is sufficient to use harmonic oscillator potential for both the terms– bond stretching and angle bending.

Angle Bending: Angle bending terms are also modeled with a harmonic oscillator potential. The functional form can be written as

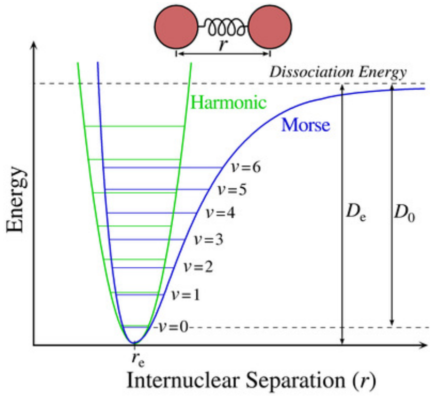


Figure 2.1: The harmonic oscillator potential (green) and Morse potential (blue) [6]. Copyright notice can be found in Appendix D.

$$V_{angle} = K_\theta\left(\theta - \theta_0\right)^2 \tag{2.6}$$

where K_θ is the force constant and θ_0 is the equilibrium bond angle. Accuracy can be improved by considering higher order terms. K_θ is much lower than K_l since the energy needed to distort an angle from its equilibrium is much less compare to the energy needed to distort a bond length from its equilibrium.

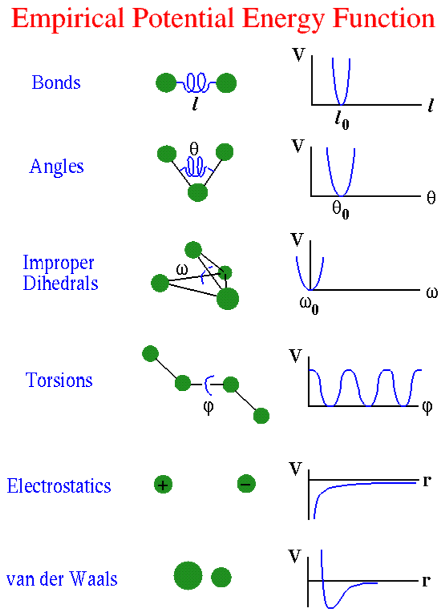


Figure 2.2: Graphical representation of empirical potential energy function [7]. Reproduced with permission (see Appendix D, Fig D.3).

Torsional Term: Two types of torsional potentials are used in biomolecular forcefields. They are the dihedral angle potential and improper torsional potential. Both potentials depend on a quartet of atoms, bonded in one way or the other. A proper dihedral angle potential depends on four consecutive bonded atoms, while the improper torsion potential relies on three atoms centered around a fourth atom. The proper dihedral angle potential is mostly used to constrain the rotation around a bond while the improper torsion term is used to maintain chirality on a tetrahedral extended heavy atom or to maintain planarity of certain atoms. The main difference between both torsion potentials is the definition of the torsional angle and the functional form of the potential function (see Fig.2.2). The functional form of both potentials are given by Eqn.2.7 and Eqn.2.8.

$$V_{torsion} = \sum_{n=0} \frac{V_n}{2}[1 + cos(n\phi - \delta)] \tag{2.7}$$

where ϕ is the torsional angle and δ is the phase and V_n is the barrier height [43] and n is the multiplicity. Multiplicity is a positive, nonzero integer number.

$$V_{improper} = K_{\omega}(1 - cos 2\omega) \tag{2.8}$$

where ω is the improper torsion angle and K_{ω} is the force constant [43]. Most of the variation in structure and relative energies is due to the complex interplay between the torsional and non-bonded contributions [43].

2.2.1 Nonbonded Interactions

In the previous section, I have described different covalent or bonded interactions and their mathematical models. These covalent forces provide stabilization to the primary structures of proteins. But for higher structural levels of proteins, the stabilizing forces are nonbonded in nature. Non-covalent or nonbonded forces include electrostatic interaction, van der Waals interaction and hydrogen bonding and hydrophobic interaction. Sometimes disulfide bonds provide stabilization to the tertiary/quaternary proteins. Since nonbonded forces are numerous, they are predominant forces in proteins. So it is essential to model these nonbonded forces for successful biomolecular simulations. These interactions range from 4 to 30 kJ/mol that can not bind two atoms together. They are usually modeled as a function of some inverse power of distance [23, 22, 43, 25].

Electrostatic Interaction: The electrostatic interaction term involves the interaction between two partial atomic charges q_i and q_j separated by a distance r_{ij}. We know that like charges repel and opposite charges attract each other. The interaction potential is given by Coulomb's law

$$V_C = \frac{q_i q_j}{\varepsilon r_{ij}} \tag{2.9}$$

where ε is the dielectric constant of the medium.

The strength of an electrostatic force depends on the environment. In vacuum the dielectric constant is 1 and the interaction force is the highest. For water the dielectric constant is 80 and the force is 80 times weaker than vacuum. Much computational power is devoted to compute this term. In molecular dynamics, a Particle Mesh Ewald method is used to compute this term efficiently.

van der Waals Interaction

The van der Waals interaction and static repulsion are treated with the Lennard-Jones (LJ)

6-12 potential of the form

$$V_{LJ} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.10}$$

The LJ 6-12 potential contains just two adjustable parameters: the collision diameter σ_{ij} (the separation for which the energy is zero) and the well depth ε_{ij} [43]. The Lennard-Jones potential is characterized by two parts: an attractive part that is proportional to r^{-6} and a repulsive part that varies as r^{-12} [43]. The r^{-6} variation is the same power-law relationship found for the leading term in theoretical treatment of the dispersion energy such as the Drude model [43]. Although the r^{-12} is reasonable for rare gases, but is too steep for other systems such as hydro carbons [43]. However, the LJ 6-12 potential is widely used in biomolecular forcefields.

Hydrogen Bonding: In proteins, a hydrogen bond is formed when a donor (hydrogen) is bonded to a strong electronegative partner like oxygen in water or nitrogen in the backbone of a polypeptide chain. Hydrogen bonds are directional. It is extremely important for protein folding since it stabilizes the formation of secondary structures such as α-helices and β-sheets. The positively charged hydrogen can interact with a negatively polarized partner like oxygen or nitrogen. This interaction is often modeled as pure electrostatic or as a dipole-dipole interaction [43, 65]. Some of the functional forms are as following

$$V_{HB1} = \frac{q_i q_j}{\varepsilon r_{ij}} \tag{2.11}$$

$$V_{HB2} = \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{10}} \tag{2.12}$$

$$V_{HB3} = cos(\theta)\left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}\right) + (1 - cos(\theta))\left(\frac{C}{r_{ij}^{12}} - \frac{D}{r_{ij}^{10}}\right) \tag{2.13}$$

where q_i and q_j are the charges on atoms i and j, r_{ij} is the distance between atoms i & j, θ is the angle of hydrogen bonds and A, B, C, D are parameters determining the strength of the potential [65, 43].

Apart from these forces, solvent interactions are also included in forcefields. I have described these solvent interactions and their modeling in the next chapter.

2.2.2 Popular Forcefields

Some of the popular forcefields which are commonly used in biomolecular simulations are– Assisted Model Building with Energy Refinement (AMBER) [18], CHemistry at Harvard Macromolecular Mechanics (CHARMM) [33], Optimized Potentials for Liquid Simulations (OPLS) [66] and GROningen MOlecular Simulation (GROMOS) [67]. Here we will show the functional form of forcefields for AMBER only .

AMBER is a family of force fields developed by Kollman's group. It can also refers to the simulation package that also uses these forcefields (FF). The latest version is AMBER 10. For my study, I used AMBER 9 molecular dynamic package. The AMBER force field consists of five energy terms in which the first term represents the covalent bond stretching energy; the second term is the angle bending energy ; the third term represents the energy barrier for rotating a bond; the fourth term describes the van der Waals energy; and the last term is the electrostatic energy. The functional form of CHARMM is similar to AMBER but with a different parameter set. CHARMM forcefields are also included in NAMD [68] molecular dynamic package.

$$\begin{aligned} E_{AMBER} &= E_{bond} + E_{angle} + E_{torsional} + E_{electronic} + E_{vdw} \\ &= \sum_{bond} k_b(r - r_0)^2/2 + \sum_{angle} k_a(\theta - \theta_0)^2/2 \\ &\quad + \sum_{torsion} V_n[1 + cos(n\omega - \gamma)]/2 \\ &\quad + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} (4\varepsilon_{i,j}[(\frac{\delta_{ij}}{r_{ij}})^{12} - (\frac{\delta_{ij}}{r_{ij}})^6]) \\ &\quad + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \end{aligned} \tag{2.14}$$

OPLS force field has been developed by Jorgensen's group and is implemented in the software packages-BOSS (molecular dynamic package), MCPRO (Monte Carlo based protein simulation package), GROMACS (molecular dynamic package) [69] and TINKER (molecular dynamic package) [70].

Scheraga and his coworkers have developed the ECEPP (Empirical Conformational Energies of Polypeptide and Proteins) forcefield for peptides and proteins [71]. Here the fixed geometries of amino acid residues are used to simplify the potential energy surface. Energy minimization is carried out in protein torsional angle space. This forcefield is also used in the protein simulation software package SMMP [72]. The ECEPP force field describes the energy function of a protein as a sum E_{ECEPP} consisting of electrostatic energy E_C, Lennard-Jones energy E_{LJ}, hydrogen-bonding energy E_{HB} and a torsional energy E_{Tor}:

$$\begin{aligned} E_{ECEPP} &= E_C + E_{LJ} + E_{HB} + E_{Tor} \\ &= \sum_{(i,j)} \frac{332 q_i q_j}{\varepsilon r_{ij}} \\ &\quad + \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ &\quad + \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ &\quad + \sum_{l} U_l (1 \pm \cos(n_l \xi_l)) \ , \end{aligned} \tag{2.15}$$

where r_{ij} is the distance between the atoms i and j, ξ_l is the l-th torsion angle. The prefactor 332 in the electrostatic energy term follows from the fact that the units of the energy terms are in kcal/mol. Apart from these force fields, there are also many other forcefields– MM2 [73], MM3 [74], CFF [75], polarizable force fields [76, 77].

2.3 Simulation Method

The energy landscape of a protein is generally very rough with a lot of high barriers and low regions. This rugged free energy landscape (FEL) makes it difficult to sample the energy space thoroughly. Simulations are hampered due to slow relaxation. Numerous simulation methods have been developed to address this protein folding problem. All the methods are based on either deterministic method (molecular dynamics) or stochastic (Monte Carlo) approach. Here we will discuss only conventional molecular dynamics, Monte Carlo (MC), and replica exchange methods (REM).

Simulations are performed at microscopic level. But we are interested in macroscopic observable such as pressure, temperature, energy and heat capacities etc. Statistical mechanics helps us to calculate these macroscopic properties from microscopic simulations. A microscopic state of N-particle system refers to a point in phase space. Phase space has 6N dimension and is characterized by 3N coordinates of position $\vec{r}^N$ and 3N coordinates of momentum $\vec{p}^N$.

An ensemble is defined as a sum of all possible systems which have different microscopic states but the macroscopic/thermodynamic states are identical. Four common ensembles are listed below.

Microcanonical Ensemble (NVE): This ensemble is described by the fixed number of particles (N), constant volume (V) and constant energy (E). This refers to an isolated system.

Canonical Ensemble (NVT): As the name suggests, it refers to an ensemble where the total number of particles N, the volume V and the temperature T are kept constant. Temperature is kept constant by a heat bath. At equilibrium, the probability of being in a microscopic state with energy E_i is

$$P_C = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} \tag{2.16}$$

where the denominator is called the canonical partition function Z [78, 79] and $\beta = 1/k_B T$. k_B is the Boltzmann constant and T is the absolute temperature. Equilibrium is characterized by the minimum Helmholtz free energy. Helmholtz free energy is given by following equation.

$$F = -k_B T \ln(Z) \tag{2.17}$$

Entropy (S) of the system can be obtained from the Helmholtz free energy.

$$S = -\left(\frac{\partial F}{\partial T}\right)_{V,N} \tag{2.18}$$

The internal energy and the heat capacity are obtained from following two formulas.

$$U = F + TS \tag{2.19}$$

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V \tag{2.20}$$

where U is the internal energy and C_V is the heat capacity. Generally, protein simulations are performed in NVT ensemble.

Isobaric-Isothermal Ensemble (NPT): In this ensemble, not only the total number of particles N is fixed,but also pressure and temperature of the system are kept constant. The equilibrium state of the system is characterized by the minimum of Gibbs free energy [25, 43].

Grand Canonical Ensemble (μVT): In grand canonical ensemble the total number of particle N is allowed to change, but volume V, temperature T and chemical potential μ are kept constant. The probability of finding the system in a microscopic state i characterized by energy E_i is given by following equation

$$P_G = \frac{e^{-(E_i - \mu N_j)/k_B T}}{Z_G} \tag{2.21}$$

where Z_G is grand canonical partition function [79, 78]. The entropy and grand canonical potential are given by

$$S_G = (E - \mu N)/T + k_B ln Z_G \tag{2.22}$$

and

$$\Phi = -k_B T ln Z_G = E - T S_G - \mu N \tag{2.23}$$

respectively. In statistical mechanics, we express average values by their ensemble averages. The ensemble average of an observable A is given by Eqn.2.24

$$A = \int\int d\vec{p}^N d\vec{r}^N A(\vec{p}^N, \vec{r}^N) \rho(\vec{p}^N, \vec{r}^N) \tag{2.24}$$

where the probability density of the ensemble is given by the following eqn.2.25

$$\rho(\vec{p}^N, \vec{r}^N) = \frac{1}{Z} e^{-\beta H(\vec{p}^N, \vec{r}^N)} \tag{2.25}$$

where $\beta = 1/K_B T$, K_B is the Boltzmann's factor, Z is the partition function and H is the Hamiltonian of the system.

The partition function Z is given by the eqn.2.26

$$Z = \int\int d\vec{p}^N \vec{r}^N e^{-\beta H(\vec{p}^N,\vec{r}^N)} \tag{2.26}$$

In molecular dynamics simulation, the points in ensemble are calculated sequentially in time. So in molecular dynamics, we calculate the time average properties only. The time average of an observable A is

$$< A >= \lim_{\tau \to \infty} \int_{t=0}^{\tau} dt A(\vec{r}^N, \vec{p}^N) \tag{2.27}$$

When τ approaches to infinity, the value of Eqn.2.27 will be the true average of A. This is called the 'Ergodic Theory'. This theory tells us that ensemble average is equal to the time average, i.e., $< A >_{ensemble} =< A >_{time}$.

2.3.1 Molecular Dynamics

Molecular dynamics (MD) simulation is deterministic in nature. In MD, Newton's equations are solved. From the knowledge of force, we can calculate the potential of each particle of the system. Integration of the equation then gives us a trajectory that describes the position, velocity and acceleration of the particles as a function of time. From this trajectory, the average properties of the system are estimated. Molecular dynamic simulations are computationally expensive compared to Monte Carlo.

The force acting on a particle of mass m_i is

$$\vec{F}_i = m_i \vec{a}_i \tag{2.28}$$

where $\vec{a}_i$ and $\vec{F}_i$ are the acceleration and force of the i^{th} particle of mass m_i. In MD simulation, the potential energy function is provided. Force is just the gradient of the potential energy.

$$\vec{F}_i = -\nabla_i V \tag{2.29}$$

Combining Eqn.2.28 and Eqn.2.29 we get

$$m_i \frac{d^2\vec{r}_i}{dt^2} = -\frac{dV}{d\vec{r}_i} \tag{2.30}$$

So by integrating the above equation, we get the time evolution of position and velocity. This integration is done by many algorithms. Here I will describe the mathematical formulation of only one integrator– Leapfrog algorithm. This algorithm is most popular and is used in AMBER molecular dynamics package also. In this algorithm, we first evaluate velocities at time *t + dt/2*. These velocities are then used to calculate the positions $\vec{r}$ at time t+dt. This means that the velocities leap over the position, then positions leap over the velocities, hence, the name leapfrog.

$$\vec{r}(t+dt) = \vec{r}(t) + \vec{v}(t+dt/2)\,dt \tag{2.31}$$

$$\vec{v}(t+dt/2) = \vec{v}(t-dt/2) + \vec{a}(t)dt \tag{2.32}$$

It is to be noted here that positions and velocities are not calculated at the same time. There is *dt/2* time-step difference between them. The velocity at time t is approximated by the following equation.

$$\vec{v}(t) = \frac{1}{2}\left[\vec{v}(t-dt/2) + \vec{v}(t+dt/2)\right] \tag{2.33}$$

Other popular integrators are velocity Verlet's and Beeman's integrator. Both algorithms enable us to calculate positions and velocities at the same time. But both methods are complex and computationally expensive compared to leapfrog.

2.3.2 Monte Carlo

Monte Carlo (MC) is based on exploring the energy landscape by random changes in the geometry of the molecule under study. Monte Carlo enables us to search larger configuration space of the system. Monte Carlo can be described in following steps:

1. First choose a start conformation randomly (current state i) and calculate its energy E_i.

2. Generate a new configuration (*j*) just by making random changes to the initial conformation (*i*). Estimate the energy E_j of the new conformation (*j*).

3. Compare E_j with E_i. If $\Delta E = E_j - E_i$ is less than zero, choose the new conformation

(*j*) as the current conformation and go back to step 2.

4. If $\Delta E = E_j - E_i$ is greater than zero, then accept the state j as current state only if $exp[-\Delta E/(k_B T)]$ is greater than a random number r (between 0 and 1). Otherwise reject the new state j, and take the state i as the current state again and go back to step 2.

5. Iterate until the total conformations generated are sufficient.

It is to be noted here that the simple Monte Carlo method underestimates the contributions from the configurations with extremely small Boltzmann factors (corresponding to low temperature), and overestimates the contributions from the configurations with larger Boltzmann factors (corresponding to high temperatures). To avoid this problem, **importance sampling** strategy is used. This strategy allows us to sample the configurations not completely at random, but is preferentially biased towards the equilibrium conformations at temperature T by sampling the phase space according to their importance. This was proposed by Metropolis et al. and is thus named as Metropolis Monte Carlo algorithm [42].

2.3.3 Replica Exchange Method

The replica exchange method (REM) [80, 81] is the most popular method to study the protein folding problem. Replica exchange method is also known as parallel tempering

[80, 81, 82, 17, 83, 84]. The system for replica exchange method (REM) is composed of M non-interacting replicas of the original system in the canonical ensemble at M different temperatures T_m where m runs from 1 to M [85]. Each replica corresponds to a particular temperature. Replicas and temperatures are related by one-to-one correspondence. This method can be realized into two steps:

i. each replica corresponding to their fixed temperature is simulated simultaneously and independently for a certain Monte Carlo or molecular dynamics steps, and

ii. Choose a pair of replicas and exchange them with the acceptance rate given by the Metropolis criterion.

$$P_{accept} = min\left(1, e^{-\Delta E/T}\right) \tag{2.34}$$

Detailed balance is ensured in replica exchange method. In replica exchange/parallel tempering method, exchanges are attempted between neighboring temperatures only to get good acceptance rate. Random walk in temperature space is realized in this method. Random walk in temperature space (hence potential energy landscape) helps us to avoid local minima and reach to the global minima faster compared to conventional canonical Monte Carlo or molecular dynamics. Large computational power is needed for this method. The number of replicas increase with the system size and hence limit the system size that we can study with this method. Now-a-days large scale of computational power is accessible. So we can think of simulating proteins of larger size with this replica exchange method. The acceptance is low for this method. To increase the acceptance rate, large set of replicas should be used. To overcome this problem we have developed a new variant of parallel tempering method. We call our method as MREMD (Microcanonical Replica Exchange Molecular Dynamics). Our method is described in Chapter 7.

Chapter 3

Solvent Model

3.1 Introduction

Many important processes (e.g., folding, binding, etc.) in biochemistry involve proteins in solution. The surface of the proteins separates between solvent and solute. The interior of the protein has a low dielectric constant and a large set of peptide charges, often found in an arranged manner, such as α -helices and β -sheets. The outside of the molecule has a high dielectric constant. If we want to understand the structural and mechanical basis of these biophysical processes theoretically, then we need to consider solvent properly. Other wise the system under study will become unphysical and will produce unreliable results.

Electrostatic interactions play a crucial role in determining the structure, dynamics and functional properties of bio-polymers, particularly for charged molecules, such as DNA and other poly-electrolytes. When we insert a solute in a solvent of high dielectric constant, the solvent is polarized by the solute's charge, dipole or higher multi-pole moment, which in turn produces a field at the solute which is known as the **reaction field**. In the absence of a solvent with high dielectric constant, the interaction between charges is described by the direct Coulomb term. But in the presence of a solvent with high dielectric constant (e.g., water, $\varepsilon = 80$), a reaction field is produced at the solute which modifies the interactions between charges significantly beyond the direct Coulomb term [55]. Solubility properties are also influenced by the reaction field realized by Max Born [86]. For an electrolyte solution, ions follow the Boltzmann distribution and hence, additional screening occurs due to the charge density in the medium. In 1938 Kirkwood first by using simple geometries apply the screened reaction field to model the solvation of molecules [87, 88] and proteins [89, 90].

Treating the electrostatic interactions in a proper way is critical for molecular dynamics or Monte Carlo simulations of solvated macromolecules. We can treat the solvent effect in two ways. We can add sufficiently large numbers of explicit solvent molecules at the cost of large computer power or we can treat it implicitly, which is less accurate but demands much less computational power.

In my doctoral research, I concentrated on developing reliable, accurate and fast implicit solvation model. In the following sections, I will briefly describe different explicit and implicit solvent models. In later chapters my contributions to the development of Poisson-Boltzmann based implicit solvent model are described.

3.2 Explicit Solvent Model

Explicit solvation models are much more accurate but computationally costly. There are several models which capture the solvent effects explicitly. These models differ from each other depending on the factors, such as number of points used to define the model (atom & dummy sites), whether the structure is flexible or rigid, whether the polarization effect is included in the model.

In general water models could be 3-site, 4-site, 5-site, and 6-site. It is worth mentioning that the computational cost increases with the number of interaction sites. For a 3-site model, 9 distances are required for each pair of water molecules while a 10 distances are required for a 4-site model. While a 5-site model requires 17 distances, for a 6-site model, 26 distances are required. This is tabulated in Table 3.1 In a molecular dynamic simulation,

Water Model	**Distances**
3-site	9
4-site	10
5-site	17
6-site	26

Table 3.1: Different water models and corresponding distances.

if we use a rigid water model then we may introduce additional cost just by constraining the structure. However, by constraining the bond lengths we may be able to use a larger time-step which reduces computational time.

3-Site Model: The simplest water model is the 3-site. This model has three interaction sites and each site corresponds to each atom of the water molecule. Each atom is assigned a point charge and Lennard-Jones (LJ) parameters are assigned to oxygen atoms. Since 3-site models are very simple and computationally efficient, they are widely used in molecular dynamic simulations. Popular 3-site models are- TIPS [91], SPC [92], TIP3P [66] and SPC/E [93]. Apart from the SPC model, most of the 3-site water-models consider rigid geometry of water that matches with the known geometry of the water molecule. The SPC model is based on the assumption of an ideal tetrahedral shape ($\angle HOH = 109.47°$), although we observe an angle of 104.5°.

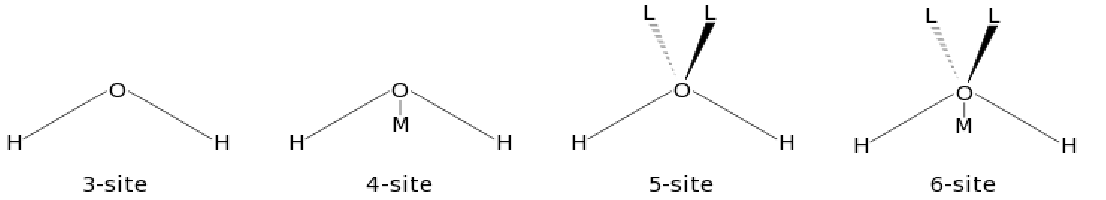


Figure 3.1: Different explicit water models [8]. Copyright notice can be found in Appendix D

The potential function in TIP3P and TIP4P is modeled as

$$E_{ab} = \sum_i \sum_j \frac{k_C q_i q_j}{r_{ij}} + \frac{A}{r_{oo}^{12}} - \frac{B}{r_{oo}^{6}} \tag{3.1}$$

where k_C = 332.1 Å kcal/mol, an electrostatic constant; q_i and q_j are partial charges relative to the charge of an electron; r_{ij} = distance between two atoms or charge sites; A & B are LJ parameters; r_{oo} = distance between two oxygen atoms.

The charged sites may be on the atoms or on dummy sites. In most water models, LJ terms are applied only to the interaction between oxygen atoms, except the TIP3P model implemented in CHARMM [94]. In CHARMM, the Lennard-Jones parameters are placed on hydrogen atoms, although charges are unmodified.

The molecular dynamics package GROAMCS uses SPC and SPC/E water models for explicit solvation. The SPC/E water model adds the following polarization correction term to the potential energy function:

$$E_{pol} = \frac{1}{2} \sum_{i} \frac{(\mu - \mu^0)^2}{\alpha_i} \tag{3.2}$$

where μ = 2.35 D, the dipole of the effectively polarized water molecule; μ^0 = 1.85 D, the dipole moment of an isolated water molecule; and $\alpha_i = 1.608 \times 10^{-40} Fm$, an isotropic polarizability constant. Compared to SPC model, the SPC/E model gives better density and diffusion constant.

4-Site Water Model: In a 4-site water model, the negative charge is placed on the dummy atom placed near the oxygen atom along the bisector of the HOH angle. Better electrostatic distribution is achieved through this arrangement. Different 4-site models are BF [95], TIPS2 [96], TIP4P [66], TIP4P-Ew [97], TIP4P/Ice [98] and TIP4P2005 [99]. The 4-site model was initially adopted by Bernal-Fowler in the year 1938. However, that model

(BF) failed to reproduce the bulk properties of water, such as density and heat of vaporization. Other TIP4P models are just subsequently reparameterized with the aim of specific application.

5-Site Water Model: In the 5-site water model, the negative charge is placed on dummy atoms representing the lone pairs of the oxygen atom with a tetrahedral-like geometry. Different 5-site water models are BNS [100], ST2 [100], TIP5P [101] and TIP5P-E [102]. Ben-Naim and Stillinger first proposed [100] this 5-site model, known as the BNS model, in 1971. Then Stillinger modified [100] this model and proposed the ST2 model in 1974. TIP5P model, proposed by Mahoney and Jorgensen [101] in 2000, gives better results in improvements in the geometry of water dimmer, and the experimentally obtained radial distribution functions are also reproduced quite well by this model. Experimentally obtained the temperature of maximum density of water is also reproduced by TIP-5P model. The TIP5P-E model was developed for use with the Ewald sums.

6-Site Water Model: This model was developed by Nada and van der Eerden [102]. This model combines all the sites of the 4- and 5- site models. The structure and melting of ice are described better by this model compared to the rest of the explicit water models.

3.3 Implicit Solvent Model

The goal of continuum solvent models is to approximate the solute potential of mean force (PMF) [47]. The statistical weight of solute conformations is determined by potential of mean force. The weight can be obtained by averaging over the degrees of freedom (DOF) of the solvent [47]. In this section we describe the different implicit models and their underlying statistical mechanics basis. Roux and Simonson [47] have provided a rigorous formulation of implicit solvent from the perspective of statistical mechanics. Let us consider a protein (solute) immersed in a solvent and the temperature is T. The system will fluctuate over a large number of conformations. The statistical properties of the system can be best characterized by the probability function given by [103]

$$P(\mathbf{X},\mathbf{Y}) = \frac{e^{-U(\mathbf{X},\mathbf{Y})/k_BT}}{\int d\mathbf{X}d\mathbf{Y}e^{-U(\mathbf{X},\mathbf{Y})/k_BT}} \tag{3.3}$$

where X & Y represent the conformations of the solute (protein) and solvent atoms respectively. The potential energy U is decomposed into three terms:

$$U(\mathbf{X},\mathbf{Y}) = U_p(\mathbf{X}) + U_s(\mathbf{Y}) + U_{ps}(\mathbf{X},\mathbf{Y}) \tag{3.4}$$

where $U_p(\mathbf{X})$ = intramolecular solute potential; $U_s(\mathbf{Y})$ = potential due to solvent-solvent interactions; and $U_{ps}(\mathbf{X},\mathbf{Y})$ = potential due to solute-solvent interactions.

For a molecular system, all the physically relevant properties are related to averages weighted by the probability function $P(\mathbf{X},\mathbf{Y})$. The expectation value of any physical quantity $A(\mathbf{X},\mathbf{Y})$ is obtained from the relation

$$<A>= \int d\mathbf{X}d\mathbf{Y}Q(\mathbf{X},\mathbf{Y})P(\mathbf{X},\mathbf{Y}) \tag{3.5}$$

But we are mainly interested in the protein's behavior, not the solvent's. We can define a reduced probability function $\bar{P}(\mathbf{X})$ that depends solely on the solute protein's configuration. The probability distribution for the protein is given by the following equation:

$$\bar{P}(\mathbf{X}) = \int d\mathbf{Y} P(\mathbf{X},\mathbf{Y}) \tag{3.6}$$

It is clear from the expression of $\bar{P}(X)$ that we have been able to get rid of explicit dependence on solvent degrees of freedom. However, at the same time the average influence of the solvent is considered. So we can see that here the solvent coordinates have been 'integrated out'. For a canonical ensemble, the reduced probability $\bar{P}(X)$ takes the following form

$$\bar{P}(\mathbf{X}) = \frac{e^{-W(\mathbf{X})/k_BT}}{\int dX e^{-W(\mathbf{X})/k_BT}} \tag{3.7}$$

where

$$e^{-W(\mathbf{X})/k_BT} = \int dY e^{-[U_p(\mathbf{X})+U_s(\mathbf{Y})+U_{ps}(\mathbf{X},\mathbf{Y})]/k_BT} \tag{3.8}$$

The function $W(\mathbf{X})$ is known as the Potential of Mean Force (**PMF**). Kirkwood first introduced this concept of PMF to describe the average structure of liquids [104]. It is to be noted that the Potential of Mean Force (PMF) is not simply equal to the mean potential energy, i.e., $W(X) \neq < U >_{(x)}$, rather, PMF is the reversible work done by the average force [47]. The average force can be obtained from the gradient of W($\mathbf{X}$).

$$\partial W(\mathbf{X}) = < \partial U/\partial x_i > \; = - \; < \mathbf{F}_{x_i} >_{(x)} \tag{3.9}$$

where x_i denotes the position of the i^{th} solute atom and the symbol $< ... >_{(x)}$ is for the average over all coordinates of the solvent [47].

All the solvent effects are taken into account in $W(\mathbf{X})$ as well as in $\bar{P}(\mathbf{X})$. If we want to express the average of a quantity A($\mathbf{X}$) which depends only on the solute configurations, then we can write

$$< A >= \int A(\mathbf{X})\bar{P}(\mathbf{X}) = \int d\mathbf{X}d\mathbf{Y}A(\mathbf{X})P(\mathbf{X},\mathbf{Y}) \tag{3.10}$$

which is the same as equation 3.5. This equation ensures that an effective potential $W(\mathbf{X})$ exists which makes no explicit reference to the degrees of freedom of the solvent, and the influence of the solvent on the equilibrium properties of the solute is also captured. Generally we can write $W(\mathbf{X})$ as $W(\mathbf{X})$ = $U_p(\mathbf{X})$ + $\Delta W(\mathbf{X})$, where $U_p(\mathbf{X})$ is the solute-solute potential and $\Delta W(\mathbf{X})$ accounts **implicitly but exactly** for the solvent's effect on the protein (solute). The primary goal and challenge of any implicit model is how accurately and efficiently we can model ΔW.

3.3.1 Decomposition of Solvation Free Energy

Among the different intermolecular forces, the dominant ones are– i) short-range repulsive interactions, and ii) long-range electrostatic interactions. Short range forces arise from the Pauli's exclusion principle and the long-range forces arise from the non-uniform distribution of solute charges. Solute-solvent interactions are represented by solvation energies—the free energy of transferring the solute from vacuum to the solvent. This is a three step process:

(i) solute gradually becomes neutral in the vacuum,

(ii) uncharged solute is immersed into solvent, and

(iii) solute gains the normal values of the charges in solvent.

We call the free energy change in step [ii] as **nonpolar solvation energy** and the sum of the energies associated with the step [i] and [ii] is known as **charging or polar solvation free energy** and describes the solvent's effect on the solute charging process. By decomposing the solute-solvent potential, we can write for a solute in conformation **X**

$$U_{ps}(\mathbf{X},\mathbf{Y}) = U_{ps}^{np}(\mathbf{X},\mathbf{Y}) + U_{ps}^{elec}(\mathbf{X},\mathbf{Y}) \tag{3.11}$$

and the total PMF as

$$W(\mathbf{X}) = U_p(\mathbf{X}) + \Delta W^{np}(\mathbf{X}) + \Delta W^{elec}(\mathbf{X}) \tag{3.12}$$

The net solvation energy is a sum of nonpolar contribution and electrostatic component. In general, the polar and nonpolar solvation terms have opposing effect. The polar solvation favors the maximum solvent exposure for all polar groups in the solute, while nonpolar solvation favors compact structures with small areas and volumes. The nonpolar solvation contribution can be written as

$$e^{-\Delta W^{np}(\mathbf{X})/k_BT} = \frac{\int d\mathbf{Y} e^{-[U_{ss}(\mathbf{Y})+U_{ps}^{np}(\mathbf{X},\mathbf{Y})]/k_BT}}{\int d\mathbf{Y} e^{-U_{ss}(\mathbf{Y})/k_BT}} \tag{3.13}$$

and the electrostatic component can be expressed as

$$e^{-\Delta W^{elec}(\mathbf{X})/k_BT} = \frac{\int d\mathbf{Y} e^{-[U_{ss}(\mathbf{Y})+U_{ps}^{np}(\mathbf{X},\mathbf{Y})+U_{ps}^{elec}(\mathbf{X},\mathbf{Y})]/k_BT}}{\int d\mathbf{Y} e^{-[U_{ss}(\mathbf{Y})+U_{ps}^{np}(\mathbf{X},\mathbf{Y})]/k_BT}} \tag{3.14}$$

The potential energy can also be written as in terms of thermodynamic coupling constants λ_1 and λ_2 [47].

$$U(\mathbf{X},\mathbf{Y}:\lambda_1,\lambda_2) = U_p(\mathbf{X}) + U_{ss}(\mathbf{Y}) + U_{ps}^{np}(\mathbf{X},\mathbf{Y}:\lambda_1) + U_{ps}^{elec}(\mathbf{X},\mathbf{Y}:\lambda_2). \tag{3.15}$$

where I. $\lambda_1 = \lambda_2 = 0 \Longrightarrow$ non-interacting reference system, II. $\lambda_1 = \lambda_2 = 1 \Longrightarrow$ fully interacting system, and III. $\lambda_1 = 1$, $\lambda_2 = 0 \Longrightarrow$ no solute-solvent electrostatic interactions. From the perspective of thermodynamic integration (TI), we can express both the contributions for a solute at conformation **X**

$$\Delta W^{np}(\mathbf{X}) = \int_0^1 d\lambda_1 \langle \frac{\partial U^{(np)}}{\partial \lambda_1} \rangle_{(x,\lambda_1,\lambda_2=0)} \tag{3.16}$$

and

$$\Delta W^{elec}(\mathbf{X}) = \int_0^1 d\lambda_2 \langle \frac{\partial U^{(elec)}}{\partial \lambda_2} \rangle_{(x,\lambda_1=1,\lambda_2)} \tag{3.17}$$

It is noteworthy that free energy decomposition is path-dependent [105, 106]. For example, we need to first create the non-polar cavity into the solvent and then perform electrostatic charging of the solute. The reverse order will give diverging results. The decomposition of net solvation free energy in this way is very helpful to understand the role of different microscopic factors in solvation.

3.3.2 Polar Solvation

The Poisson-Boltzmann (PB) equation is the most popular choice for describing the continuum electrostatics for biomolecular system. We can derive the PB equation in several way, but we will describe here the method which starts from the Poisson's equation.

$$-\vec{\nabla}.\left[\varepsilon(\vec{x})\vec{\nabla}\phi(\vec{x})\right] = \rho(\vec{x}) \tag{3.18}$$

for $\mathbf{x} \in \Omega$ and $\phi(\mathbf{x}) = \phi_{\mathbf{0}}(\mathbf{x})$ for $\mathbf{x} \in \delta\Omega$. Here $\phi(\vec{x})$ is the potential at position $\vec{x}$ due to a charge distribution $\rho(\vec{x})$ and the position dependent dielectric constant of the medium is $\varepsilon(\vec{x})$. Now $\rho(\vec{x})$ = $\rho_f(\vec{x})$ + $\rho_m(\vec{x})$ where $\rho_f(\vec{x})$ represents the solute charge distribution and $\rho_m(\vec{x})$ is the aqueous mobile ions distribution. The solute charge distribution is a summation of a set of delta functions and is given by the following equation:

$$\rho_f(\vec{x}) = \sum_i Q_i \delta(\vec{x} - \vec{x}_i) \tag{3.19}$$

where Q_i is the solute atom's charge and $\vec{x}_i$ is the solute atom's position. If we neglect the explicit interactions between aqueous ions, the mobile charges can be modeled as a continuous "charge cloud" described by the Boltzmann distribution. For m ion species with charges q_j, bulk concentrations c_j and steric potential $V_j(\vec{x})$, the mobile ion charge distribution is written as

$$\rho_m(\vec{x}) = \sum_j^m c_j q_j exp[-q_j \phi(\vec{x})/k_B T - V_j(\vec{x})/k_B T] \tag{3.20}$$

where k_B is the Boltzmann's constant and T is absolute temperature. Therefore, we can write

$$-\vec{\nabla}.\left[\varepsilon(\vec{x})\vec{\nabla}\phi(\vec{x})\right] = \sum_{i=1}^N Q_i \delta(\vec{x}-\vec{x_i}) + \sum_{j=1}^m c_j q_j exp[-q_j \phi(\vec{x})/k_B T - V_j(\vec{x})/k_B T] \tag{3.21}$$

Now if we expand the term $exp[-q_j \phi(\vec{x})/k_B T]$ in Taylor series and retain only the first term, and assuming V_j = V for all j, then we will get the Linearized Poisson-Boltzmann (LPB) equation:

$$-\vec{\nabla}.\left[\varepsilon(\vec{x})\vec{\nabla}\phi(\vec{x})\right] + \varepsilon(\vec{x})\kappa^2(\vec{x})\phi(\vec{x}) = \sum_{i=1}^N Q_i \delta(\vec{x}-\vec{x_i}) \tag{3.22}$$

where

$$\kappa^2(\vec{x}) = exp[-\beta V(\vec{x})].2I\beta e_c^2/\varepsilon(\vec{x}) \tag{3.23}$$

where $\beta = 1/k_B T$ and

$$I = \frac{1}{2}\sum_j^m c_j q_j^2/e_c^2 \tag{3.24}$$

is the ionic strength and e_c is the unit electric charge. Now by solving this equation, we will get the electrostatic potential for the entire system. Once we have access to the electrostatic potential, we can calculate the electrostatic free energy by a variety of integral formulations. For the LPB equation, the electrostatic free energy is

$$\Delta W_{elec} = \frac{1}{2}\sum_{i=1}^{N} Q_i \phi(\vec{x}_i) = \frac{1}{2}\int \rho_f \phi(\vec{x}) d\vec{x} \tag{3.25}$$

The Poisson-Boltzmann equation is solved numerically, since the analytical solution is not available for biomolecules with realistic shape and charge distributions. For the Nonlinear Poisson-Boltzmann (NPB) equation, the electrostatic free energy is given by

$$G[\phi] = \int \left[\rho_f \phi - \frac{\varepsilon}{8\pi}(\nabla\phi(\vec{x}))^2 - 2\kappa T \bar{n} e^{-\beta V}(\cosh(\frac{e_c \phi}{\kappa T}) - 1) \right] d\vec{x} \tag{3.26}$$

For small $\phi(\vec{x})$, this equation will give the free energy due to LPB equation:

$$G[\phi] = \int \left[\rho_f \phi - \frac{\varepsilon}{8\pi}(\nabla\phi(\vec{x}))^2 - \frac{\bar{\kappa}^2}{2}\phi(\vec{x})^2 \right] d\vec{x} \tag{3.27}$$

Now if we differentiate the free energy expressions in Eq. 3.27 with respect to atomic displacements, we will get the expressions for electrostatic forces [107, 108]. It is known from the saddle-point approximation made in deriving the PB equation that $\delta G[\phi]/\delta\phi =$

0. Therefore, the force on atom *i* while we consider the nonlinear PB equation, is

$$\vec{F}_i[\phi] = -\int \left[\phi(\frac{\partial \rho_f}{\partial \vec{y}_i}) - \frac{(\nabla\phi)^2}{8\pi}(\frac{\partial \varepsilon}{\partial \vec{y}_i}) + 2\bar{n}e^{-\beta V}(\cosh(\beta e_c \phi) - 1)(\frac{\partial V}{\partial \vec{y}_i}) \right] d\vec{x} \tag{3.28}$$

and if we consider the linearized PB equation, then the force on atom *i* is

$$\vec{F}_i[\phi] = -\int \left[\phi(\frac{\partial \rho_f}{\partial \vec{y}_i}) - \frac{(\nabla\phi)^2}{8\pi}(\frac{\partial \varepsilon}{\partial \vec{y}_i}) - \frac{\phi^2}{2}(\frac{\partial \bar{\kappa}^2}{\partial \vec{y}_i}) \right] d\vec{x}. \tag{3.29}$$

It is to be noted that both the nonlinear PB and linearized PB equations are approximations. We can not apply this method blindly to the biomolecular systems. Care should be taken for highly charged systems. The Poisson-Boltzmann equation is based upon the mean field approximation (**MFA**) of the counter-ion in which we neglect the counter-ion correlations and fluctuations. But at high ion concentration and valencies, ion correlations and fluctuations become important factors. We should also keep in mind that the PB equation is based on the assumption of local and linear polarization of the solvent with respect to the applied field which can be broken down under high electric fields or in highly-ordered systems of water [47]. In a nutshell, PB equations and other implicit models work best for describing the electrostatic effects on biomolecules with low linear charge density in solutions of monovalent ions at low concentration [47].

There are couple of software available that treat the Poisson-Boltzmann equation efficiently. Examples include APBS [52], Delphi [109], GRASP [110], MEAD [111], ZAP [112], UHBD [113], MacroDox, Jaguar [114, 115], CHARMM [116] and AMBER [18].

3.3.3 Generalized Born Model

Generalized Born Model or GB Model is a very popular implicit solvent model. This model is widely used in molecular dynamics simulation. This model is very efficient and reliable. AMBER [18] uses GB model [117].

For a simple spherical ion of radius R_{ion} and charge Q_{ion}, the electrostatic component of the solvation free energy is given by the well-known Born formula [86]

$$\Delta W^{elec} = \Delta G_{pol} = -\frac{Q_{ion}^2}{2R_{ion}}\left(1 - \frac{1}{\varepsilon_s}\right) \tag{3.30}$$

where ε_s is the dielectric constant of the solvent. Now let us consider a molecule consisting of charges $Q_1 ... Q_N$ embedded in spheres of radii, $a_1 .... a_N$ and we also assume that the separation r_{ij} between any two spheres is sufficiently large in comparison to the radii. Then the electrostatic component of the solvation free energy is given by

$$\Delta G_{pol} \simeq \sum_i^N -\frac{Q_i^2}{2a_i}\left(1 - \frac{1}{\varepsilon_s}\right) + \frac{1}{2}\sum_i^N \sum_{i \neq j}^N \frac{Q_i Q_j}{r_{ij}}\left(\frac{1}{\varepsilon_s} - 1\right) \tag{3.31}$$

where the first term corresponds to the sum of individual Born terms and the second term corresponds to pairwise Coulombic terms [47]. Coulombic interactions are rescaled by a pre-factor $\left(\frac{1}{\varepsilon_s} - 1\right)$ because of change of dielectric constant upon going from the vacuum to solvent.

The general goal of the GB model is to get a closed form semi-analytical formula that will mimic equation 3.31 and capture the essential physics of the Poisson equation for a realistic protein geometries. The GB theory says that

$$\Delta G_{pol} \simeq -\left(1-\frac{1}{\varepsilon_s}\right)\frac{1}{2}\sum_{ij}\frac{Q_iQ_j}{f_{ij}^{GB}}. \tag{3.32}$$

For $i = j$, f^{GB} can be thought of as "effective Born radii," and for off-diagonal terms, it could be thought of as an effective interaction distance. The most common formula for f_{ij}^{GB} is given by Still et. al [118].

$$f_{ij}^{GB}(r_{ij}) = \left[r_{ij}^2 + R_iR_je^{\left(-\frac{r_{ij}^2}{4R_iR_j}\right)}\right]^{\frac{1}{2}} \tag{3.33}$$

Here R_i are the effective Born radii of the atoms, which depends on the radius (a_i) of the atom i and also is influenced by radii and relative position of all other atoms. An effective way of calculating the approximated Born radii rapidly is now needed. In terms of electric displacement vector $\mathbf{D}$, we can write

$$G_{pol} = \frac{1}{2}\int_{\Omega}\rho_f(\mathbf{x})\phi(\mathbf{x})d\mathbf{x} = \frac{1}{8\pi}\int_{\Omega}\mathbf{E}.\mathbf{D}d\mathbf{x} \tag{3.34}$$

Now using the Coulomb Field Approximation ($\mathbf{CFA}$), we can write the displacement due to the charge of atom i is,

$$\mathbf{D_i} \approx \frac{Q_i\mathbf{r}}{r^3}. \tag{3.35}$$

$$G_i = \frac{1}{8\pi}\int (\mathbf{D}/\varepsilon).\mathbf{D}d\mathbf{x} \approx \frac{1}{8\pi}\int_{interior}\frac{Q_i^2}{r^4\varepsilon_p}d\mathbf{x} + \frac{1}{8\pi}\int_{exterior}\frac{Q_i^2}{r^4\varepsilon_s}d\mathbf{x}. \tag{3.36}$$

The electrostatic component of the solvation free energy is

$$\Delta G_{pol,i} = \frac{1}{8\pi}\left(\frac{1}{\varepsilon_s} - 1\right)\int_{exterior}\frac{Q_i}{r^4}d\mathbf{x} \tag{3.37}$$

Now comparing this equation with the Born formula, one can write

$$R_i^{-1} = \frac{1}{4\pi}\int_{exterior}\frac{1}{r^4}d\mathbf{x} \tag{3.38}$$

This could also be written as

$$R_i^{-1} = a_i^{-1} - \frac{1}{4\pi}\int_{interior, r>a_i}\frac{1}{r^4}d\mathbf{x}. \tag{3.39}$$

For monatomic ion, $R_i = a_i$ and the Born formula is restored exactly. If we consider the molecule is composed of a set of non-overlapping spheres of radius a_j at positions r_{ij} relative to atom i, then the above equation can be rewritten as

$$R_i^{-1} = a_i^{-1} - \frac{1}{4\pi}\sum_j\int_{|\mathbf{r}-\mathbf{r_{ij}}|<a_j}\frac{1}{r^4}d\mathbf{x}. \tag{3.40}$$

The integrals over spheres can then be calculated analytically giving

$$R_i^{-1} = a_i^{-1} - \sum_j \frac{a_j}{2\left(r_{ij}^2 - a_j^2\right)} - \frac{1}{4r_{ij}}\log\frac{r_{ij} - a_j}{r_{ij} + a_j} \tag{3.41}$$

Hawkins et al. [119] have proposed a formula to calculate the effective Born radii R_i, as

$$R_i^{-1} = a_i^{-1} - \sum_j H\left(r_{ij}, S_j a_j\right), \tag{3.42}$$

where H is a complex function.

Several variations of the GB model are also available now of which GB/SA (Generalized Born/Surface Area) and GB/MV (Generalized Born/Molecular Volume) [120, 121] are most frequently used in molecular dynamics simulations. AMBER [18] uses both variants. GB/SA and GB/MV are created to take into account both polar and non-polar effects in solvation free energies. GB/SA is the most popular choice for biomolecular simulation in implicit solvent.

3.3.4 Non-polar Solvation

When we insert a solute in a solvent, we need to create a cavity in the shape of the solute to accommodate the solute protein. The energy associated with this creation of a cavity is known as the **cavitation energy**. The attractive van dan Waals solute-solvent interaction gives rise to the dispersion term. These are the two non-polar terms we should consider in implicit model. In the context of non-polar molecules, these two terms become dominant in solvation free energies. A common way to treat these non-polar terms is to introduce a Solvent Accessible Surface Area (SASA)-term. But this SASA approach is a subject of debate. It can capture neither cavitation nor dispersion terms. We need to treat both terms individually. The cavitation term can be best treated by scaled particle theory (SPT) [60, 61, 122]. SPT is a statistical mechanics based approach proposed by Reiss, Stillinger and Pierotti to estimate the free energy associated with inserting a non-polar repulsive sphere into a solvent [47]. The radius of the repulsive sphere is scaled in this approach and hence, the name SPT [47].

We can estimate the required reversible work $W(R)$ to create a spherical cavity for a hard-sphere liquid of bulk density $\bar{\rho}$ as

$$W(R) = -k_B T ln \left(1 - \frac{4}{3}\pi R^3 \bar{\rho}\right) \tag{3.43}$$

for $2R \leq a$ and a = 2.75 Å for a non-polar solute in liquid water [122, 47]. According to Tolman [123], in the limit of a large cavity, eqn. 3.43 takes the form

$$W(R) = \frac{4}{3}\pi R^3 p + 4\pi R^2 \gamma_v \left(1 - \frac{4\delta}{R}\right) + ... \tag{3.44}$$

where p is the isotropic pressure, γ_v is the surface tension of the solvent and δ is a molecular length scale. For water, the value of δ is approximately 0.5 Å[122]. The value of γ_v could be obtained from the experiments.

From the concept of Scaled Particle Theory (SPT), we can easily relate the non-polar free energy contribution to the solvent-exposed surface area. From eqn. 3.44, we can see that the microscopic surface tension coefficient depends on the radius of curvature, i.e.,

$$\gamma(R) = \gamma_v \left(1 - \frac{4\delta}{R}\right) \tag{3.45}$$

The length scale δ is chosen such a way that the curvature dependency becomes prominent only if the radius R is very small [47]. Ignoring the curvature effect, one can write

$$\Delta W^{np}(\mathbf{X}) = \gamma_{\mathbf{v}} \mathbf{A}_{\mathbf{tot}}(\mathbf{X}) \tag{3.46}$$

Such a treatment of the non-polar contribution to the solvation free energy is very popular in biophysical applications and has been used extensively [124, 125, 105, 126, 127, 128] because of its simplicity. There are several models which are just a slight variation of the solvent-exposed area model. Among those models are the shell model of Scheraga [129, 130], the solvent excluded-volume model of Colonna and Sander [131, 132] and the Gaussian model of Lazaridis and Karplus [133]. Computation of accurate molecular surface and its analytical derivative with respect to atomic position is computationally demanding. Motivated by this, Janin and Wodak [134] developed an approximate expression for the molecular surface area in macromolecules and this has been parameterized by Fraternali et.al. for MD simulations of proteins [128].

3.4 Our Model

To treat the solvent implicitly we decompose the net solvation free energy into three terms following the quantum mechanical Polarizable Continuum Model (PCM)[46]

$$\Delta G_{net} = \Delta G_{pol} + \Delta G_{disp} + \Delta G_{cav} \tag{3.47}$$

We solve the Poisson-Boltzmann (PB) equation to compute the polarization term. The dispersion term is computed using the Caillet-Claverie [57] formula. For the cavitation term, we adopt rPA [61] (revised Pierotti Approach) formalism. We call our model Enhanced Solvation Model since our approach adopts polar as well as non-polar terms also. Many of the continuum models contain only the polarization term. This model is developed within the Boundary Element Method [54] (BEM) framework. For my doctoral studies, I concentrated on the first two terms– ΔG_{pol} and ΔG_{disp}. My contributions to this field are described in next three chapters.

Chapter 4

Boundary Composition in Poisson Boltzmann Calculations

This chapter is reproduced from our original paper - P. Kar, Y. Wei, U. H. E. Hansmann, S. Höfinger. 2007. Systematic Study of the Boundary Composition in Poisson Boltzmann Calculations, J. Comp. Chem., 28, 2538-2544. Copyright Wiley (2007).

4.1 Introduction

A common way of describing solvation effects to biomolecular structure is to treat the solvent as a continuum of characteristic dielectric constant. The biomolecule of interest, i.e. a protein, DNA, RNA, glycolipid, etc. is considered in full atomic detail, while the surrounding medium is represented as structureless continuum interacting primarily via polarization, dispersion, repulsion and cavitation effects [44, 45, 46, 47, 48]. The underlying physics concerned with polarization is then often expressed in terms of solutions to the Poisson-Boltzmann equation (PB) [49, 50, 51, 52, 53, 67, 54, 55]. Approximations to the PB — motivated by simplified computational protocols — are standard practice e.g. the Generalized Born model (GB) [118, 117]. However, PB and GB are dealing with the polarization term only, and the other above mentioned interactions are usually treated by either first-principle [135] or semi-empirical [66] character.

Solutions to the PB are computed either by the finite difference method (FDPB) [49, 50, 51, 52] or by the boundary element method (PB/BEM) [54, 55]. The latter is particularly intriguing since it reduces a three-dimensional integral over the entire volume to a two-dimensional surface integral, leading to considerable savings in computational time. Both approaches depend fundamentally on the exact definition of the boundary between solute and solvent. All definitions are based on the area of the atoms exposed to the solvent, for instance the solvent accessible surface area (SASA), the solvent excluded volume, or the molecular surface [9], which all depend on a chosen set of van der Waals (vdW) radii [136, 137, 138, 139] assigned to the center of the atoms.

Given the dependence on the exact geometry and quality of the boundary it appears necessary to study the geometric factors that influence the outcome of PB calculations in greater detail. This is particularly appropriate for semi-quantitative approaches [140] where the demand on accuracy is a very sensitive issue [141]. Particularly we draw our attention to the following factors such as i) surface type and surface resolution, ii) dependence on atomic model parameters, i.e. van der Waals radii, iii) generality and physicochemical significance.

In this present work we provide such an analysis by focusing on each of these three points separately. At first, we employ different surface generation algorithms to a subset of randomly chosen protein structures of variable size and shape. PB/BEM calculations are carried out with increasing resolution of the boundary. Optimal surface resolution and surface generation parameters that guarantee numerical convergence and methodic stability are derived. Next, we use these optimized parameters for a set of model peptides and vary the van der Waals radii in a systematic way. The reference set of model peptides is considered at a high level of quantum chemical theory, i.e. PCM [46] using the Becke-98 density functional [142] and the basis set of Sadlej [143]. The aim of this second step is to identify optimal van der Waals radii within the PB/BEM approach that will lead to boundaries and solute geometries of similar size and shape as those used in the high-level PCM calculations. Finally, with the optimized parameters determined in the initial two stages we compute actual PB/BEM polarization energies in order to obtain a close match with the quantum chemical results obtained from the reference set.

4.2 Methods

4.2.1 Sample Selection, Preparation and Set Up of Structures and Computation of Molecular Surfaces with Different Programs

A set of different protein structures is randomly selected from the Protein Data Bank [38]. The actual download site used is the repository PDB-REPRDB [144]. Default options are applied with the following exceptions: i) **Number of residues less than 40 excluded – NO**, ii) **Include MUTANT – NO**, iii) **Exclude COMPLEX**, iv) **Exclude FRAGMENT**, v) **Include NMR – NO**, vi) **Include Membrane Proteins – NO**. A total of 28 structures of different protein sizes and shapes (see Table 4.1) are chosen. The PDB codes of the samples are, 2ERL, 1P9GA, 1FD3A, 1N13E, 1BRF, 1PARB, 1K6U, 1AVOA, 1SCMA, 1OTFA, 1DJTA, 1KU5, 1K3BC, 1R2M, 1CC8, 1L9LA, 1ZXTD, 1GYJA, 1T8K, 1XMK, 1YNRB, 1EZGA, 1C5E, 1SAU, 1WN2, 1JBE, 1C7K and 1WKR.

Two different programs to calculate molecular surfaces have been employed: the Connolly program **MSROLL** [9] and the **SIMS** program [10].

Downloaded PDB structures are cleaned from multichain entries, HETATM lines CONNECT lines, ANISOU lines, counter ions, water molecules and the footer section. Program MOLDEN [145] is used to visualize the downloaded PDB structures after cleaning and the force field **Tinker Amber** is selected before a new PDB file is written out from within

MOLDEN using option 'Write_With_Hydrogens'. Since MOLDEN always uses the default HIP type in AMBER jargon, HIS residues need to be converted to HIP types, as well as CYS residues engaged in disulfide bonds need to be converted to CYX-type residues. Occasional cases with PRO being the initial residue are manually edited and initial PROs removed. AMBER non-bonded parameters [146], i.e. charges and van der Waals radii are assigned to all the atoms in the protein structures. In this first part of the study, the vdW-radii are increased by a factor of 1.12 and atomic partial charges are scaled down by another factor of 0.9 [59].

The MSROLL program is used with varying choices of the fineness value (the **-f** command line argument) which defines the resolution of the surface. With smaller values the resolution of the surface becomes better but computational cost will increase. The probe radius (the **-p** command line argument) is set to 1.5 Å. Analytically calculated SASA and molecular volumes are recorded, and the data file containing triangulation details is translated into a human readable format, and critical items (for example almost coinciding triangles) removed. The SIMS program is used with identical arguments to those employed in MSROLL. Similarly, varying the resolution of the surface triangulation into small sized triangles means adjusting the **dot-density** parameter in SIMS. Higher values for this parameter will yield higher surface resolutions but also increase the computational demand. We record the number of BE, number of iterations, SASA and volume for comparison.

4.2.2 Computation of Polarization Free Energies, ΔG^{Pol}, Based on solutions to the Poisson Boltzmann Equation

Inner/outer dielectric constants at the molecular boundary are set to 1.0 and 80.0 respectively. The serial version of the PB/BEM program **POLCH** [147] is used. Critical cases with additional secondary cavities located in the interior part of the proteins are excluded. AMBER van der Waals radii and partial charges [146] are applied. Using our own tool chain for the assignment allows us to conveniently scale these data, as well as to write out in the same instance the corresponding parameter files required by the molecular surface programs.

The most prominent combinations of peptidic Φ, Ψ-angles [148] are used to construct different conformations of dipeptides. We only consider homodimers. All 20 types of different amino acids are used for this combinatorial approach. Zwitterionic forms are built and 9 conformations per class of amino acid are taken into account leading to all in all 180 structures. Program "protein.x" from the TINKER package version 4.2 is employed [70]. Each of these reference structures is subjected to Polarizable Continuum Model (PCM) [149] calculations at the Becke-98 [142] level of density functional theory (DFT) using the high-quality basis set of Sadlej [143] within the Gaussian-03 suite of programs [150]. Geometric properties, i.e. the molecular volume and the molecular surface area, as well as polarization free energies are extracted from the reference calculations and used as a base line when comparing to PB/BEM data. The computational demand of these reference calculations is significant. For example, WW-conformations require on the order of 6 weeks (and beyond) single-processor time on modern computing architectures.

4.3 Results

4.3.1 Stage I: Rather small-sized BEs are Needed to Obtain Consistently Convergent Polarization Free Energies ΔG^{Pol}

We start with PB/BEMA calculations for a set of protein structures (PDB codes summarized in Table 4.1). The boundary discretization is achieved with two independent programs, MSROLL [9] and SIMS [10]. Boundary resolution into BEs is steadily increased with either program and independent PB/BEM results are computed for each particular boundary decomposition. A typical plot of the trend of ΔG^{Pol} as a function of number of BEs is shown in Figure 4.1 for the protein structure with PDB code 1C5E. Both approaches converge to identical results in the limit of large numbers of BEs. The importance of well-resolved boundaries becomes clear from Figure 4.1. Errors on the order of $\pm 40\frac{kcal}{mol}$ are easily introduced when working in the non-converged domain. Connolly's MSROLL program (red triangles in Figure 4.1) reaches a plateau value in a continuous manner, while the SIMS program (blue spheres in Figure 4.1) finds its limit value within an alternating

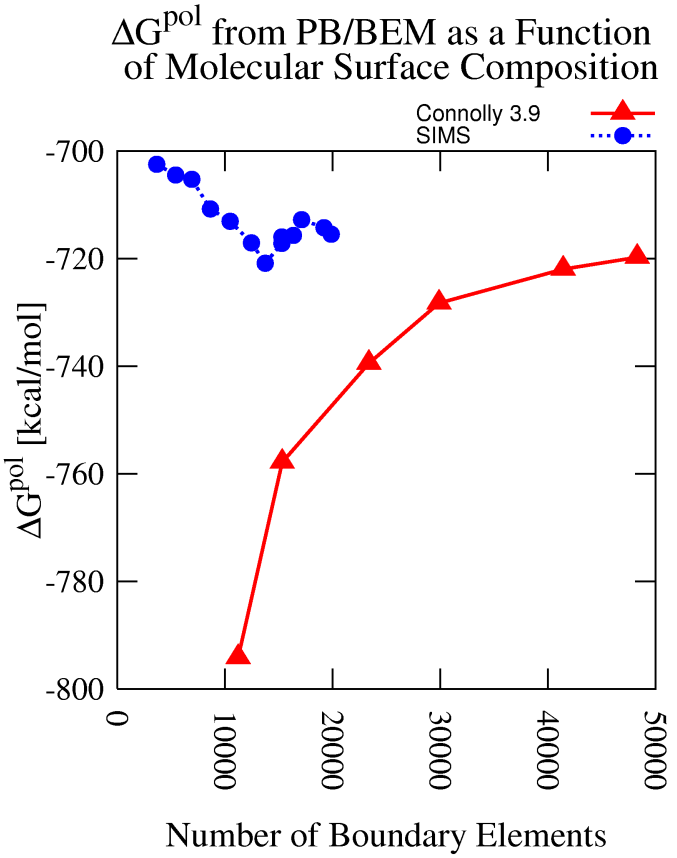


Figure 4.1: PB/BEM derived ΔG^{Pol} as a function of BE obtained from two independent programs MSROLL [9] and SIMS [10]. The example represents results for PDB structure 1C5E [11].

sequence. The SIMS program reaches convergence much faster than the Connolly program. The quality of the computed molecular boundaries is comparable, see, for instance, the values of molecular surfaces and volumes (final two columns in Table 4.1) obtained with either program. SIMS seems to overestimate the volume by a small margin of roughly 1%. The recommended average size of BEs for converged results using MSROLL is on the order of 0.11 Å^2 while SIMS would require an average size of 0.31 Å^2. Both numbers are close to the value of 0.4 Å^2 advocated in Quantum Chemistry [151].

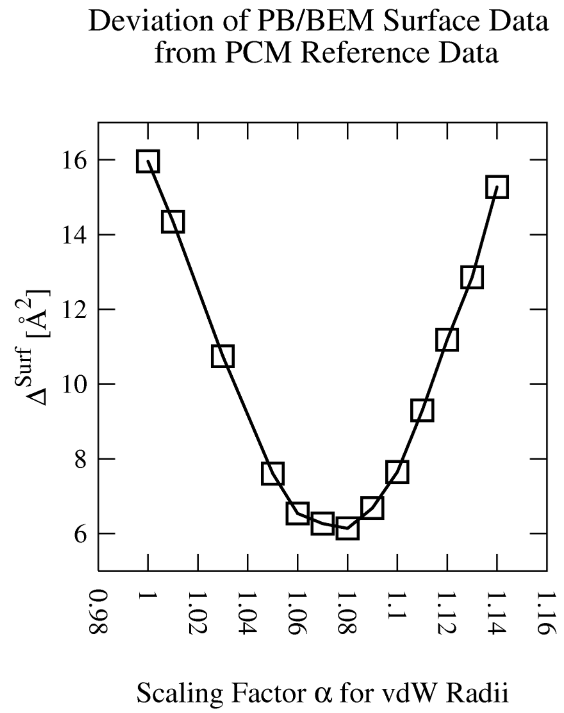


Figure 4.2: Comparison of employed molecular surfaces in the PB/BEM series based on scaling the AMBER default vdW radii by a factor α to the reference data obtained from PCM calculations [11].

4.3.2 Stage II: Systematic Geometric Comparison to High Level Quantum Chemistry Calculations Suggests a Uniform Scaling of AMBER van der Waals Radii by a Factor of 1.07

A reference set of dipeptides in different conformations (9 per species) is constructed. Only homodipeptides comprising all 20 types of naturally occurring amino acids are considered. Thus a total number of 180 dipeptidic reference structures is set up. The zwitterionic form is used throughout. Each of these structures is computed at the Becke-98 level of theory [142] using the basis set of Sadlej [143] and the PCM model [149] for solvation free en-

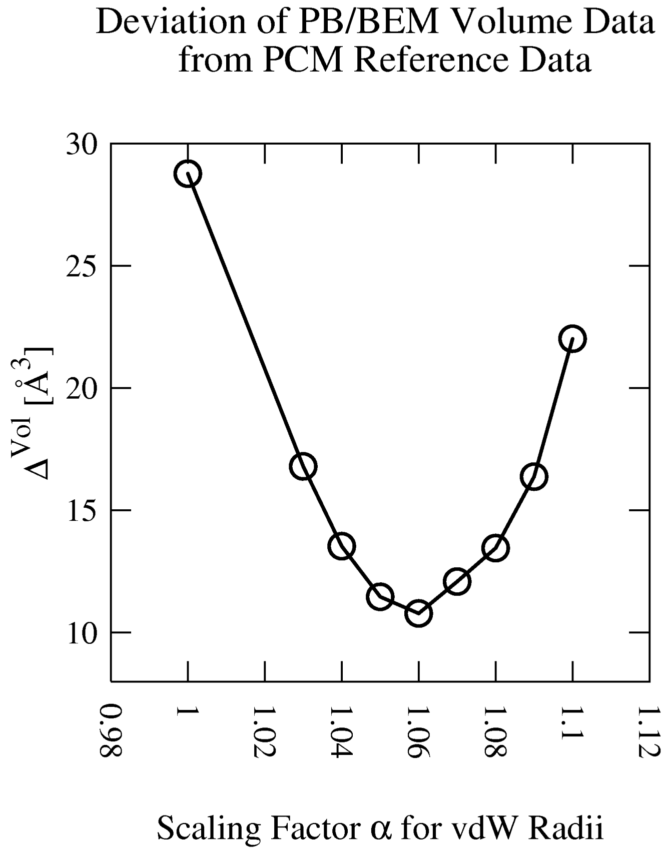


Figure 4.3: Comparison of employed molecular volumes in the PB/BEM series based on scaling the AMBER default vdW radii by a factor α to the reference data obtained from PCM calculations [11].

ergies. Geometric properties such as the cavity volume and the cavity surface area are extracted from each of the reference calculations. All 180 structures are also computed within the PB/BEM approach using optimized parameters for the boundary resolution determined in Stage I of this study. However, only the SIMS program is used. We define a global deviation from the reference data by

$$\Delta^{Surf} = \frac{1}{20}\sum_{i=1}^{20}\frac{1}{9}\sum_{j=1}^{9}\sqrt{(Surf_{i,j}^{PCM} - Surf_{i,j,\alpha}^{PB/BEM})^2} \tag{4.1}$$

where j runs over the conformations and i over the different types of homodipeptides, i.e. GG, AA, VV, etc. The parameter α refers to a specific scaling factor used when constructing the boundaries within the PB/BEM approach. In particular this scaling makes the van der Waals radii larger or smaller by a certain fraction. The AMBER default set of van der Waals radii is used [146]. A similar criterion is used for comparing molecular volumes,

$$\Delta^{Vol} = \frac{1}{20}\sum_{i=1}^{20}\frac{1}{9}\sum_{j=1}^{9}\sqrt{(Vol_{i,j}^{PCM} - Vol_{i,j,\alpha}^{PB/BEM})^2} \tag{4.2}$$

and the dependence on the scaling factor α is shown in Figures 4.2 and 4.3.

As becomes clear from Figures 4.2 and 4.3 the best match to the reference data is obtained when scaling the AMBER van der Waals radii by a factor of 1.07. Detailed data with respect to conformational averages per type of dipeptide are shown in Table 4.2 and Table 4.3.

4.3.3 Stage III: Charge Scaling is Not Required

Using the optimized parameters obtained in the previous two stages leads us to the final step of directly comparing polarization free energies ΔG^{Pol} computed within the PB/BEM approximation and at the PCM level of theory. The idea is to identify another uniform scaling factor β which applied to the AMBER default charges would result in an optimal match to the reference polarization free energies. Thus another deviation criterion is introduced,

$$\Delta^{\Delta G^{Pol}} = \frac{1}{20}\sum_{i=1}^{20}\frac{1}{9}\sum_{j=1}^{9}\sqrt{(\Delta G_{i,j}^{Pol,PCM} - \Delta G_{i,j,\beta}^{Pol,PB/BEM})^2} \tag{4.3}$$

that allows to identify the optimal value of β. The dependence of the PB/BEM polarization free energies on the charge scaling factor β is shown in Figure 4.4.

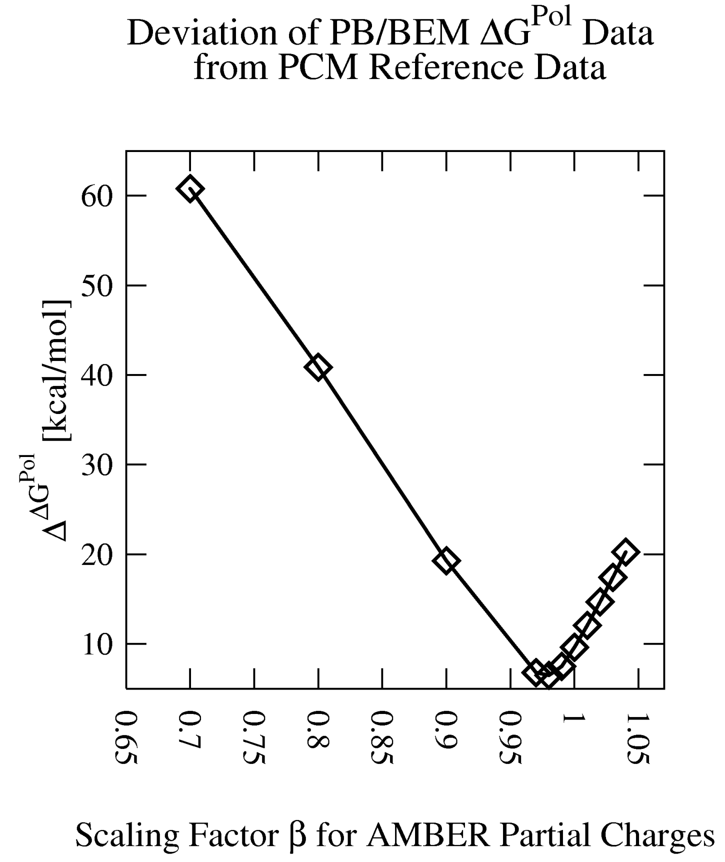


Figure 4.4: Comparison of PB/BEM polarization free energies ΔG^{Pol} based on scaling the AMBER default charges by a factor β to the reference data obtained from PCM calculations [11]

The trend shown in Figure 4.4 suggests an optimal value of β very close to 1.0, hence no charge scaling is required. This result i) emphasizes the broad applicability of AMBER partial charges and ii) circumvents conceptual difficulties that would arise when charges had to be scaled, i.e. modified net charges in proteins, non-neutral forms, etc. A detailed analysis with respect to the magnitude of the average deviation of each particular type of dipeptide studied is shown in Table 4.4.

4.4 Discussion

Motivated by recent high-performance solution to Poisson Boltzmann calculations [147] we have tested the influence of the many critical parameters involved. One obvious issue is the exact choice and composition of the boundary between solute and solvent. At first, we have to ensure the numerical stability within the selected level of approximation. In order to address this problem we have carried out PB/BEM calculations on a large sample of different proteins. When using different programs to create the boundary surface and increasing systematically the resolution of these surfaces into small-sized boundary elements, a recommended threshold size of about 0.31 Å^2 for the average BE is identified when using program SIMS [10] which showed faster convergence than the well-known Connolly program [9]. Although giving rise to very fine-resolved boundary surfaces, hence large numbers of BEs, this value is close to the corresponding value of 0.4 Å^2 frequently advised in Quantum Chemical models [151]. As a consequence, even proteins of modest size thus require consideration of vast numbers of BEs (see for example Table 4.1), and the importance of efficient means of solving the computational problem is underlined again. After having established the necessary degree of boundary partitioning in the first stage, we performed a systematic comparison against a reference set of dipeptides computed at a high level of Quantum Chemical theory. Consideration of geometric factors revealed that when applying a scaling factor of about 1.07 to AMBER default van der Waals radii, rather good agreement can be reached between the reference geometries and the geometries in the PB/BEM approach. The recommended value of 1.07 is somewhat smaller than a factor found previously by Höfinger et.al. (1.12 of ref [59]) and reflects the much finer resolved boundary surfaces used in this present work.

The final step was to compare actual calculations of the polarization free energies to each other. Following previous attempts, we wanted to derive another scaling factor that, when applied to AMBER partial charges, would yield a close match to the reference polarization free energies. The trend visible in Figure 4.4 indicates that no scaling of the charges is necessary: they are already close to optimal. This is an unexpected — but very welcome — result, as it eliminates potential secondary problems that would emerge with modifying charges. Again, this is another consequence of the much finer resolved boundary surfaces in this present work as opposed to previous results by Höfinger [59] where a scaling factor of 0.9 had been found.

4.5 Conclusion

Combined employment of small-sized BEs (≈ 0.3 Å^2 on average), slightly increased AMBER van der Waals radii (by a factor of 1.07), and default AMBER partial charges leads to good quality estimates of the polarization free energy, ΔG^{Pol}, for proteins within the PB/BEM framework.

Table 4.1: PDB codes of studied structures and the number of BEs needed to reach converged PB/BEM results using molecular surface algorithms MSROLL [9] and SIMS [10] respectively [11].

PDB	No. of	No. of BEs	No. of BEs	Molecular	Molecular
	Residues	Using	Using	Surface Area	Volume
		MSROLL	SIMS	(Difference)	(Difference)
		[9]	[10]	[Å^{2}]	[Å^{3}]
2ERL	40	15661	9807	2370 (+1)	5653 (-43)
1P9GA	41	22302	5751	2091 (-5)	5055 (-72)
1FD3A	44	25865	6699	2408 (+7)	5819 (-56)
1N13E	52	18419	10353	3750 (+11)	6542 (-69)
1BRF	53	33879	11810	2796 (-7)	7734 (-77)
1PARB	53	42336	11006	3968 (-8)	8509 (-108)
1K6U	58	24220	13406	3195 (+12)	8603 (-65)
1AVOA	60	43916	13335	4777 (-3)	9325 (-186)
1SCMA	60	54464	14603	5131 (-13)	10601 (-179)
1OTFA	62	40128	10610	3767 (+6)	8942 (-86)
1DJTA	64	35828	9134	3331 (+26)	9422 (-43)
1KU5	66	46390	12208	4310 (+3)	10153 (-133)
1K3BC	69	61667	16297	5768 (+19)	18193 (-165)
1R2M	71	39316	13659	3244 (-1)	9596 (-74)
1CC8	73	27668	15091	3644 (+2)	11094 (-65)
1L9LA	74	20278	11636	4182 (+5)	11728 (-112)
1ZXTD	76	37335	11259	4089 (+8)	10809 (-93)
1GYJA	76	44770	13665	4885 (-3)	11464 (-118)
1T8K	77	35978	13846	3925 (+4)	11410 (-119)
1XMK	79	56033	16468	4294 (+9)	12288 (-98)
1YNRB	80	31529	12630	4417 (+16)	11911 (-157)
1EZGA	84	34628	9122	3258 (+3)	10103 (-95)
1C5E	95	48306	19880	4480 (0)	13285 (-110)
1SAU	115	47613	17765	5197 (+25)	17897 (-116)
1WN2	121	51325	21555	5614 (+15)	17836 (-118)
1JBE	128	58119	16729	5409 (+22)	18905 (-188)
1C7K	132	54104	16675	5389 (0)	18858 (-182)
1WKR	340	74167	55378	11008 (-41)	47105 (-299)

Table 4.2: Comparison of average molecular surfaces based on unscaled and scaled AMBER vdW radii with data from PCM calculations [11].

Dipeptide Type	Mean Surface AMBER Unscaled [Å^2]	Mean Surface AMBER Scaled [Å^2]	Mean Surface PCM Reference [Å^2]
AA	191.764 (5.205)	204.388 (3.697)	214.747 (4.955)
CC	214.592 (4.622)	229.274 (5.461)	232.910 (6.522)
DD	228.687 (6.032)	242.724 (5.972)	240.839 (6.776)
EE	273.402 (5.835)	287.557 (6.951)	283.025 (5.766)
GG	150.255 (4.460)	161.637 (4.936)	167.764 (3.249)
II	279.535 (10.274)	294.402 (11.331)	302.173 (12.233)
KK	314.712 (6.273)	332.593 (7.677)	340.545 (7.599)
LL	276.435 (10.012)	290.497 (12.440)	293.172 (10.465)
MM	301.384 (6.691)	318.033 (8.261)	329.815 (8.374)
NN	232.466 (6.072)	247.553 (7.325)	247.040 (7.310)
QQ	276.939 (5.957)	293.663 (7.531)	292.648 (6.582)
RR	354.636 (6.925)	377.310 (7.568)	380.408 (6.739)
SS	196.528 (4.982)	207.904 (4.610)	212.107 (4.695)
TT	224.181 (8.047)	238.570 (8.359)	239.112 (10.124)
VV	251.913 (8.574)	265.008 (8.296)	276.423 (8.140)
YY	340.272 (17.100)	356.042 (17.293)	346.378 (14.704)
FF	329.058 (17.123)	343.245 (17.402)	326.947 (15.489)
WW	355.790 (26.425)	377.209 (27.865)	361.573 (25.182)
HH	282.802 (13.007)	299.235 (12.829)	296.885 (11.461)
PP	224.999 (10.768)	237.525 (10.500)	233.118 (9.900)

Table 4.3: Comparison of average molecular volumes based on unscaled and scaled AMBER vdW radii with data from PCM calculations [11].

Dipeptide	Mean Volume	Mean Volume	Mean Volume
Type	AMBER Unscaled	AMBER Scaled	PCM Reference
	[Å^3]	[Å^3]	[Å^3]
AA	191.804 (4.553)	215.661 (3.355)	228.591 (2.885)
CC	223.713 (3.799)	252.135 (4.069)	255.902 (3.594)
DD	242.406 (4.242)	270.928 (5.037)	260.904 (4.723)
EE	296.037 (4.888)	327.303 (3.864)	311.645 (3.810)
FF	381.346 (5.875)	417.696 (7.992)	388.363 (6.223)
GG	136.315 (3.565)	154.314 (2.631)	164.287 (2.048)
HH	317.309 (3.605)	353.465 (5.228)	340.591 (5.324)
II	323.590 (6.689)	357.080 (7.564)	367.182 (6.520)
KK	343.720 (3.747)	382.391 (4.384)	388.641 (3.903)
LL	313.537 (5.063)	346.507 (6.450)	344.950 (7.458)
MM	325.073 (5.069)	363.563 (5.484)	377.789 (3.905)
NN	248.729 (4.628)	278.449 (4.968)	271.305 (5.415)
PP	242.861 (8.154)	269.561 (9.116)	263.128 (9.957)
QQ	301.279 (3.985)	336.468 (5.590)	325.660 (5.591)
RR	384.833 (4.089)	431.811 (4.444)	424.874 (3.559)
SS	198.981 (3.533)	221.590 (3.162)	225.193 (2.964)
TT	244.397 (5.673)	272.848 (7.051)	273.546 (6.533)
VV	282.296 (6.114)	311.383 (6.895)	330.378 (8.172)
WW	431.248 (14.910)	480.974 (17.118)	447.693 (15.019)
YY	393.622 (5.433)	433.845 (7.433)	407.695 (5.466)

Table 4.4: Comparison of average PB/BEM polarization free energies ΔG^{Pol} using AMBER default charges to corresponding data obtained from PCM calculations [11].

Dipeptide Type	Mean $\Delta G^{Pol,PB/BEM}$ AMBER Default Charges [kcal/mol]	Mean $\Delta G^{Pol,PCM}$ PCM Reference [kcal/mol]	Mean $\Delta\Delta G^{Pol}$ Deviation [kcal/mol]	Number of References
AA	-91.36 (8.56)	-83.89 (10.12)	7.47	9
CC	-115.11 (11.02)	-96.80 (12.83)	18.31	9
DD	-296.25 (17.08)	-285.27 (18.24)	10.98	9
EE	-266.54 (14.09)	-259.29 (13.76)	7.25	9
GG	-96.52 (10.09)	-89.41 (11.36)	7.11	9
II	-82.72 (7.49)	-75.97 (8.70)	6.77	9
KK	-249.63 (16.51)	-236.37 (19.64)	13.26	9
LL	-85.54 (7.20)	-64.51 (8.64)	21.03	9
MM	-88.82 (7.55)	-82.10 (9.42)	6.72	9
NN	-105.11 (8.19)	-101.80 (12.20)	4.25	9
QQ	-119.08 (10.88)	-115.33 (12.60)	3.89	9
RR	-235.39 (17.79)	-228.71 (21.45)	6.68	6
SS	-112.78 (13.38)	-105.47 (13.90)	7.32	9
TT	-106.87 (12.06)	-100.61 (12.88)	6.55	9
VV	-85.17 (7.44)	-77.46 (8.73)	7.70	9
YY	-93.35 (4.36)	-90.11 (8.48)	3.59	5
FF	-89.92 (10.55)	-82.51 (13.94)	7.41	6
HH	-237.74 (19.08)	-236.06 (22.15)	3.66	9
PP	-79.15 (5.73)	-82.71 (7.50)	3.56	9
WW	-100.50 (4.27)	-88.11 (12.93)	12.39	2

Chapter 5

Implementation and Analysis of Dispersion Term

This chapter is reproduced in part from our paper- P. Kar, M. Seel, U. H. E. Hansmann, S. Höfinger. 2007. Dispersion Terms and Analysis of Size- and Charge- Dependence in an Enhanced Poisson-Boltzmann Approach, J. Phys. Chem. B, 111, 8910-8918. Copyright American Chemical Society (2007).

5.1 Introduction

The stabilizing effect of water on biomolecules is an intensively studied area in contemporary biophysical research. This is because many of the key principles governing biological functionality result from the action of the solvent, and thus water is often regarded as the “matrix of life”.

In theoretical work, the important factor “solvent” needs to be taken into account too, or the studied system will be unphysical. There are two main ways of solvent treatment in biophysical research. One is to embed the biomolecule of interest into a box of explicit solvent molecules resolved into full atomic detail [66, 93, 152]. The alternative form is to consider the solvent as a structureless continuum and describe the response of the environment with implicit solvation methods [46, 48, 47, 153, 45]. Basics of both approaches are discussed in Chapter 2. Both approaches have their own merits and demerits. While explicit solvent approaches are much more accurate than the implicit solvent approach but implicit solvent requires less computational power compared to explicit models. Much effort has been devoted to describing the electrostatic component within implicit solvation models. Efficient solutions have become popular in the form of *Generalized Born* (GB) models [118, 117, 154, 155] as well as *Poisson-Boltzmann* models (PB) [49, 50, 51, 52, 53, 67, 54, 55]. Solutions to the PB are computed either by the finite difference method (FDPB) [49, 50, 51, 52] or by the boundary element method (PB/BEM) [54, 55]. Considerable computational savings are expected from the latter because the problem can be reduced from having to solve a volume integral in FDPB to solving a surface integral in PB/BEM. Either approach is sensitive to the degree of discretization into grid elements or boundary elements [156, 11, 157].

Aside from the electrostatic component there are also apolar contributions to consider [46, 158]. Especially in the context of nonpolar molecules, such factors often become the dominant terms in the solvation free energy. A common way to treat these nonpolar contributions is to introduce a SASA-term, which means measuring the solvent accessible surface area (SASA) and weighing it with an empirically determined factor. Although commonly employed, this procedure has become the subject of intensive debates [159, 160, 161, 162]. Not only were SASA terms found to be inappropriate for representing the cavitation term [59, 162], but also is the weighing factor — usually associated with surface tension — completely ill-defined in an atomic scale context [56]. While the short range character of dispersion and repulsion forces occurring at the boundary between solute and solvent would imply that SASA can describe these kinds of interactions, a recent careful analysis has shown that, at least for dispersion, such a relationship is not justified [159].

The discrepancy arising with SASA-terms has been recognized by many groups and its persistent employment may be largely due to the sizeable cancellation of error effects. Waggoner and Baker [162] have divided the non-polar contributions into repulsive and attractive components and compared their approach to the mean forces obtained from simulation data on explicitly solvated systems. The specific role of the volume to account for repulsive interactions (cavitation) was clearly identified. Further inclusion of a dispersion term resulted in a satisfactory model of high predictive quality. Levy and Gallicchio devised a similar decomposition into a SASA-dependent cavitation term and a dispersion term within a GB scheme [163, 164, 165]. Their model makes use of atomic surface tensions and a rigorous definition of the molecular geometry within the GB framework. Particularly attractive is their efficient implementation and straightforward interfacing with Molecular Dynamics codes. Zacharias has already noted that a decompositon into a dispersion term and a SASA-based cavity term greatly benefits the quality of predictions of apolar solvation [158]. His approach uses distinct surface layers for either contribution. Hydration free energies of a series of tested alkanes agreed very well with data from explicit simulations [166] and from experiment. The striking feature in this approach is the improvement in hydration free energies of cyclic alkanes. Methodic advancement has recently been reported within the newest release of AMBER [18] where GB was augmented by a volume term [167] and the inclusion of dispersion terms was found to significantly improve the general predictive quality of PB. Of particular interest are systematic and physics-based decompositions that allow for separate consideration of each of the terms involved. In Quantum Mechanics (QM) such a technique has long been established with the Polarizable Continuum Model (PCM) [46]. It therefore seems advisable to use techniques like PCM (Polarizable Continuum Model) as a reference system whenever additional method development is performed, especially when regarding the multitude of technical dependencies continuum solvation models are faced with [156, 11].

In the present work we describe a systematic process to introduce dispersion terms in the context of the PB/BEM approach. The PCM model, that treats dispersion and repulsion terms from first-principles, is used as a reference system along with experimental data. Different ways of calculating dispersion-, repulsion contributions in PCM have recently been compared [135]. For our purposes the Caillet-Claverie method [57, 58] was implemented since it seems to offer a good compromise between accuracy and computational overhead. This method was also chosen in earlier versions of PCM [168] and thus represents a proven concept within the BEM framework. The fundamental role of dispersion and the potential danger of misinterpreting hydrophobicity related phenomena by ignoring it has been underlined recently [169, 170].

Given the fundmental nature of hydrophobicity and the potential role of dispersion within it, together with the current diversity seen in all the explanatory model concepts [171, 172], it seems to be necessary to advance all technical refinements to all solvation models (implicit as well as explicit) just to facilitate an eventual understanding of the factors governing these basic structure-forming principles.

After determination of appropriate dispersion constants used in the Caillet-Claverie approach, we apply our model to a series of proteins of increasing size. In this way we can analyze the relative contribution of the individual terms as a function of system size. Moreover, we have carried out semi-empirical calculations on the same series of proteins and can therefore compare effects resulting from different charge assignments to each other.

The semi-empirical program LocalSCF [19] also allowed for estimation of the polarization free energies according to the COSMO model [173], which could be readily used for direct comparison to PB/BEM data.

5.2 Methods

5.2.1 Theoretical Concepts

We use the following decomposition of the solvation free energy

$$\Delta G^{solv} = \Delta G^{pol} + \Delta G^{cav} + \Delta G^{disp} \tag{5.1}$$

where the individual terms represent polarization, cavitation and dispersion contributions. Explicit consideration of repulsion is not necessary as the cavitation term includes these interactions. PB/BEM methodology is used for ΔG^{pol} at the boundary specification described previously [174]. The cavitation term is expressed via the revised Pierotti approximation (rPA) [59, 56] (rPA), which is based on the Scaled Particle Theory [60, 61]. The major advantage with this revised approximation is a transformation property involving the solvent excluded volume. Hence after having identified the basic rPA-coefficients from free energy calculations the rPA-formula may be applied to any solute regardless of its particular shape or size [59]. ΔG^{disp} is computed from the Caillet-Claverie formula [57, 58] projected onto the boundary elements as suggested by Floris et al. [168]

$$\Delta G^{disp} = \sum_i^I \rho^{slv} \omega_i \sum_j^J \sum_k^K \underbrace{-0.214 \kappa_i \kappa_j \frac{64 (R_i^W)^3 (R_j^W)^3}{{R_{ij}}^6}}_{Caillet-Claverie} \frac{1}{3} \left(\vec{R}_{ij} \cdot \vec{n}_k \right) \Delta\sigma_k \tag{5.2}$$

where the first sum is over different atom types, i, composing one molecule of solvent, the second sum is over all solute atoms, j, and the sum over k is over all surface elements resulting from an expansion of the molecular surface by the dimension of radius R_i^W of a particular solvent atom, see Figure 5.1 for a graphical representation. Here solvent atoms are shown in grey and solute atoms are represented as white circles. The scheme corresponds to one particular choice of i. For example, if the solvent molecule in Figure 5.1 is

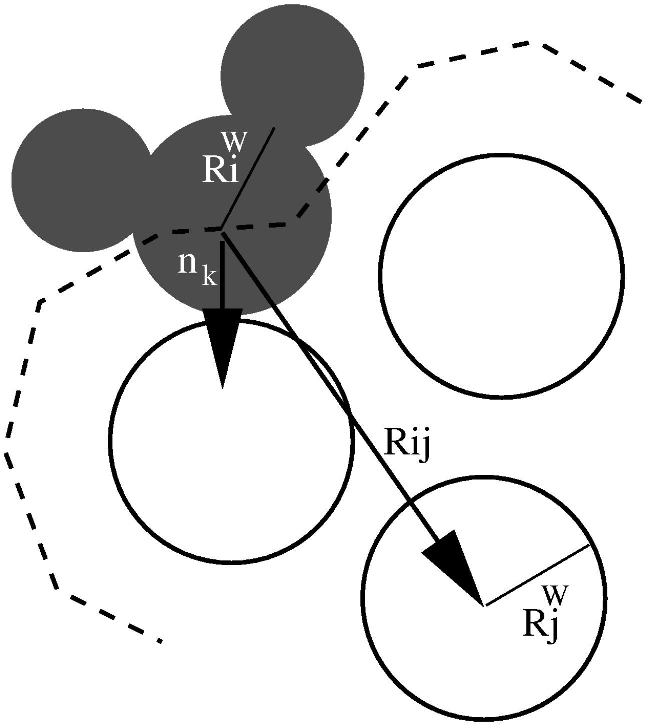


Figure 5.1: Graphical representation of the geometrical elements needed for computing the dispersion energy [12].

water, then the scenery depicts the first of two possibilities where i refers to the oxygen atom. After i is set, all atom radii of the solute are increased by the amount of the atomic radius of oxygen and the molecular surface (dashed line in Figure 5.1) is reconstructed. Next, the inner double sum is carried out where J is the total number of solute atoms and K is the total number of BEs forming the interface. Note that index j serves for a double purpose, looping over all solute atoms as well as defining the type of atomic radius to use. At every combination j,k of solute atoms with BEs, the expression emphasized by the curly bracket in eq. 5.2 must be evaluated. Here κ_i and κ_j are dispersion coefficients and R_i^W, R_j^W are atomic radii, all of them determined empirically by Caillet-Claverie [57, 58]. The corresponding values are summarized in Table 5.1.

Table 5.1: Summary of the data used for Caillet-Claverie style of dispersion treatment as outlined in eq. 5.2 [12].

Caillet-Claverie Dispersion Coefficients, κ, and Atomic Radii, R^W, in Å[57, 58]

κ_H	κ_C	κ_N	κ_O	κ_F	κ_{Na}	κ_P	κ_S	κ_{Cl}	κ_K	κ_{Br}	κ_J
1.00	1.00	1.18	1.36	1.50	1.40	2.10	2.40	2.10	2.90	2.40	3.20
R_H^W	R_C^W	R_N^W	R_O^W	R_F^W	R_{Na}^W	R_P^W	R_S^W	R_{Cl}^W	R_K^W	R_{Br}^W	R_J^W
1.20	1.70	1.60	1.50	1.45	1.20	1.85	1.80	1.76	1.46	1.85	1.96

R_{ij} is the distance between the center of some BE, k, and the center of a solute atom, j.

After the expression in the curly bracket of eq. 5.2 has been evaluated, it must be multiplied with a scalar product between the vectors $\vec{R}_{ij}$ and $\vec{n}_k$, the inwards pointing normal vector corresponding to the k^{th} BE. The remainder of eq. 5.2 is multiplication with a constant factor $\frac{1}{3}$ and multiplication with $\Delta\sigma_k$, the partial area of the BE, k. After all possible combinations j,k have been considered, the procedure is repeated with an incremented i, now referring to the H-atom, the second type of atom in a molecule of water. The molecular surface is recomputed, extended by the dimension of the atomic radius of hydrogen, and the entire inner double sum will be repeated as outlined for the case of oxygen. However, since both H-atoms in the solvent molecule are identical, this step needs to be done only once and ω_i, the number of occurrences of a particular atom type i, will take care of the rest (in the case of water $\omega_1 = 1$ for oxygen and $\omega_2 = 2$ for hydrogen). Finally, ρ^{slv} in eq. 5.2 represents the value of number density (ρ^{slv} = 0.033 for water at 298 K) of the solvent. We have restricted the approach to just the 6^{th}-order term in the expression derived by Floris et al. [168]. Note that we consider molecular surfaces as defined by Connolly [9]. The partial term listed after the curly bracket in eq. 5.2 is the actual consequence of mapping the classical pair interaction terms onto a boundary surface [168]. The partial expression enclosed in the curly bracket can be substituted with any other classic pair potential, for example using AMBER style of dispersion [18],

$$\Delta G^{disp} = \sum_{i}^{I} \rho^{slv} \omega_i \sum_{j}^{J} \sum_{k}^{K} \underbrace{-2\sqrt{\varepsilon_i \varepsilon_j} \left(\frac{R_i^W + R_j^W}{R_{ij}} \right)^6}_{AMBER} \frac{1}{3} \left(\vec{R}_{ij} \cdot \vec{n}_k \right) \Delta\sigma_k \tag{5.3}$$

with similar meanings of the variables used above and ε_i being the van der Waals well depth corresponding to homogeneous pair interaction of atoms of type i.

5.2.2 Model Calibration

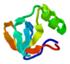
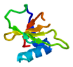
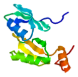
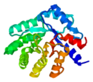
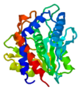
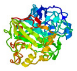
The algorithm covering computation of dispersion is implemented in the PB/BEM program POLCH [147] (serial version). Proper functionality was tested by comparing dispersion results of 4 sample molecules, methane, propane, iso-butane and methyl-indole, against results obtained from GAUSSIAN-98 [175] (PCM model of water at user defined geometries). Deviations were on the order of ± 1.8 % of the G98 value, so the procedure is assumed to work correctly. The small variations are the result of employing a different molecular surface program in PB/BEM [10]. Next, the structures of amino acid side-chain analogues are derived from standard AMBER pdb [38] geometries by making the C_α-atom a hydrogen atom, adjusting the C-H bond length and deleting the rest of the pdb structure except the actual side chain of interest. In a similar process, zwitterionic forms of each type of amino acid are constructed. PB/BEM calculations are carried out and net solvation free energies for solvent water are stored. A comparison is made against the experimental values listed in [13] as well as results obtained from the PCM model in GAUSSIAN-03 [150]. AMBER default charges and AMBER van der Waals radii increased by a multiplicative factor of 1.07 are used throughout [11]. Initial deviation from the reference set is successively improved by introducing a uniform scaling factor to the dispersion coefficients κ_i of eq. 5.2. The optimal choice of this dispersion scaling factor is identified from the minimum mean deviation against the reference data set. The initially derived optimal scaling factor is applied to the zwitterionic series, a subset of molecules for which experimental values have been compiled [14], and a set of 180 dipeptide conformations studied previously. When new molecules are parameterized, we use ANTECHAMBER from AMBER-8 and RESP charges based on MP2/6-31g* grids of electrostatic potentials [146]. Molecular geometries are optimized in a two-step procedure, at first at B3LYP/3-21g* and then at MP2/6-31g* level of theory and only the final optimized structure becomes subject to the RESP calculation.

Extensions are pursued in two directions. First, the PB/BEM approach is used with solvents other than water, and the question is raised whether the optimized scaling factor for dispersion in water is of a universal nature or needs to be re-adjusted for each other type of solvent considered. Secondly, we tested the introduced change when the Caillet-Claverie specific formalism of dispersion treatment is changed to AMBER-style dispersion as indicated in eqs. 5.2 and 5.3.

5.2.3 Study of Size- and Charge Dependence

Crystal structures of 10 proteins of increasing size are obtained from the Protein Data Bank [38]. The actual download site is the repository PDB-REPRDB [144]. Structures are purified and processed as described previously [11, 157]. The PDB codes together with a characterization of main structural features of the selected test proteins are summarized in Table 5.2. Two types of calculations are carried out using the semi-empirical model PM5

Table 5.2: PDB codes and structural key data of a series of proteins used for comparison [12]

Shape Sketch	PDB-Code	Number of Residues	Number of Atoms	Charge [a.u.]
	1P9GA	41	517	+3
	2B97	70	981	+1
	1LNI	96	1443	-5
	1NKI	134	2082	+5
	1EB6	177	2570	-11
	1G66	207	2777	-2
	1P1X	250	3813	0
	1RTQ	291	4287	-16
	1YQS	345	5147	+2
	1GPI	430	6164	-12

[176] and the fast multipole moment (FMM) method [177]. A single point vacuum energy

calculation is followed by a single point energy calculation including the COSMO model [173] for consideration of solvent water. The difference between the two types of single point energies should provide us with an estimate of the solvation free energy. Furthermore, the finally computed set of atomic partial charges is extracted from the PM5-calculation and feeded into the PB/BEM model to substitute standard AMBER partial charges. In this way we can examine the dependence on a chosen charge model as well as compare classic with semi-empirical QM approaches to the solvation free energy.

5.2.4 Computational Aspects

The sample set of 10 proteins listed in Table 5.2 is analyzed with respect to computational performance regarding the calculation of the dispersion term as defined in eq. 5.3. It is important to note that for this particular task the surface resolution into BEs may be lowered to levels where the average size of the BEs becomes ≈ 0.45 Å^2. CPU times for the two steps, ie creation and processing of the surface and evaluation of the expression for ΔG^{disp} are recorded and summarized in Table E.12 of the Appendix E. As can be seen clearly from these data, the major rate-limiting step is the production of the surface, which can reach levels of up to 20 % of the total computation time. Evaluation of the dispersion term itself is of negligible computational cost. Since the surface used for the polarization term is defined according to Connolly (see section 5.2.1), we could not use this molecular surface directly for a SASA-based alternative treatment of the non-polar contributions. Rather we had to compute a SASA from scratch too, and were facing identical computational constraints as seen with the approach chosen here.

5.3 Results

5.3.1 A universal scaling factor applied to Caillet-Claverie dispersion coefficients leads to good overall agreement with experimental solvation free energies of amino acid side-chain analogues in water

Since our main focus is on proteins, our first goal is to optimize our approach for proteins in aqueous solution. We can resort to the experimental data for amino acid side-chain analogues (see [13] and references therein). At first we seek maximum degree of agreement between experimental and PB/BEM values of the solvation free energy, ΔG^{solv}, by multiplying a scaling factor, λ, to the Caillet-Claverie dispersion [57, 58] coefficients, κ_i. The remaining terms in eq. 5.1 are computed at the optimized conditions reported previously [174, 56]. We define a global deviation from the experimental data by

$$\Delta\Delta G^{solv} = \frac{1}{13}\sum_{i=1}^{13}\sqrt{\left(\Delta G_i^{solv,Exp} - \Delta G_{i,\lambda}^{solv,PB/BEM}\right)^2} \tag{5.4}$$

where *i* refers to a particular type of amino acid side-chain analogue included in the reference set of experimental values and λ is the introduced scaling factor applied to the Caillet-Claverie dispersion [57, 58] coefficients. The trend of $\Delta\Delta G^{solv}$ for different choices of λ is shown in Figure 5.2. As becomes clear from Figure 5.2, a scaling factor of 0.70 establishes the best match to the experimental data. A detailed comparison of individual amino acid side-chain analogues at this optimum value is given in Table 5.3. We achieve a mean unsigned error of 1.15 $\frac{kcal}{mol}$, hence come close to the accuracy reported recently by

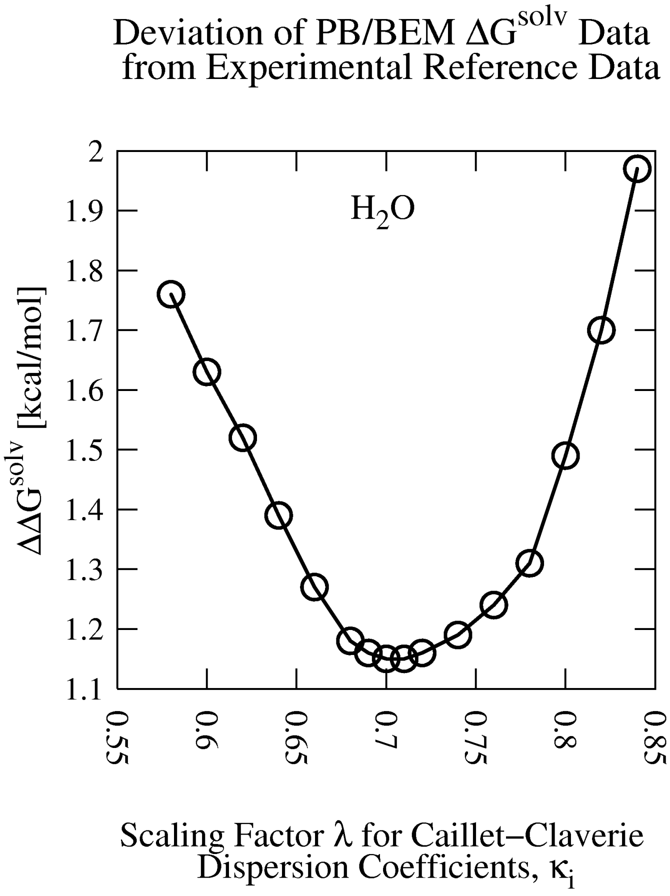


Figure 5.2: Deviation of the PB/BEM ΔG^{solv} from experimental values tabulated in [13] as a function of λ [12].

Chang et al. [13], a study that agreed very well with earlier calculations carried out by Shirts et al. [178] and MacCallum et al. [179].

Table 5.3: Comparison of PB/BEM-computed versus experimental total solvation free energies, ΔG^{solv}, of amino acid side-chain analogues in water. A scaling factor, λ, of 0.70 has been uniformly applied to all dispersion coefficients, κ_i, in eq. 5.2 [12].

Species	$\Delta G^{solv,PB/BEM}$ $\left[\frac{kcal}{mol}\right]$	$\Delta G^{solv,Exp}$ $\left[\frac{kcal}{mol}\right]$	Deviation $\left[\frac{kcal}{mol}\right]$
acetamide	-10.97	-9.68	1.29
butane	1.92	2.15	0.23
ethanol	-4.58	-4.88	0.30
isobutane	1.74	2.28	0.54
methane	0.72	1.94	1.22
methanethiol	-3.57	-1.24	2.33
methanol	-6.58	-5.06	1.52
methyl-ethyl-sulfide	-0.30	-1.48	1.18
methylindole	-4.19	-5.88	1.69
p-cresol	-3.56	-6.11	2.55
propane	1.72	1.99	0.27
propionamide	-9.34	-9.38	0.04
toluene	1.05	-0.76	1.81

Several computed solvation free energies in Table 5.3 still show significant deviation from the experimental value, e.g. p-cresol and methanethiol. A comparison to results with a simple SASA-based model is included in the Appendix E (Table E.11). This comparison reveals a certain improvement for the most critical components, but no indication of a general amelioration of the situation. The somewhat special character of methanethiol has been noticed before [160].

5.3.2 Component-wise juxtaposition of PB/BEM and PCM approaches reveals a difference in individual contributions but similarity in net effects

As interesting as total solvation free energies are the constituting partial terms and how they compare to their analogous counter parts in a high-level QM model such as PCM. We therefore studied all amino acid side-chain analogues with PCM [46] calculations at the Becke-98 [142] level of density functional theory (DFT) using the high-quality basis set of Sadlej [143] and program GAUSSIAN-03 [150]. A summary of these data is given in Table 5.4.

Since in PB/BEM we do not consider repulsion explicitly, the PB/BEM dispersion term is compared to the sum of ΔG^{disp} and ΔG^{rep} of PCM. It becomes clear from Table 5.4 that there is rather general agreement in polarization terms but sizeable divergence in the apolar terms. However, the sum of all apolar terms, ie. ΔG^{cav} and ΔG^{disp}, appears to be again in good agreement when comparing PB/BEM with PCM. The reason for the difference in the apolar terms is largely due to a different cavitation formalism used in PB/BEM, which we currently believe to represent a very good approximation to this term [56].

Table 5.4: Analysis of individual contributions to the net solvation free energy for solvent water as computed by PB/BEM or by PCM [12].

Species	ΔG^{cav}	ΔG^{cav}	ΔG^{disp}	ΔG^{disp}_{rep}	ΔG^{pol}	ΔG^{pol}	ΔG^{solv}	ΔG^{solv}
	PB/BEM	PCM	PB/BEM	PCM	PB/BEM	PCM	PB/BEM	PCM
	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$
acetamide	5.10	12.71	-4.26	-7.45	-11.81	-14.13	-10.97	-8.88
butane	7.05	15.46	-4.41	-8.48	-0.72	-0.45	1.92	6.54
ethanol	4.89	11.79	-3.71	-6.74	-5.76	-6.42	-4.58	-1.37
isobutane	7.17	15.94	-4.44	-8.12	-0.98	-0.55	1.74	7.28
methane	3.08	9.98	-2.10	-3.03	-0.26	-0.07	0.72	6.88
methanethiol	4.19	10.95	-4.12	-6.77	-3.64	-4.35	-3.57	-0.17
methanol	3.39	9.53	-3.05	-4.88	-6.91	-6.02	-6.58	-1.37
methyl-ethyl-sulfide	7.00	16.37	-5.30	-9.49	-2.00	-3.02	-0.30	3.86
toluene	8.54	17.40	-5.51	-11.17	-1.98	-3.73	1.05	2.51
methylindole	10.00	20.67	-7.09	-14.10	-7.10	-10.07	-4.19	-3.50
p-cresol	8.92	18.93	-6.11	-12.16	-6.37	-10.48	-3.56	-3.70
propane	5.80	13.58	-3.68	-6.92	-0.40	-0.34	1.72	6.31
propionamide	6.34	14.56	-4.83	-9.05	-10.84	-13.05	-9.34	-7.54

5.3.3 The identified scaling factor of 0.70 applied to Caillet-Claverie dispersion coefficients yields good quality estimates of the solvation free energy in water for many molecules

In order to test the PB/BEM approach further we used the initially determined scaling factor for dispersion coefficients of 0.70 to compute water solvation free energies of a series of other molecules. The procedure for obtaining atomic partial charges is described in section 5.2.2. It is important to note that the electron density used for RESP fitting must be of MP2/6-31G* quality to achieve maximum degree of compatibility to standard AMBER charges, which have been found to mimic high quality calculations very well [174]. Experimental reference values have been obtained from the extensive compilation by Li et al. [14]. The data comprising 18 arbitrarily selected molecules are summarized in Table 5.5. The mean unsigned error of 1.18 $\frac{kcal}{mol}$ for this set of molecules comes close to PCM quality and must be considered very satisfactory again. Another class of molecules we looked into are amino acids in their zwitterionic form, where due to the charges at the amino/carboxy groups the net solvation free energies become larger by about an order of magnitude. A comparison against the recently reported data by Chang et al. [13] is given in Table 5.6. The degree of agreement is still considerably high and there is no obvious indication of a systematic deviation. A final comparison is made against a series of 180 molecules that has been used in a previous study [174]. These structures include all 20 types of naturally occurring amino acids in 9 different conformations (zwitterionic forms assumed). The set of dipeptides has been subjected to PCM [46] calculations at the Becke-98 DFT level [142] using Sadlej's basis set [143]. Average net solvation free energies are formed from all 9 different conformations per type of amino acid (or the number of available reference calculations) and the results are presented in Table E.1 of the Appendix E. Considering the variation with respect to conformational flexibility the match must still be considered to be reasonably good. It is interesting to note that the variability of the

Table 5.5: Individual contributions to the water net solvation free energy as computed from PB/BEM or PCM for a series of arbitrary small molecules [12].

Species	ΔG^{cav}	ΔG^{cav}	ΔG^{disp}	ΔG^{disp}_{rep}	ΔG^{pol}	ΔG^{pol}	ΔG^{solv}	ΔG^{solv}
	PB/BEM	PCM	PB/BEM	PCM	PB/BEM	PCM	PB/BEM	PCM
	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$
propanal	6.04	13.26	-3.92	-7.67	-4.32	-6.35	-2.21	-0.76
butanoic acid[(a)]	7.54	16.98	-5.31	-10.3	-8.18	-10.85	-5.94	-4.16
cyclohexane	8.77	16.45	-5.29	-11.56	0.00	-0.58	3.48	4.31
acetone	5.77	14.30	-3.96	-7.04	-4.67	-6.05	-2.85	1.21
propene	5.55	12.60	-3.58	-6.48	-0.98	-1.24	0.99	4.88
propionic acid[(a)]	6.31	14.57	-4.60	-8.73	-8.38	-10.42	-6.67	-4.59
propyne	4.87	12.07	-3.14	-5.75	-2.36	-3.33	-0.62	2.99
hexanoic acid[(a)]	9.97		-6.76		-8.33		-5.12	
anisole	8.66		-6.12		-3.27		-0.73	
benzaldehyde	8.47	17.26	-5.74	-11.99	-5.05	-9.38	-2.32	-4.12
ethyne	4.16	9.78	-2.76	-4.92	-0.96	-1.05	0.44	3.81
butanal	7.18	15.75	-4.63	-9.33	-4.55	-6.77	-2.00	-0.36
benzene	7.24	14.21	-4.84	-10.27	-2.76	-4.04	-0.36	-0.10
bromobenzene	8.67	16.96	-5.91	-12.73	-2.46	-4.76	0.29	-0.53
acetic acid[(a)]	4.89	12.38	-3.92	-7.02	-8.41	-10.49	-7.44	-5.13
bromoethane	5.95	13.09	-4.25	-8.35	-1.61	-2.77	0.09	1.98
ethylbenzene	9.57	19.57	-6.11	-12.68	-1.92	-3.62	1.54	3.27
diethylether	7.49	17.71	-5.12	-9.47	-1.41	-2.48	0.97	5.76

[(a)] protonated form

dispersion contributions alone, considered isolated per se as a function of conformational flexibility is much less pronounced than what we see for the net solvation (see Table E.2 of the Appendix E).

Table 5.6: Comparison of PB/BEM computed solvation free energies of zwitterionic amino acids in water against data by Chang et al. [13] obtained from Monte Carlo Free Energy simulations [12].

Species	$\Delta G^{solv,PB/BEM}$	$\Delta G^{solv,MC}$	Deviation
	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$	$\left[\frac{kcal}{mol}\right]$
Gly	-55.73	-56.80	1.07
Ala	-51.75	-57.70	5.95
Val	-48.79	-56.20	7.41
Leu	-49.05	-57.30	8.25
Ile	-47.55	-55.70	8.15
Ser	-60.82	-55.30	5.52
Thr	-61.33	-54.40	6.93
Cys	-60.86	-54.70	6.16
Met	-50.88	-57.30	6.42
Asn	-58.63	-60.10	1.47
Gln	-65.82	-59.60	6.22
Phe	-51.46	-55.90	4.44
Tyr	-55.16	-61.60	6.44
Trp	-58.00	-64.60	6.60

5.3.4 The scaling factor of 0.70 applied to Caillet-Claverie dispersion coefficients in the case of water is not of a universal nature but must be re-optimized for any other type of solvent.

An important aspect of the PB/BEM approach is how the identified scaling factor for Caillet-Claverie dispersion coefficients — 0.70 in the case of water — translates into other situations of non-aqueous solvation. We have therefore repeated the studies for identifying optimal boundaries [174] for solvents methanol, ethanol and n-octanol. Again we consider PCM cavities of the set of 180 dipeptide structures as reference systems and search for the best match in volumes and surfaces dependent on slightly enlarged or shrinked standard AMBER van der Waals radii. We again employ the molecular surface program SIMS [10]. Detailed material of this fit is included in the Appendix E (Table E.3-E.8 and Figure E.1-E.3). We find to have to marginally increase AMBER van der Waals radii by factors of 1.06 in solvents methanol and ethanol and 1.05 in solvent n-octanol. Based on these conditions for proper locations of the solute-solvent interface we then repeat the search for appropriate scaling factors of dispersion coefficients that result in close agreement to experimental solvation free energies (see section 5.3.1). Results are presented in Figures 5.3 and 5.4. It becomes clear that the factor of 0.70, optimal for water, is not universally applicable. Rather, we find for ethanol 0.82 and for n-octanol 0.74 to be the optimal choices. A detailed comparison against experimental values at optimized conditions is given in Tables 5.7 and 5.8. We achieve mean unsigned errors of 1.38 $\frac{kcal}{mol}$ for ethanol and 1.27 $\frac{kcal}{mol}$ for n-octanol.

Cavitation terms of similar quality to the ones presented in [56], which are needed in PB/BEM, are available for methanol and ethanol (unpublished work in progress) or obtained from [180]. Unfortunately, we cannot do the calculations for methanol because of the lack of experimental values and the non-systematic trend in dispersion scaling factors of the other alcoholic solvents. All optimized parameter sets for the various types of solvents are summarized in Table 5.9.

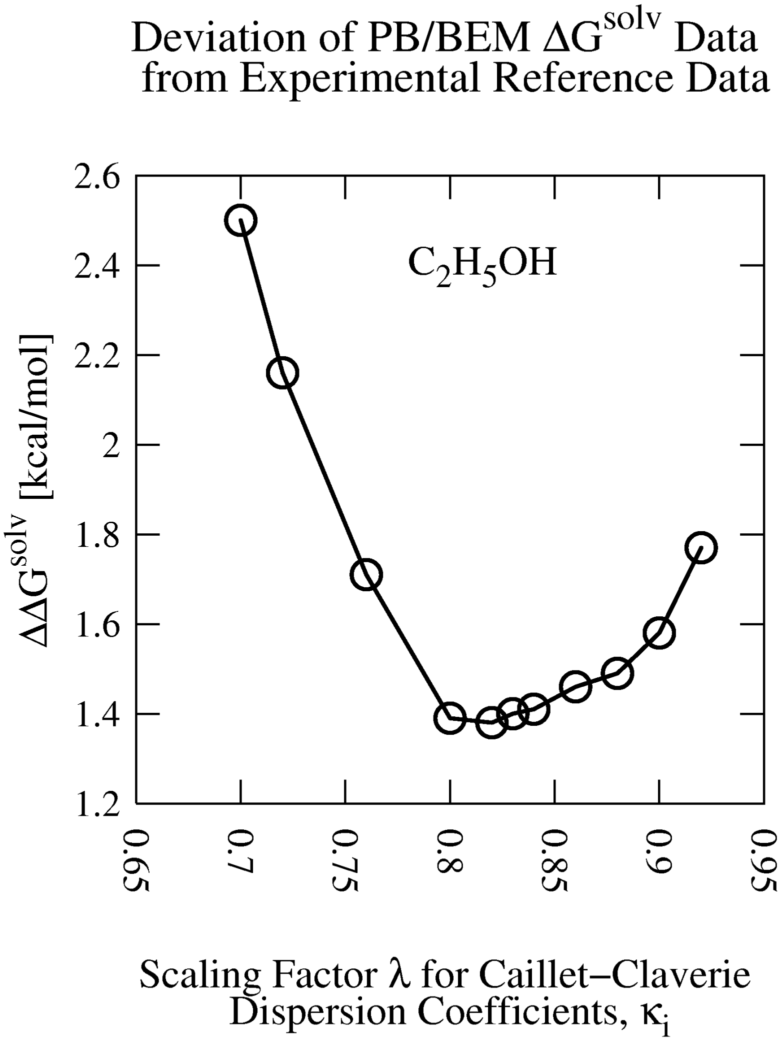


Figure 5.3: Ethanol: Deviation of the PB/BEM ΔG^{solv} from experimental values tabulated in [14] as a function of λ [12].

5.3.5 Switching from Caillet-Claverie-style of dispersion to AMBER-style requires a re-adjustment of scaling factors.

An obvious question is how the described approach will change when substituting the Caillet-Claverie formalism with the corresponding AMBER-dispersion formula, ie replac-

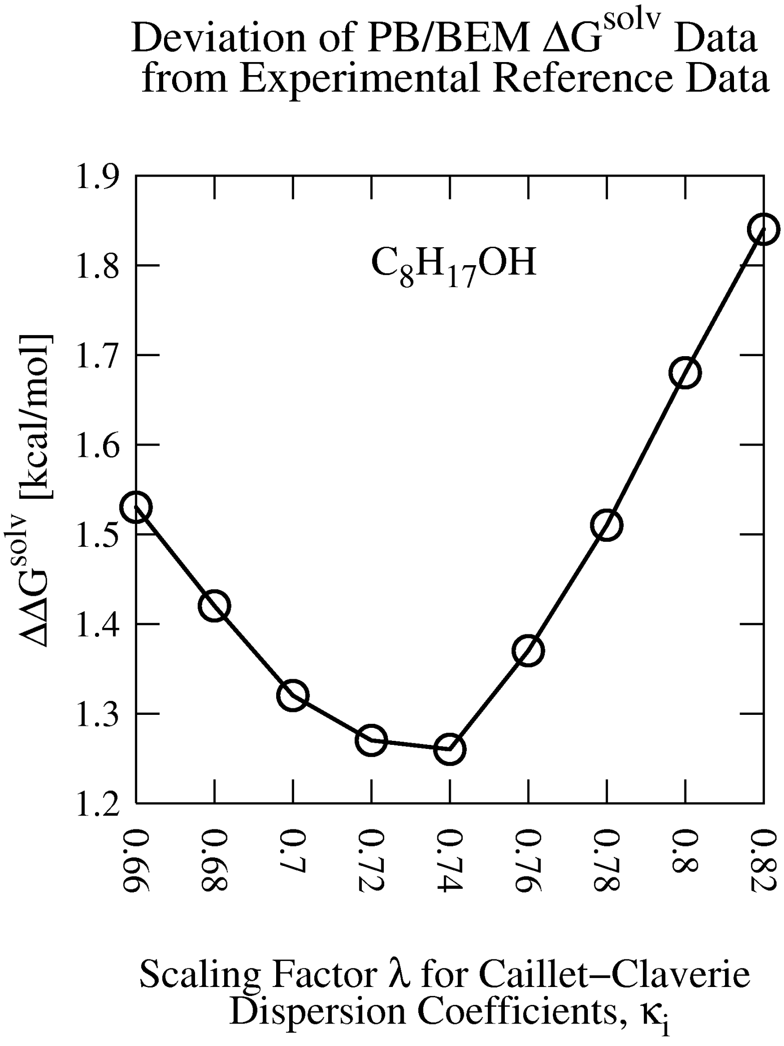


Figure 5.4: n-Octanol: Deviation of the PB/BEM ΔG^{solv} from experimental values tabulated in [14] as a function of λ [12].

ing eq. 5.2 with eq. 5.3. We therefore implemented a variant where we use eq. 5.3 together with standard AMBER van der Waals radii (slightly increased as done for the definition of the boundary and indicated in table 5.9) and standard AMBER van der Waals potential well depths. Similar to the Caillet-Claverie treatment we find that a uniform scaling factor, λ, applied to the AMBER van der Waals potential well depths, ε_i, is sufficient to lead to

Table 5.7: Comparison of PB/BEM-computed versus experimental total solvation free energies, ΔG^{solv}, of various substances in ethanol [12].

Species	$\Delta G^{solv,PB/BEM}$ $\left[\frac{kcal}{mol}\right]$	$\Delta G^{solv,Exp}$ $\left[\frac{kcal}{mol}\right]$	Deviation $\left[\frac{kcal}{mol}\right]$
n-octane	-0.70	-4.23	3.53
toluene	-3.30	-4.57	1.27
dioxane	-6.03	-4.68	1.35
butanone	-4.83	-4.32	0.51
chlorobenzene	-3.52	-3.30	0.22

good agreement with experimental data. An identical strategy to the one presented in section 5.3.1 for determination of appropriate values of λ may be applied. The optimal choice of λ turns out to be 0.76 for solvent water as indicated in Figure E.6 of the Appendix E. Corresponding detailed data is shown in Table 5.10.

The mean unsigned error amounts to 1.01 $\frac{kcal}{mol}$ at optimized conditions. While in the case of water similar scaling factors are obtained for Caillet-Claverie as well as AMBER type of dispersion, for the remaining types of solvents a less coherent picture arises (see Table 5.9). Identification of scaling factors for solvents ethanol (λ=0.94) and n-octanol (λ=2.60) is shown in Figure E.3 of the Appendix E and corresponding detailed data listed in Tables E.9 and E.10 of the Appendix E. Mean unsigned errors are 1.21 $\frac{kcal}{mol}$ for ethanol and 1.00 $\frac{kcal}{mol}$ for n-octanol respectively. Either approach is competitive and comes with its own

Table 5.8: Comparison of Poisson-Boltzmann/Boundary Element Method (PB/BEM)-computed versus experimental total solvation free energies, ΔG^{solv}, of various substances in the solvent n-octanol [12].

Species	$\Delta G^{solv,PB/BEM}$ $\left[\frac{kcal}{mol}\right]$	$\Delta G^{solv,Exp}$ $\left[\frac{kcal}{mol}\right]$	Deviation $\left[\frac{kcal}{mol}\right]$
acetone	-5.28	-3.15	2.13
anisole	-4.80	-5.47	0.67
benzaldehyde	-6.16	-6.13	0.03
benzene	-3.87	-3.72	0.15
bromobenzene	-3.75	-7.47	3.72
butanal	-5.02	-4.62	0.40
butanoic acid$^{(a)}$	-8.74	-7.58	1.16
cyclohexane	-0.64	-3.46	2.82
acetic acid$^{(a)}$	-8.96	-6.35	2.61
ethylbenzene	-2.94	-5.08	2.14
ethylene	-1.57	-0.27	1.30
hexanoic acid$^{(a)}$	-8.89	-8.82	0.07
propanal	-4.71	-4.13	0.58
propionic acid$^{(a)}$	-8.75	-6.86	1.89
propene	-1.61	-1.14	0.47
propyne	-2.81	-1.59	1.22
bromoethane	-2.69	-2.90	0.21

$^{(a)}$ protonated form

merits. Caillet-Claverie coefficients are more general and specific to chemical elements only, hence no distinction between for example sp3-C atoms and sp2-C atoms needs to be made. Employment of AMBER parameters on the other hand appears to be straightforward in the present context since the geometry of the boundary is already based on AMBER van der Waals radii.

Table 5.9: Summary of optimized parameters to be used in Poisson-Boltzmann/Boundary Element Method (PB/BEM) for different types of solvents. Average sizes of boundary elements (BEs) are given as pairs of values employed for calculation of ΔG^{pol} and ΔG^{disp} respectively [12].

Parameter Class	Water	Methanol	Ethanol	n-Octanol
BE Average Size [Å^2]	0.31/0.45	0.31/0.45	0.31/0.45	0.31/0.45
Probe Sphere Radius [Å]	1.50	1.90	2.20	2.945
AMBER vdW Radii Scaling	1.07	1.06	1.06	1.05
AMBER Partial Charges Scaling	1.00	1.00	1.00	1.00
Caillet-Claverie Dispersion Coefficients Scaling	0.70	–	0.82	0.74
AMBER vdW Potential Well Depth Scaling	0.76	–	0.94	2.60

5.3.6 Replacement of static AMBER partial charges with semiempirical PM5 charges introduces a rise in solvation free energies by about 20 % of the classic result regardless of the size or total charge state of the system.

A series of proteins of different size, shape and total net charge (see Table 5.2) is computed within the PB/BEM approach at optimized conditions for aqueous solvation, that is using a Caillet-Claverie dispersion coefficient scaling factor of 0.70, slightly increased AMBER van der Waals radii by a factor of 1.07 and standard AMBER partial charges. In addition to

Table 5.10: Effect on total solvation free energies for water as Poisson-Boltzmann/Boundary Element Method (PB/BEM)-computed with AMBER style of dispersion (eq. 5.3) versus Caillet-Claverie style of dispersion (eq. 5.2) and comparison to the experimental value [12].

Species	$\Delta G^{solv}_{Caillet-Claverie}$ $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{AMBER} $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{Exp} $\left[\frac{kcal}{mol}\right]$
propanal	-2.21	-1.71	-3.44
butanoic acid[(a)]	-5.94	-6.00	-6.47
cyclohexane	3.48	-1.33	1.23
acetone	-2.85	-2.42	-3.85
propionic acid[(a)]	-6.67	-6.57	-6.47
propyne	-0.62	-2.09	-0.31
hexanoic acid[(a)]	-5.12	-5.81	-6.21
anisole	-0.73	-3.49	-2.45
benzaldehyde	-2.32	-3.22	-4.02
butanal	-2.00	-1.86	-3.18
benzene	-0.36	-2.78	-0.87
bromobenzene	0.29	-1.63	-1.46
acetic acid[(a)]	-7.44	-6.76	-6.70
bromoethane	0.09	-0.32	-0.70
ethylbenzene	1.54	-1.25	-0.80
diethylether	0.97	-0.68	-1.76

[(a)] protonated form

this classic approximation we also carry out semi-empirical QM calculations with the help of program LocalSCF [19] using the PM5 model. From the semi-empirical calculation we extract atomic partial charges and use these instead of AMBER partial charges within the PB/BEM approach. Results of these calculations are presented in Table 5.11 and Figure 5.5. In general one can observe rather a constant change of about 20 % of the classic AMBER

Table 5.11: Analysis of partial term contributions to Poisson-Boltzmann/Boundary Element Method (PB/BEM)-computed solvation free energies for a series of proteins of increasing size using either molecular dynamic package AMBER [18] standard partial charges or semi-empirical PM5 charges obtained from program LocalSCF [19, 12].

Species	$\frac{Surface}{Volume}$ $\left[\frac{1}{-}\right]$	ΔG^{cav} $\left[\frac{kcal}{mol}\right]$	ΔG^{disp} $\left[\frac{kcal}{mol}\right]$	ΔG^{pol}_{AMBER} $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{AMBER} $\left[\frac{kcal}{mol}\right]$	ΔG^{pol}_{PM5} $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{PM5} $\left[\frac{kcal}{mol}\right]$
1P9GA	0.42	114.6	-66.2	-339.9	-291.6	-251.6	-203.3
2B97	0.34	181.3	-91.3	-548.0	-457.9	-517.5	-427.4
1LNI	0.34	237.5	-126.8	-1418.6	-1307.9	-1140.3	-1029.6
1NKI	0.37	305.4	-199.2	-1652.1	-1546.0	–	–
1EB6	0.28	353.0	-182.7	-2571.9	-2401.7	-2312.2	-2141.9
1G66	0.26	369.0	-187.5	-1193.6	-1012.1	-881.5	-700.0
1P1X	0.25	459.3	-235.4	-2434.6	-2210.7	-2106.8	-1882.9
1RTQ	0.22	506.4	-238.2	-4077.5	-3809.3	-3172.4	-2904.1
1YQS	0.23	566.9	-286.4	-2133.4	-1852.9	-1680.8	-1400.3
1GPI	0.24	651.8	-342.6	-3961.3	-3652.1	-3252.0	-2942.8

based ΔG^{solv} estimate when switching to PM5 charges. This is independent of the size, shape or net charge of the system (compare red bars with purple bars in Figure 5.5). The polarization term constitutes the major contribution but apolar terms are far from negligible (compare magnitude of blue and black bars to green and grey bars in Figure 5.5). When using the COSMO approximation within the semi-empirical method and deriving solvation free energies from that we get entirely uncorrelated results for the solvation free energy, ΔG^{solv} (data not shown). It is important to note that the surface to volume ratio drops to a value around 0.25 with increasing protein size, whereas typical values in the range of 0.80

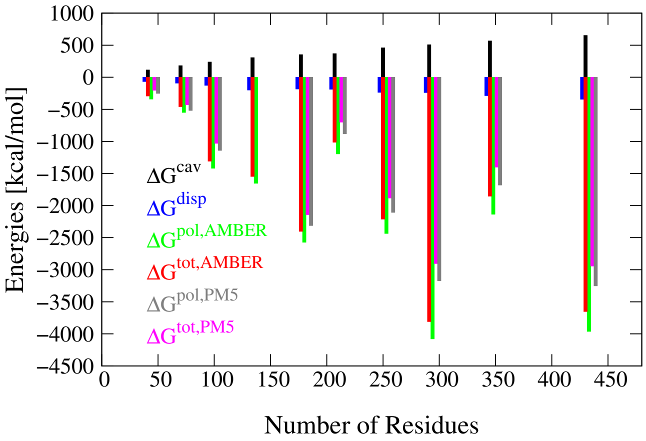


Figure 5.5: Classic versus semi-empirical charge assignments to atoms of proteins of various size used in PB/BEM calculations [12].

to 1.0 are maintained in the initial calibration phase, hence care must be taken with large scale extrapolations from small molecular reference data.

5.4 Discussion

Motivated by the recent high-performance implementation of Poisson-Boltzmann calculations [147] we now complement this approach with a systematic inclusion of apolar effects. In particular the important dispersion contribution is introduced and fine-tuned against available experimental data. This is based on physics-based terms, that have long been considered in a similar fashion within QM models [46]. The resulting model is applied to a series of protein structures, and size and charge effects are examined.

Direct assessment of the predictive quality of the PB/BEM approach after calibration has revealed rather good performance indicators for PB/BEM. This was based on suggested scaling factors applied to Caillet-Claverie dispersion coefficients. Since the original aim of Caillet-Claverie was to explain crystal data, we would expect a need for re-adjustment in this present implementation. Moreover, since the boundary and the rest of the PB/BEM model is based on AMBER parameterization it does not come as a surprise that one has to adjust a non-related second set of van der Waals parameters in order to achieve general agreement to a reference data set. Related to this point it seems particularly encouraging that when replacing the scaled Caillet-Claverie part with standard AMBER-dispersion terms for water no further refinement is necessary and similar levels of precision are established automatically. In the case of water, this brings in a second advantage. Because the employed TIP3P model assigns van der Waals radii of zero to the H-atoms, so the effective sum over i in eq.5.3 may be truncated already after the oxygen atom. The second cycle considering H-atoms in water would add only zeros.

A somewhat critical issue is the determination of missing parameters or the estimation of solvent probe sphere radii for different types of solvents. In this present work we found it convenient to make use of electron density grids and corresponding iso-density thresholds to define the boundary of molecules. For example to determine the probe sphere radius of methanol we compute the volume of a single molecule of methanol up to an electron density threshold of 0.0055 a.u. and derive an effective radius assuming spherical relationships. The same threshold criterion is applied to all other solvents leading to the data summarized in Table 5.9. Electron grids are based on B98/Sadlej calculations. Similarly we determine atomic van der Waals radii for Cl- or Br-containing substances from isodensity considerations. However in these latter cases the threshold criterion is adapted to a level that re-produces proper dimensions of well-known types, ie neighboring C-, O-, N-atoms and at this level the unknown radius is determined. In the case of n-octanol the assumption of spherical geometry is certainly not justified. On the other hand the concept of an over-rolling probe sphere representing approaching solvent molecules will remain a hypothetical model construct anyway. Complying with this model construct it may be argued that over time the average of approaching solvent molecules will hit the solute with all parts (head, tail or body regions of the solvent molecule) equally often and thus the idealized spherical probe is not entirely unreasonable.

Another interesting aspect is the fact that the present PB/BEM approach is all based on molecular surfaces rather than SASAs. This is of technical interest and the consequence of that is a greatly reduced sensitivity to actual probe sphere dimensions. A graphical explanation is given in the Appendix E (Figure E.4). While SASA based surfaces would see significant changes when probe spheres are slightly modified (blue sphere replaced by red sphere in Figure E.4 of the Appendix E) the molecular surface itself faces only a minor change in the reentrant domain (green layer indicated in Figure E.4 of the Appendix E).

Large scale extrapolations resulting from a calibration process done with small sized reference structures have to be taken with care. Because of the drop of surface to volume ratios the most important requirement for such a strategy is to have the individual terms properly analyzed whether they scale with the volume, or the surface. For PB/BEM the question reduces to the cavitation term, since the remainder is mainly a function of Coulombic interactions. As that particular aspect has been carefully analyzed in previous studies [59] we are confident that a large-scale extrapolation actually works in the way suggested in eq. 5.1.

A final remark may be relevant with regard to the discrepancy seen in using classic AMBER partial charges versus semi-empirical PM5 charges. Intuitively, one is tempted to believe stronger in the PM5 results. There might however also be a small drift in energies introduced by PM5/PM3 models as has been observed within an independent series of single point calculations (see Appendix E).

5.5 Conclusion

Consideration of dispersion effects within a physics-based continuum solvation model significantly improves accuracy and general applicability of such an approach. The proposed method follows a proven concept [168] and is easily implemented into existing models. Generalization to different treatments of dispersion as well as extension to non-aqueous solvents is straightforward.

Chapter 6

Algorithmic Refinement & Application

This chapter is reproduced in part from my following two publications–1. P. Kar, M. Seel, U. H. E. Hansmann, S. Höfinger. Algorithmic refinements to an Enhanced Poisson-Boltzmann Approach Used in Biomolecular Simulations. NIC Publication Series, Vol. 36, 173-176(2007) [15] and 2. P. Kar, M. Seel, U. H. E. Hansmann, S. Höfinger. Comparing Semiempirical versus Classical Charge Assignments in Biomolecules and Their Effect on Electrostatic Potentials. NIC Publication Series, Vol. 36, 155-158 (2007) [20]

6.1 Algorithmic Refinements

6.1.1 Introduction

Biological molecules typically reside in aqueous environments. Reliable consideration of the effect of water on structure and dynamics of biomolecules is among the key factors governing accurate descriptions of biological matter [50]. Here we focus on an implicit solvation model. Among other methods, e.g. SASA, GB, FDPB, the Poisson - Boltzmann (PB) approach [49] within the Boundary Element Method (BEM) [54] is frequently chosen due to its intermediate position regarding computational cost versus achievable accuracy. In our recent series of publications [11, 12] we have outlined a generalization of the Polarizable Continuum Model (PCM) [46] applied to biomolecular structure. Each of the considered terms represents a separate portion of distinct physical interaction,

$$\Delta G^{sol} = \Delta G^{pol} + \Delta G^{disp,rep} + \Delta G^{cav} \tag{6.1}$$

which are polarization, dispersion and cavitation. The latter plays an important role in hydrophobicity related phenomena [181]. Care has been taken to operate the model at conditions that guaranteed a maximum level of numerical accuracy. However, a number of internal parameters could still profit from further optimization.

6.1.2 Aim

In this present study, we address the following factors and examine their consequences on run-time performance with regard to a series of test proteins of increasing size that we have studied earlier [12],

(i) the exact value of the exit criterion used to terminate the calculation of the polarization term, Δ G^{pol},

(ii) the array dimension regulating the allowed number of consecutive DIIS [182] steps,

(iii) the switch criterion used to move from the pre-DIIS stage to the DIIS stage,

(iv) the dependence on system size of the number of necessary iterations to achieve convergence,

(v) the dependence on renormalization factors applied to the net sum of polarization charges,

(vi) the influence of very small-sized boundary elements (BEs), or the introduced change when merging these very small-sized elements to larger ones from the neighborhood,

(vii) the necessary degree of surface resolution for accurate calculation of the dispersion term, Δ G^{disp}.

6.1.3 Procedure

We select 10 proteins of different size (number of residues reaching from 41 to 430). Initially we run the PB/BEM program POLCH [59] at default conditions. The run time for all the 10 cases is recorded and forms a reference set. At first, we adjust the parameter MAXNIT which defines the maximum number of successive DIIS steps, hence determines the size of the DIIS matrix, and compute the run time deviation from the reference set for all the 10 proteins. Once parameter MAXNIT is optimized, we rerun the entire test set and extract net solvation free energies, Δ G^{sol}, which serve as a new reference. ACCURA is the second parameter to be optimized. It defines the threshold criterion used for termination of the iterative process when computing the polarization term, Δ G^{pol}. For optimizing ACCURA we require the deviation from the reference set not to exceed ± 0.05 kcal/mol for any of the proteins. Once ACCURA is optimized we redo the whole set of test proteins at optimized conditions for either parameter, ACCURA as well as MAXNIT. We extract the number of iterations needed for completion and use these as a new reference. In our next step we optimize the parameter DSNTRC. This parameter sets the switch criterion used to move from a pre-DIIS stage to the DIIS stage. It represents the mean square deviation of two successive sets of polarization charges. We keep changing DSNTRC and optimize for a minimum number of necessary iterations. The next point is concerned with the renormalization of the polarization charges according to Gauss' Law. We study the effect this has on the net solvation free energies. The solvation free energies obtained after renormalization form another reference set for our next investigation. Here, we study the influence of very small-sized BEs. We will merge these very small-sized elements to larger ones from the neighborhood. We change the parameter REQSZ (the required minium size of a BE) and compute the deviation of solvation free energies from the reference set. We again do not

Table 6.1: Sensitivity to total system size, total charge and renormalization attempts [15].

Protein PDB Code	No. of Residues	Molecular Charge (a.u.)	No. of Iterations	Δ G^{sol} Without Norm. (kcal/mol)	Δ G^{sol} Including Norm. (kcal/mol)	Δ G^{sol} Deviation (unsigned) (kcal/mol)
1P9GA	41	+3	9	-319.73	-321.54	1.81
2B97	70	+2	8	-40.22	-39.26	0.96
1LNI	96	-5	10	-534.64	-536.39	1.75
1NKI	134	+5	10	-456.20	-454.94	1.26
1EB6	177	-11	10	-1224.17	-122.75	1.42
1G66	207	-2	9	-118.26	-119.01	0.75
1P1X	250	+3	11	-636.41	-636.41	0.00
1RTQ	297	-16	11	-1998.72	-2011.23	12.51
1YQS	345	+2	11	-217.22	-217.22	0.04
1GPI	430	-12	12	-1259.54	-1271.59	12.05

allow the energy to change more than by ± 0.05 kcal/mol in all test runs. Finally, we use all previously optimized parameters for a final test focusing on the dispersion term. We change the resolution of the boundary used for calculation of Δ G^{disp} which need not be maintained at such rigorous levels as identified for the polarization term [11].

6.1.4 Results and Conclusions

Sensitivity to total system size, total charge and renormalization attempts is represented in Table 6.1. Variation of the termination criterion is graphically represented in Figure 6.1. In summary we find that the following parameters lead to a reasonable degree of numerical accuracy.

(1) Best performance is achieved when the DIIS matrix is dimensioned 7x7,

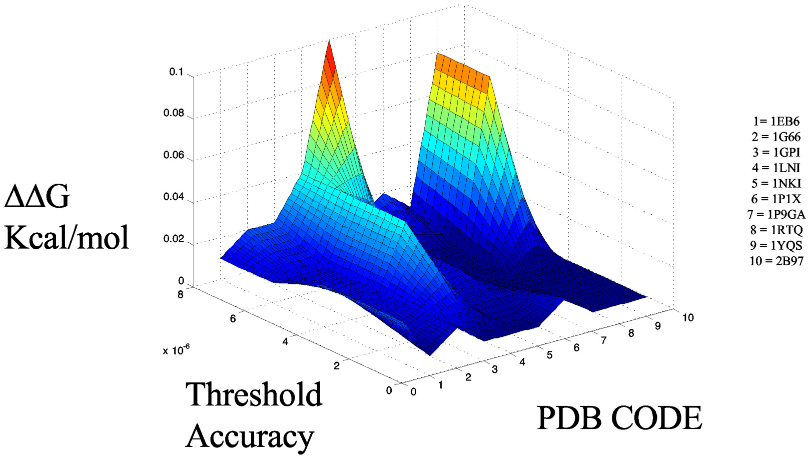


Figure 6.1: Numerical sensitivity of the employed enhanced Poisson-Boltzmann approach to the threshold criterion used for termination of the iterative sequence to calculate the polarization term, ΔG^{pol} [15].

(2) Using a threshold criterion of 4.0x10^{-6} for termination of the iterative sequence occurring in ΔG^{pol} computation leads to stable numerical results.

(3) The best switch criterion to move from the pre-DIIS stage to the DIIS stage is given when the root mean square deviation between two successive sets of polarization charges falls below 0.05 a.u.

(4) The number of iterations necessary to achieve convergence does not depend on system size.

(5) A renormalization process will affect the net solvation free energies, ΔG^{sol}, on the order of ± 1-2 % of their total values. Systems with large net charges are more sensitive to renormalization.

(6) If we merge small sized BEs to larger ones then no significant changes will occur when this procedure is limited to elements smaller than 8 % of the mean size (0.31 Å^2). A reduction in number of BEs will lower the computational cost and foster numerical stability.

(7) For calculation of the dispersion term, ΔG^{disp}, we can reduce the discretization of the boundary into BEs of average size 0.45 Å^2 without loss of accuracy.

6.2 Applications

Poisson-Boltzmann based implicit solvent models have numerous applications in biomolecular simulations. We have applied our enhanced solvent model to estimate the Electrostatic Potential (ESP) of an antifungal protein. We have described the importance of electrostatic potential in structural biology and our findings.

6.2.1 Electrostatic Potential

6.2.2 Introduction

Electrostatics plays an integral part in the study of structure and function of proteins at physiological conditions [50]. Theoretical considerations of the electrostatics in proteins are usually based on solutions to the Poisson-Boltzmann (PB) equation [49, 54]. All these theoretical descriptions will involve a certain type of charge assignment to the atoms of the protein. Since the result of the PB calculation will inevitably depend on the particular choice made for the charges, it might be of interest to study the influence and variation resulting from different charge assignments. Of particular interest will be the comparison between a set of classic charges, ie from force fields commonly employed in the simulation of biomolecules, and charges derived from *ab-inito* calculations performed at a certain level of Quantum Mechanical (QM) theory.

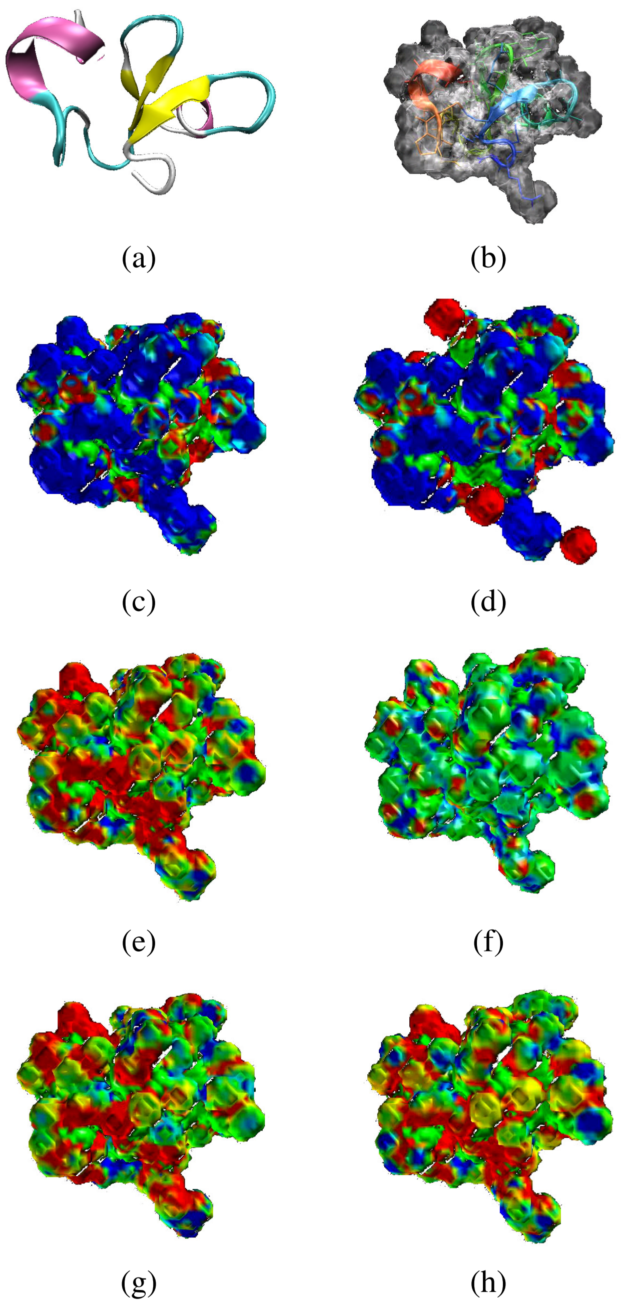
A convenient method to compare different charge assignments to each other is to study the shape and appearance of electrostatic potential (ESP) maps. These ESP maps describe the way the protein will represent itself to its environment in electrostatic terms. Since the solution to the PB equation is included, ESP maps render a reasonably complete picture of the protein in its native environment, ie at physiological conditions. Moreover, ESP maps are a useful tool with many direct applications in structural biology. For example, from ESP maps we can learn whether a protein,
(i) is likely to migrate to the membrane [183],
(ii) will potentially bind RNA or DNA [184, 185],
(iii) belongs to a certain family [186, 187, 188],
(iv) offers a chemically attractive binding site to ligands and other proteins. In this present study, we, therefore, comepare ESP maps based on classic charge assignments using AMBER paramters [146] with ESP maps resulting from semi-empirical charges computed with program LocalSCF [19] at several levels of semi-empirical theory, ie AM1, MNDO, PM3 and PM5. The PB program POLCH [147] is used throughout.

6.2.3 Methods

After download of the protein with pdb code EAFP2 from the pdb data bank, a PB calculation is performed using program POLCH [147] and classic AMBER partial charges [146]. Inner/outer dielectric constants are set to 1 and 80 respectively. The net charge is +4 due to the four Arg residues. ESP maps are computed on the molecular surface and on a cubic grid superimposing the protein. Only ESP maps directly mapped onto the molecular surface are used for further analysis. Semi-empirical calculations are then carried out on the protein EAFP2 using LocalSCF [19] and finally computed partial charges are extracted from the output. The net charge is +2 due to different treatment of lone-pairs in the semi-empirical models. AM1, MNDO, PM3 and PM5 methods are applied. Classic AMBER partial charges are then replaced with either charge set derived from the semi-empirical calculations and PB calculations are repeated with the changed charge assignment. Resulting ESP maps are compared in the form of difference ESP maps.

6.2.4 Results and Conclusions

A structural sketch of the antifungal protein EAFP2 is shown in Table 6.2 (a) with corresponding representation of the molecular surface (b). Here the N-terminal end is colored in red while the C-terminus is given in blue. The ESP map based on classic AMBER charge assignment after PB calculation is represented in Table 6.2 (c). ESP levels are color-coded



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Table 6.2: Electrostatic Potential (ESP) maps for the antifungal protein EAFP2 (pdb code). Major structural elements are shown in (a) and a corresponding representation of the molecular surface is shown in (b). The ESP mapped onto the molecular surface after solution of the PB equation based on AMBER charge assignment is shown in (c). Blue patches correspond to the +5 kT/e level, green regions represent neutral ESP and red domains indicate -5 kT/e level. The marginal change when including 4 explicit Cl^- counter ions is shown in (d). A differential ESP map representing the difference between ESP(AM1) and ESP(AMBER) is shown in (e) with the same color-coding scheme used in (c). Further differential maps are ESP(AM1)-ESP(MNDO) (f), ESP(PM3)-ESP(AMBER) (g) and ESP(PM5)-ESP(AMBER) (h) [20]

as +5 kT/e (blue), 0 kT/e (green) and -5 kT/e (red). It becomes clear that the major appearance of EAFP2 in aqueous solution is that of a macroscopic particle of largely positive ESP, hence the tendency to migrate to the membrane can be explained straightforwardly [183] (which also implies the antifungal mode of action). An initial test regarding the sensitivity to counter ions is shown in Table 6.2 (d). Here explicit Cl^- counter ions have been included in the PB calculation and corresponding ESP maps produced. The change in major ESP patterns introduced by counter ions is only marginal, thus the rest of the analysis is performed without consideration of counter ions. A differential ESP map representing ESP(AM1) - ESP(AMBER) is shown in Table 6.2 (e). Identical color-coding is used as mentioned above. It becomes clear that the AM1-based ESP map is comparable in sign, but significantly different in magnitude (individual ESP values have become less positive). Extended red patches mark off regions of most severe difference. Contrary to the change seen in the AM1-AMBER differential map, when comparing AM1 with MNDO we obtain essentially only green patches (see Table 6.2 (f)). Thus AM1 and MNDO deliver essentially the same ESP properties. Comparison of PM3 with AMBER is represented in the differential ESP map shown in Table 6.2 (g). The trend is similar to the one seen with AM1, but the difference is less severely pronounced (ie certain extended red regions turn yellow or green). Switching further to PM5 description is continuing the trend, ie lessening the deviation from the AMBER-based map again (see Table 6.2 (h)). Closer examination of the residues lying beneath the red-colored patches (indicating most severe deviation) reveals a specific role of Arg residues and the charges assigned to the N-atoms of Asn and Gln.

In summary, semi-empirical charge assignments deliver a consistent picture of significant differences seen for the charged residues. However, individual semi-empirical models differ considerably amongst each other. With increasing sophistication of the semi-empirical model the deviation from the classic AMBER results becomes less severe.

Chapter 7

Enhanced Sampling- Microcanonical Replica Exchange

This chapter is reproduced from our paper– P. Kar, W. Nadler and U. H. E. Hansmann; Microcanonical Replica Exchange Molecular Dynamic Simulation of Proteins, Phys. Rev. E 80, 056703(2009). Copyright Americal Physical Society (2009). The author has the right to use the article or a portion of the article in a thesis or dissertation without requesting permission from APS, provided the bibliographic citation and the APS copyright credit line are given on the appropriate pages ($http : //forms.aps.org/author/copyfaq.html$).

7.1 Introduction

In the last years we have seen remarkable progress in modeling the folding, aggregation and interaction of proteins. For instance, a recent investigation of a 49-residue C-terminal fragment of the artificial protein TOP7, relying on an all-atom force field and an implicit solvent, found not only a lowest energy configuration within 2 Å to the experimentally determined structure, but also a novel folding mechanism that relies on "caching" of a N-terminal "chameleon" segment [189]. These successes are mainly due to the advances in sampling techniques. Generalized-ensemble and replica exchange techniques [80] are now routinely used to enhance the sampling of low-energy configurations, and — especially in their optimized forms [81, 82, 17, 83, 84] — have led to much faster convergence at physiological temperatures than achieved in regular Monte Carlo or molecular dynamics simulations.

While these techniques have alleviated the sampling problem, a number of difficulties remain. Most prominent here are simulations of proteins with explicit water. This is because in replica exchange the probability for an exchange between two temperatures decreases not only with the temperature difference ΔT between two replicas but also with the number of degrees of freedom N. Hence, because of the large number of water molecules needed in protein simulations, the temperature intervals ΔT have to be chosen small, and therefore a large number M of replicas is needed to cover the range between the temperature of interest (the lowest one) and the highest temperature which should correspond to the largest relevant barrier in the system. On the other hand, the number of round trips between lowest and highest temperature, and back, defines a lower bound on the number of independent configurations sampled at the lowest temperature (i.e. the one of interest). However, the number of round trips decreases as $\sqrt{M}$ with the number of replicas M. As a consequence, protein simulations with explicit water do not only require a large number of replicas but also long simulation times for each replica in order to reach equilibrium and obtain sufficient statistics.

In a recent brief communication [17] Nadler and Hansmann suggested to circumvent this problem of low acceptance rate and resulting large number of replicas through use of a novel microcanonical replica exchange method that is rejection free, and therefore optimizes the flow along the temperature ladder. Molecular dynamics simulations are usually done in the canonical ensemble ($T = const$) instead of a microcanonical ensemble ($E = const$). One reason is that the canonical ensemble is often more closer to the experimental settings (albeit not always, constant energy surface simulations are of interest in their own right [190, 191], e.g. for comparison with recent molecular beam experiments [192]). The other reason is that integration errors can accumulate in microcanonical molecular dynamics and easily lead to numerical instabilities and uncontrolled behavior; the use of a thermostat usually washes out the effect of these errors. Our assumption is that these integration errors are also averaged out in microcanonical replica exchange molecular dynamics through the exchange moves and velocity re-weighting. As it is possible in principle to connect back from a microcanonical ensemble to the canonical ensemble, the rejection-free microcanonical replica exchange molecular dynamics becomes a promising alternative in cases such as simulations in explicit solvent that otherwise suffer from low acceptance rates.

The purpose of the present work is to test the suitability of this idea in a practical application. We have chosen as our system the trp-cage protein [193, 194] as it has become a common model to test numerical methods [195, 181]. As the present work describes a proof-of-concept study, we simulate the molecule with an implicit solvent allowing for a faster evaluation of our approach. In the following section we first describe our method in detail before presenting our results. We finally discuss possible applications and modifications of our approach.

7.2 Methods

7.2.1 Statistical physics of microcanonical molecular dynamics

In microcanonical molecular dynamics the equations of motion are solved numerically for a particular system, generating states of constant energy E for that system. Assuming ergodicity, the hypersurface of states with constant energy E is connected and all states on the constant energy hypersurface are sampled uniformly. For observables that depend only on kinetic and potential energy, $M(E_{pot}, E_{kin})$, the microcanonical averages are given by

$$\begin{aligned}\langle M\rangle_E &= \frac{1}{|\Omega_{tot}(E)|}\times \\ &\quad \int dE_1 \int dE_2\, \delta\left(E - E_1 - E_2\right)\times \\ &\qquad \Omega_{pot}(E_1)\Omega_{kin}(E_2)M(E_1,E_2) \quad , \qquad (7.1)\end{aligned}$$

with $\Omega_{pot}(E)$ and $\Omega_{kin}(E)$ being the respective densities of states for the potential energy and for the kinetic energy; the total state space volume of the energy shell at E is used as normalization

$$\begin{aligned}|\Omega_{tot}(E)| &= \int dE_1 \int dE_2\, \delta\left(E - E_1 - E_2\right)\times \\ &\qquad \Omega_{pot}(E_1)\Omega_{kin}(E_2) \quad . \qquad (7.2)\end{aligned}$$

Usually we are interested in canonical averages, i.e.

$$\langle M\rangle_\beta = Z^{-1}\int dE\, M(E)\Omega_{pot}(E)e^{-\beta E} \quad , \qquad (7.3)$$

with β the inverse canonical temperature, and the partition function Z is used as normalization,

$$Z = \int dE\, \Omega_{pot}(E)e^{-\beta E} \quad . \qquad (7.4)$$

In order to evaluate such properties from microcanonical simulations, we need to estimate the density of states for the potential energy $\Omega_{pot}(E)$ from them. Since the distribution of potential energies observed in a microcanonical simulation is given by

$$\begin{aligned} P(E_{pot};E) &\propto \int dE_1 \int dE_2\, \delta\left(E - E_1 - E_2\right) \times \\ &\qquad \Omega_{pot}(E_1)\Omega_{kin}(E_2)\delta\left(E_{pot} - E_1\right) \\ &= \Omega_{pot}(E_{pot})\Omega_{kin}\left(E - E_{pot}\right) \quad , \end{aligned} \tag{7.5}$$

the density of potential energies has to be separated from the kinetic energy part. This is straightforward as the kinetic energy is given by

$$E_{kin} = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i}, \tag{7.6}$$

with $\mathbf{p}_i$ the momentum vector and m_i the mass of atom or group i; the density of states for the kinetic energy therefore can be determined analytically:

$$\Omega_{kin}(E) \propto E^{\frac{3N-f-2}{2}} \quad , \tag{7.7}$$

where f counts the constraints on the system (i.e. the true number of degrees of freedom is not $3N-2$ but reduced by f) Hence, up to the normalization constant the distribution of potential energies observed in a microcanonical simulation is given by

$$P(E_{pot};E) \propto \Omega_{pot}(E_{pot})\left(E - E_{pot}\right)^{\frac{3N-f-2}{2}} \tag{7.8}$$

Since both functions on the *rhs* grow strongly with their arguments, $P(E_{pot};E)$ is a sharply peaked function. Consequently, microcanonical averages of the energies are given by the most probable value, e.g.

$$\left\langle E_{pot}\right\rangle_E \approx \hat{E}_{pot} \quad , \tag{7.9}$$

note that $\hat{E}_{pot}+\hat{E}_{kin}=E$ holds. A saddle point approximation of Eq. (7.8) leads to the well-known relation between kinetic energy and microcanonical temperature ($\beta_E = 1/k_B T_E \equiv d\ln\Omega_{pot}(E)/dE$)

$$\hat{E}_{kin} = \frac{M}{2}T_E \quad , \tag{7.10}$$

where $M = 3N-f-2$ is the number of degrees of freedom in the system, and one obtains:

$$\begin{aligned} P(E_{pot};E) \quad \propto \quad & \Omega_{pot}(E_{pot})\exp\Bigg\{-\beta_E E_{pot} \\ & +\beta_E^2\left(\frac{E_{pot}-\hat{E}_{pot}}{\hat{E}_{kin}}\right)^2 \\ & +O\left[\beta_E^3\left(\frac{E_{pot}-\hat{E}_{pot}}{\hat{E}_{kin}}\right)^3\right]\Bigg\} \quad . \end{aligned} \tag{7.11}$$

Therefore, to leading order, the microcanonical energy distribution is given by the Boltzmann distribution, with the canoncial temperature equal to the microcanonical temperature.

7.2.2 Microcanonical replica exchange

In canonical replica exchange [196, 197, 62] two configurations with energies E_1 and E_2, sitting at temperatures T_1 and T_2, are exchanged with probability $\exp(\Delta\beta\Delta E)$, with the inverse temperature $\beta = 1/k_B T$. In microcanonical replica exchange one uses that

$$E(x,v) = E_{pot}(x) + E_{kin}(v) \tag{7.12}$$

with

$$E_{kin}(v) = \frac{1}{2}\sum_i m_i v_i^2 \tag{7.13}$$

where the potential energy E_{pot} depends only on the coordinates x, and the kinetic energy E_{kin} solely on the velocities v. Scaling all velocities by a factor r therefore changes the kinetic energy by:

$$E_{kin}(rv) = r^2 E_{kin}(v) \tag{7.14}$$

Hence, assuming $E^{(1)} < E^{(2)}$ and choosing suitable scaling parameters r_1 and r_2, one can exchange the two configurations with probability one:

$$\begin{aligned} E^{(1)}(x^{(1)}, v^{(1)}) &= E_{pot}(x^{(1)}) + E_{kin}(v^{(1)}) \\ \longrightarrow E^{(2)}(x^{(1)}, r_1 v^{(1)}) &= E_{pot}(x^{(1)}) + E_{kin}(r_1 v^{(1)}) \\ &= E_{pot}(x^{(1)}) + r_1^2 E_{kin}(v^{(1)}) \end{aligned} \tag{7.15}$$

and

$$\begin{aligned}
E^{(2)}(x^{(2)},v^{(2)}) &= E_{pot}(x^{(2)})+E_{kin}(v^{(2)}) \\
\longrightarrow E^{(1)}(x^{(2)},r_2v^{(2)}) &= E_{pot}(x^{(2)})+E_{kin}(r_2v^{(2)}) \\
&= E_{pot}(x^{(2)})+r_2^2E_{kin}(v^{(2)})
\end{aligned} \tag{7.16}$$

where the two rescaling factors r_1 and r_2 are given by

$$r_{1,2} = \sqrt{\frac{E^{(2),(1)}-E_{pot}(_{1,2})}{E^{(1),(2)}-E_{pot}(_{1,2})}} \quad . \tag{7.17}$$

Such moves are possible for $E_{pot}(_2) < E^{(1)}$, a restriction that does not violate detailed balance. On the other hand, ergodicity is ensured because of the regular microcanonical molecular dynamics between exchange moves. The acceptance probability for an allowed move is always one, since both weight functions are constant.

7.2.3 Technical Details and Setting

We test the efficiency of this microcanonical replica exchange molecular dynamics in all-atom simulations of the 20-residue trp-cage miniprotein which has become a commonly used test system for evaluation of new sampling schemes. The AMBER9 package is used with the ff99SB forcefield, approximating the interaction between protein and surrounding solvent by the Generalized Born implicit solvent. 18 replicas are used with the total energies - and corresponding temperatures - given in Table 7.1. After generating linear configurations with the module xLEAP, and minimizing these with 500 steps of steepest descent followed by another 500 steps of conjugate gradient, we heat the molecule to the respective target temperatures of Table 7.1. Here, and in the canonical replica exchange simulations with that we compare our results, we use SHAKE and a Berendsen thermostat for temperature control (coupling constant 1.0 ps). The resulting 18 structures serve as

Energy Shell	**Total Energy (kcal/mol)**	***T* (K)**
E1	-368.5	250
E2	-360.9	260
E3	-340.0	273
E4	-311.2	290
E5	-271.7	315
E6	-252.3	325
E7	-223.2	350
E8	-192.2	373
E9	-151.2	393
E10	-119.8	413
E11	-81.9	433
E12	-46.9	450
E13	-15.2	473
E14	19.2	493
E15	51.6	513
E16	91.3	533
E17	130.3	555
E18	184.3	580

Table 7.1: 18 replicas and their corresponding total energies and temperatures used in our simulations [16].

our initial starting configurations for both microcanonical and canonical replica exchange molecular dynamics simulations. Each structure consist of 304 atoms; however, the number of degrees of freedom is not $3N - 2 = 910$ but 757 as SHAKE constraints the length of certain bonds. For each algorithm, we perform runs of 15 ns, with an exchange move attempted every 5ps. We had written an external driver script for the replica exchange scheme. Only the last 10 ns are used for analysis.

7.3 Results

The inherent roughness of protein free energy landscapes leads to slow sampling at low temperatures (or in microcanonical simulations at low energies). In order to demonstrate this sampling problem, we have performed for our test system canonical molecular dynamics runs at $T = 250$ K, and microcanonical molecular dynamics at the corresponding energy

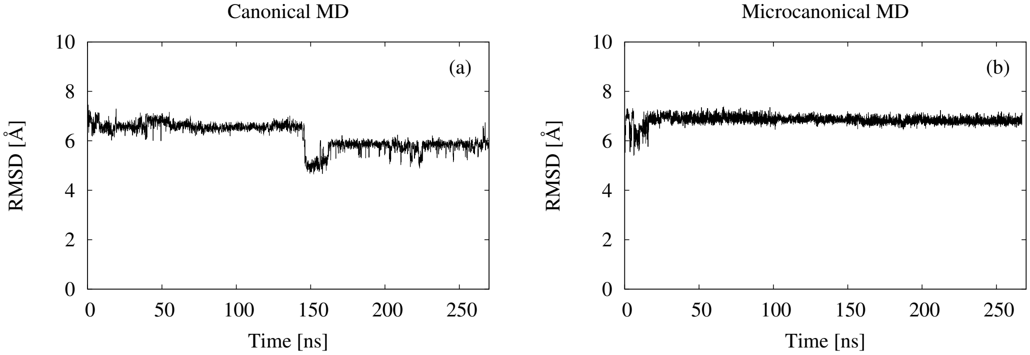


Figure 7.1: Root-mean-square deviation (rmsd) to the experimentally determined structure as function of time for (a) a canonical molecular dynamics simulation at $T = 250$ K, and (b) a microcanonical molecular dynamic simulation at the corresponding energy $E_{tot} = -368.5$ kcal/mol. [16]

($E_{tot} = -368.5$ kcal/mol). The two runs are over a time of 270 ns which corresponds to the total effort in the replica exchange simulations (18×15 ns). In Figure 7.1 we show as function of time the root-mean-square deviation (rmsd) of the actual configuration to the experimentally determined one (Protein Data Bank Id: 1L2Y). Over the whole length of the simulation, the rmsd is around or larger than 6 Å indicating that the simulations never thermalized and got stuck in local minima structurally very different from the native configuration.

A common approach to overcome this sampling problem is parallel tempering, also known as replica exchange sampling [196, 197, 62]. In Fig. 7.2 we display the resulting time series of rmsd at $T = 250$ K from a replica exchange simulation of the trp-cage protein with the temperature distribution given by Table 7.1. As in the canonical and microcanonical runs of Fig. 7.1 the rmsd starts at around 7 Å, indicating a starting configuration very different from the native one. However, the replica exchange sampling process leads soon to configurations that are within 3 Å rmsd, and therefore similar to the experimentally determined structure. Our results are comparable to the ones obtained by Simmerling et al [194] who have performed 50 ns long all-atom, fully unrestrained folding simulation of this protein at 325 K in implicit GB solvent [118, 117] using the AMBER ff99SB force field [198]. Without showing data we also remark that the transition temperature of ≈ 413 K (see also Fig. 7.9) is comparable to the melting temperatures of ≈ 400 K found by Pitera and Swope [199]. Albeit diverging from the experimentally determined transition temperature of 315 K [200], both results show that our data are comparable with previous simulations

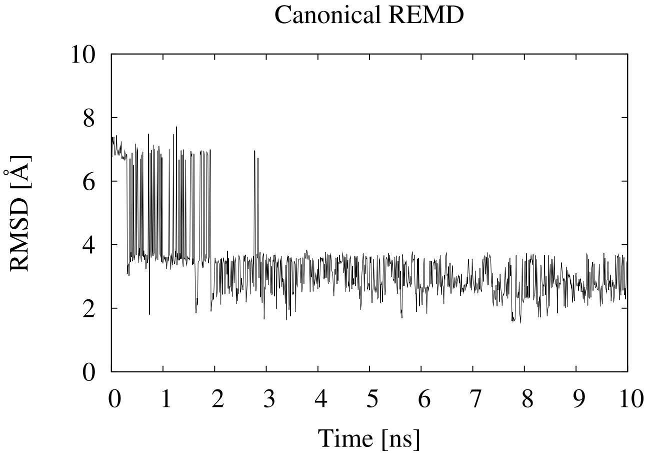


Figure 7.2: Root-mean-square deviation (rmsd) to the experimentally determined structure as function of time. The data are from a canonical replica exchange simulation with a temperature distribution given in table 7.1, and measured at $T = 250$ K [16].

relying on the Amber force field and an implicit solvent.

The reason for the enhanced sampling of low energy configurations are the excursions to high temperatures that allow a replica to escape from local minima. As an example, in Fig. 7.3a we show this walk through temperature space for one of the 18 replicas. A lower limit for the number of independent structures observed at lowest temperature $T = 250$ K is the number of round trips between this temperature and the highest temperature (in our case, $T = 580$ K), and back. In our example, only one such round trip is observed, and only a total of three round trips for all replicas together. The difficulty in ensuring a sufficient number of round trips (and therefore sufficient statistics), especially for the case of protein simulations in explicit solvent, has been described in the introduction, and is the starting point for our investigation. Our proposed new algorithm replaces a replica exchange in temperature by an exchange of replicas between different energy levels in microcanonical

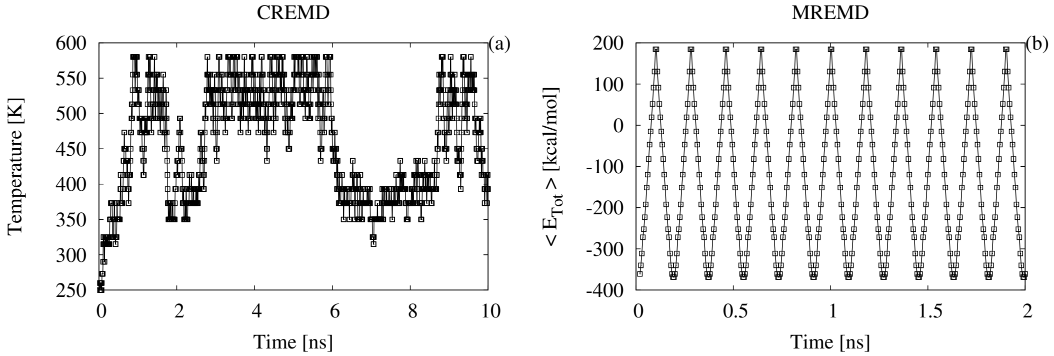


Figure 7.3: Walk of a specific replica (a) through temperature in a canonical replica exchange molecular dynamic simulation (CREMD); and (b) through energy in a microcanonical replica molecular dynamics simulation (MREMD) with the updates proposed in Ref. [17]. Note that he large number of roundtrips observed for the later case allowed us only to show a short segment of the 10ns run [16].

molecular dynamics. As the exchange move is rejection-free, it leads to much faster round trip times. This can be seen also in Fig. 7.3b where we show the walk of one replica through energy space. Note that the various energy levels correspond to the temperatures of the canonical replica exchange run, and are also listed in Table 7.1. Because of the large number of round trips we could show here only 2 ns of the 10 ns long run, for otherwise the figure would no longer be readable.

However, while the microcanonical replica exchange molecular dynamics method (MREMD) of Ref. [17] leads to a 50-fold decrease in round trip times when compared to the canonical replica exchange molecular dynamics method (CREMD), this gain in efficiency does not translate into improved sampling. This is obvious from Fig. 7.4 where we plot the average radius of gyration $< r_{gy} >$ as function of temperature T when calculated from the canonical replica exchange molecular dynamics; and as function of the corresponding total energies when calculated from the microcanonical replica exchange molecular dynamics.

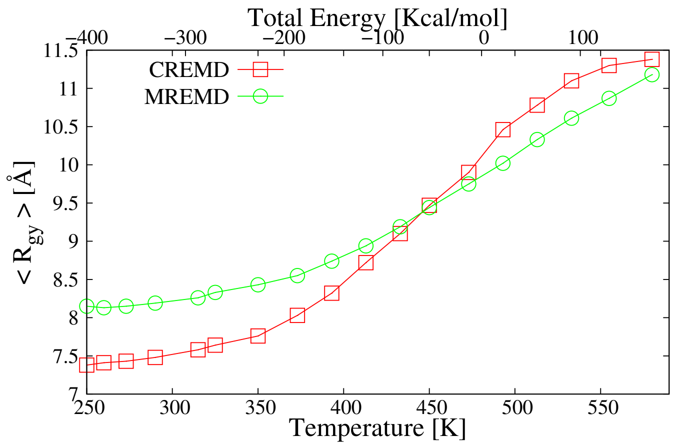


Figure 7.4: Radius of gyration $< r_{gy} >$ as function of temperature in a canonical replica exchange molecular dynamics simulation (CREMD); and the corresponding energy levels in a microcanonical replica molecular dynamics simulation (MREMD) with the updates proposed in Ref. [17] [16]

As a measure for the compactness of protein structures and its change this quantity indicates structural transitions. Clearly, the two curves differ considerably. Together with similar behavior for other physical quantities (data not shown) the difference between the two curves indicates sampling problems in the new approach.

The difference between the two simulations is puzzling as the microcanonical replica exchange method is formally correct, and therefore should yield the same results as the canonical replica exchange. Hence, this difference indicates that despite the increased flow through temperature space the sampling is still slower than in the canonical case, not faster as was expected.

A fundamental assumption behind the idea of optimizing a replica exchange simulation through maximizing the flow through temperature space is that relaxation at a given temper-

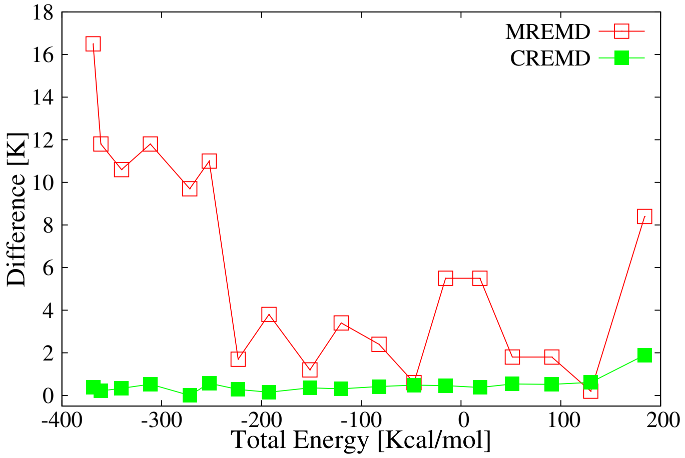


Figure 7.5: Difference between canonical temperature, see Table 7.1, and microcanonical temperature, calculated from the kinetic energy via Eq. (7.10), as function of total energy. The figure shows this quantity as measured in canonical replica molecular dynamics simulations (CREMD) as well as in the microcanonical replica molecular dynamics simulation (MREMD) with updates proposed in Ref. [17] [16].

ature is fast compared with the time scale of flow through temperatures. In the present case this seems not to be the case. An indicator for this lack of kinetic energy equilibration is the difference between the microcanonical temperature and the canonical temperature. We have plotted this quantity in Fig. 7.5 as a function of total energy, comparing data from the canonical replica exchange molecular dynamics simulation with those from microcanonical replica exchange approach. While the temperature difference fluctuates around zero for the canonical run, it differs strongly in the case of microcanonical replica exchange. Hence, the assumptions behind Eq. 7.8 and Eq. 7.11, do not hold on the time scales of our simulations. The equivalence can be expected to be restored for very long simulation times, see, for instance, in Fig. 7.6 the time evolution of the frequency of native-like configurations with simulation time; however, the required long simulation times would defy

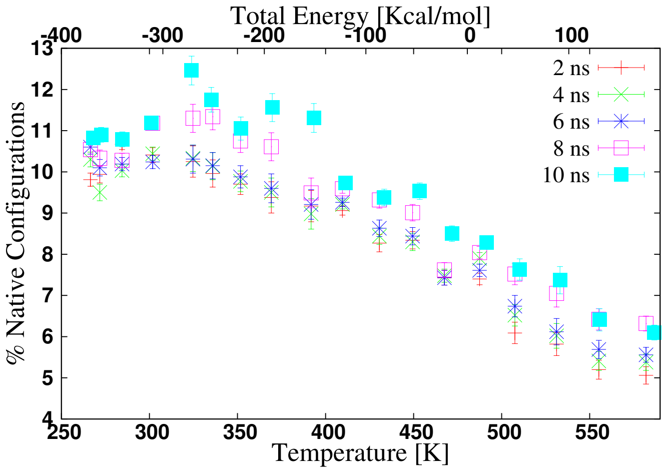


Figure 7.6: Frequency of native-like configurations (rmsd $<$ 3.7 Å) as function of simulation time as measured in a microcanonical relpica molecular dynamics simulation with the updates proposed in Ref. [17] [16]

.

the purpose of our investigation. In order to overcome this bottleneck one can think of two approaches. The microcanonical replica exchange molecular dynamic leads for finite times to quasi cyclic motions in phase space. Introducing randomness in the system will destroy these deterministic motions and allow for sampling of a wider area in phase space. One possibility to introduce this randomness is by periodic refreshing of the velocities at the highest energy shell. This is justified as the underlying assumption of replica exchange methods is that a given replica can cross any relevant barrier, and therefore looses history, once it reaches the highest temperature/energy. As our data show, this is not the case in the microcanonical replica exchange molecular dynamics (MREMD), but can be enforced by such randomization of velocities at this energy shell. We call this version randomized microcanonical replica exchange molecular dynamics (RMREMD). By the definition of the method, the walk of replicas through the various energy shells for RMREMD is still

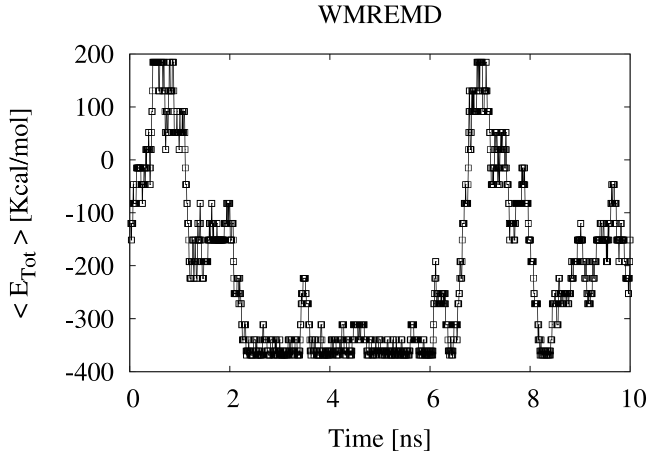


Figure 7.7: Walk of a specific replica through energy in a microcanonical replica molecular dynamics with trial of exchange moves given by Eq. 7.8[16]

deterministic, and does not differ from that of the original method (MREMD), displayed in Fig. 7.3b.

A second possibility to introduce randomness in the motion is by way of the replica exchange move, i.e. giving up the rejection-free exchange moves in microcanonical replica exchange molecular dynamics. A possible approach is to enforce validity of Eq. 7.8 by exchanging replicas between energy shells according to this distribution. We name this version of our approach weighted microcanonical replica exchange molecular dynamics (WMREMD). The resulting random walk through the energy shells is displayed in Fig. 7.7. We have performed simulations of both variants with same statistics as in the case of canonical replica exchange molecular dynamics and the original version of microcanonical replica exchange molecular dynamics. For a comparison of the various methods we show in Fig. 7.8 the percentage of native-like configurations for all four methods. Note the difference between the original MREMD and RMREMD on one side, and canonical replica

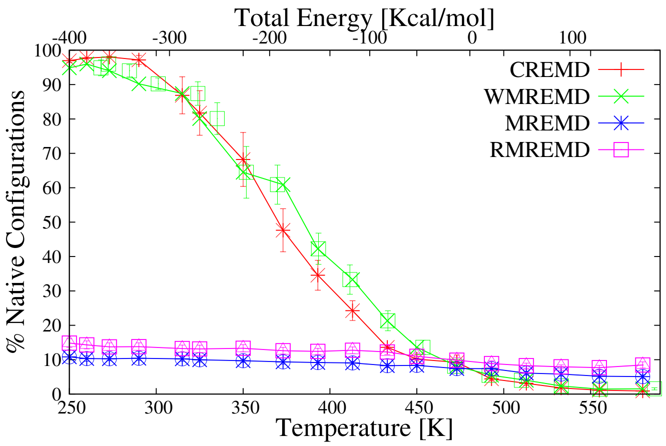


Figure 7.8: Frequency of configurations with a rmsd smaller than 3.7 Å as measured in canonical (CREMD) and various versions of microcanonical replica exchange molecular dynamics [16].

exchange molecular dynamics and WMREMD on the other side. While the first two lead at lowest energy (and corresponding temperature) to less than 20% of native-like structures, the weighted microcanonical replica exchange molecuar dynamics leads essentially to the same frequency as the canonical replica exchange molecular dynamics, i.e about 90% of native like configurations. However, while the data in canonical replica exchange molecular dynamics rely on solely 3 round trips, WMREMD let to 9 round trips, i.e. three times higher statistics.

So far, our investigation has shown that the weighted, i.e. modified, microcanonical replica exchange molecular dynamics (WMREMD) leads to correct averages, and exhibits an at least three times faster sampling than canonical replica exchange molecular dynamics. Having demonstrated the improved sampling, we want to show now how this allows us to study in detail the thermodynamics of the trp-cage protein. In Fig 7.9 we show the

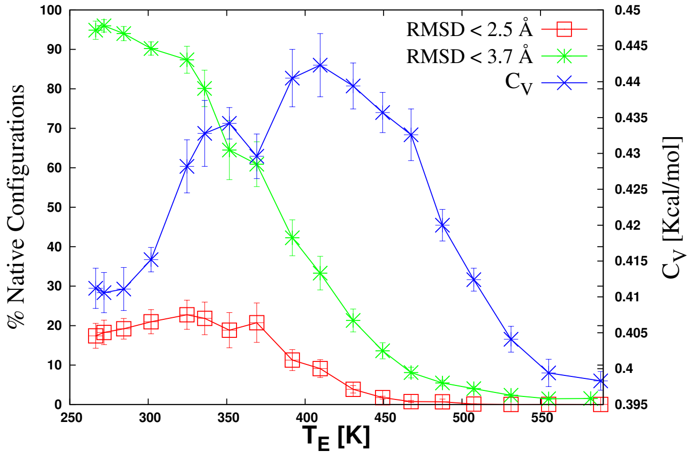


Figure 7.9: Frequency of native-like configurations as measured according to two criteria (see text), and specific heat capacity, as measured in simulations with our weighted microcanonical replica exchange molecular dynamics [16].

frequency of configurations with a rmsd smaller than 3.7 Å and those with rmsd smaller than 2.5 Å. Approaching from high energies (temperatures) a critical energy (temperature) of ≈ -120 kcal/mol (corresponding to $T \approx 413$ K), the frequency of configurations with rmsd smaller than 3.7 Å increases dramatically, and stays constant after approaching its maximum. On the other hand, configurations with rmsd smaller than 2.5 Å, i.e. those very close to the experimentally determined one, also first increase rapidly, but decrease again after reaching its maximum value at ≈ -272 kcal/mol ($T \approx 315$ K). Note that the increase in both curves is correlated with the position of the peak in specific heat capacity

$$C = \frac{< E_{pot}^2 > - < E_{pot} >^2}{k_B T^2} \tag{7.18}$$

shown also in Fig. 7.9. The decrease observed for the frequency of configurations with rmsd smaller than 2.5 Å seems to be correlated with a shoulder in the specific heat capacity

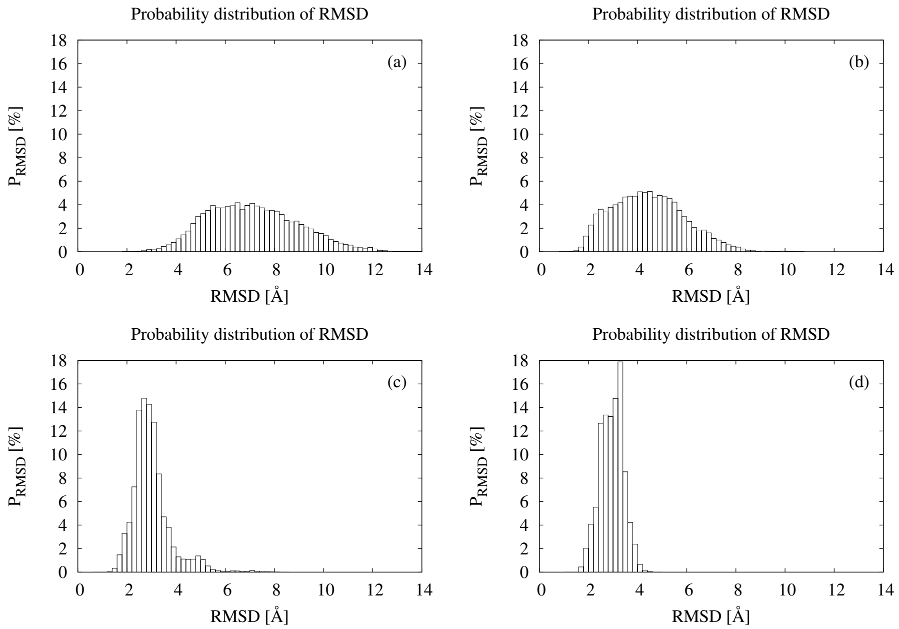


Figure 7.10: Histograms of configurations as function of rmsd calculated for four different energy levels [16].

curve. A natural interpretation for the steep increase in native-like configuration (according to both definitions) and the peak in specific heat capacity is that of a folding transition. The decrease in frequency of configurations, whose similarity to the native structure is measured according to the more stringent criteria of an rmsd smaller that 2.5 Å, requires a more detailed analysis. For this purpose, we show in Fig. 7.10 histograms of configurations as function of rmsd for four values of energy. At the highest energy shell (184.3 kcal/mol, Fig. 7.10a) we observe a broad single-peaked distribution centered around a rmsd of $\approx 6-7$ Å, indicating that at this energy (and corresponding temperature) configurations have little resemblance with the native structure. The distribution shown in (b) is drawn for $E_{tot} = -119.8$ kcal/mol, the energy level corresponding to the peak in specific heat capacity. Here, we find a distribution centered around a rmsd of ≈ 4 Å that covers both structures with large rmsd and such that resemble the native one (small rmsd). Hence at this energy level, which corresponds to a microcanonical folding temperature of ≈ 413 K, we have an equilibrium of unfolded and folded configurations (these with rmsd smaller than 3.7 Å). The third distribution (Fig. 7.10c) is calculated for $E_{tot} = -271.7$ kcal/mol ($T = 315$ K), i.e. the position of the shoulder in specific heat capacity and maximum of the curve in Fig. 7.7 that displays the frequency of configurations with rmsd smaller than 2.5 Å. Again, we observe a single peaked distribution centered around ≈ 3 Å that is almost exclusively made up of native-like structures (such with a rmsd smaller than 3.7 Å). Surprisingly, this distribution does not become narrower when going to the lowest energy level $E_{tot} = -368.5$ kcal/mol, nor does its center moves to smaller values of rmsd. Instead, the distribution becomes double-peaked with one peak around a rmsd of ≈ 2.5 Å, and the second and larger one centered around a rmsd of 3.3 Å, indicating an equilibrium between configurations with rmsd around and smaller than 2.5Å, and such with rmsd between 3 Å and 4 Å. An example for both types of configurations is shown in Fig. 7.11.

In connection with Fig. 7.9 we interpret the series of histogram as follows. At temperature of ≈ 413 K we have a folding transition that separates unfolded configurations from an ensemble of configurations that are to similar to the native structure. This ensemble is made up of two clusters of structures shown in Fig. 7.10. Both configurations are stabilized by a salt bridge between ASP9 and ARG16 that is responsible for the fast folding kinetics of this protein. Decreasing the temperature further the frequency of the configurations of Fig. 7.11a increases. The overlay with experimentally determined structure

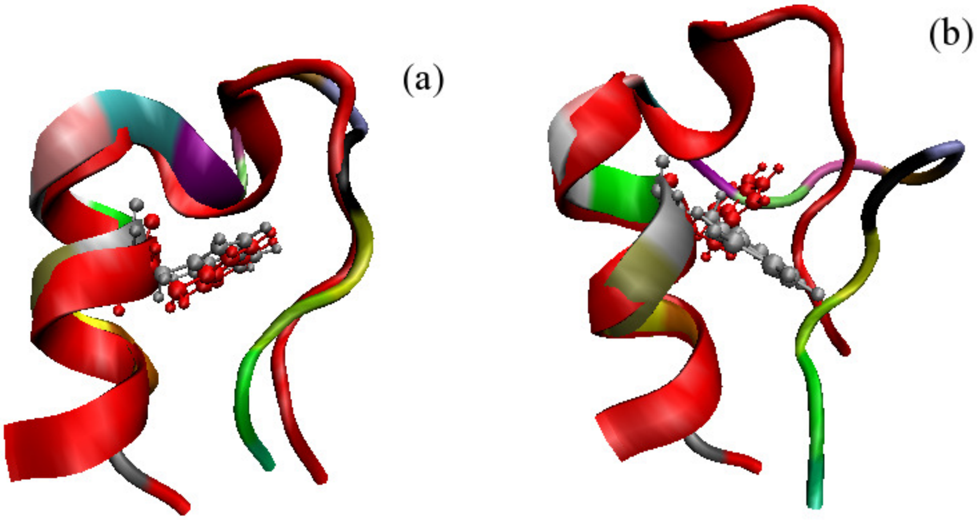


Figure 7.11: The two dominant low-energy structures (color), shown in overlay with the native structure (red) [16].

emphasizes how closely the configuration resemble the native structure (≈ 2 Å), not only in the backbone but also in the orientation of the tryptophan side chain. However, below a certain temperature, the frequency of configurations of this type decreases again, and dominant now are the slightly different configurations of Fig. 7.11b. These configurations differ from the native structure by rmsd about $3-4$ Å and are characterized by a wrongly positioned tryptophan side chain and divergent backbone orientation at residue 9 that leads to this structure. Unlike in the native structure, the chain terminals are connected by hydrogen bonds that energetically favors this structure over the native form. The increase in frequency of these structure in lieu of the native one with decreasing temperature may indicate limitations in the accuracy of our energy function (see also Ref. [201]), but could also indicate a partial "cold unfolding". In the later case this would demonstrate again the well-known fact that the native state of a protein is the global minimum in free energy at physiological temperatures, but not necessarily the global minimum in potential energy.

7.4 Conclusions

We have tested a recently proposed microcanonical replica exchange molecular dynamics approach in simulations of the trp-cage protein in implicit solvent. We evaluated the performance of this method, and introduced a variant that lead to improved sampling for this protein. Using this new sampling technique we could not only find the native structure of this protein within 2 Å rmsd, but also show that the folding thermodynamics of this protein is surprisingly rich, with not only a folding transition but also indications for a partial cold unfolding.

Chapter 8

Summary & Future Directions

In this doctoral dissertation, an enhanced implicit solvation model based on the Poisson-Boltzmann equation within boundary element method (PB/BEM) framework has been studied. We also discuss an enhanced sampling method to study the protein folding problem and its application to a Trp-cage protein in an implicit solvent.

Following the quantum mechanical Polarizable Continuum Model [46], the net solvation free energy is decomposed into three distinct physics-based terms: *polarization*, *dispersion* and *cavitation*.

$$\Delta G_{net} = \Delta G_{pol} + \Delta G_{disp} + \Delta G_{cav} \tag{8.1}$$

The cavitation term is obtained via the revised Pierotti approximation (rPA) [11, 59, 60, 61]. The polarization free energy is estimated by solving the Poisson-Boltzmann equation

[59, 50, 49, 54]. The dispersion term is handled either via the Caillet-Claverie [57, 58] approach or a revised Lennard-Jones formulation using AMBER [18] parameters.

Each term is treated individually and parameterized independently to get the maximum level of accuracy. We investigate the influence of surface type and surface resolution and dependence on atomic model parameters, such as van der Waals radii & partial atomic charges [174]. Our study shows that an error on the order of 40 kcal/mol is introduced if one does not resolve the surface properly, and work in the nonconvergent domain [174]. Our investigation also reveals the fact that rather small-sized boundary elements (BEs) (0.3 $Å^2$) are needed to obtain consistently convergent polarization free energies ΔG_{Pol} [174]. Consideration of geometric factors revealed that when applying a scaling factor of about **1.07** to AMBER default van der Waals radii, a good agreement can be reached between the reference geometries (PCM results) and the geometries in the PB/BEM approach [174]. With this small BEs and slightly increased van der Waals radii and unchanged AMBER partial charge, we can achieve a good estimate of the polarization free energy ΔG^{pol} when compared to other studies in the literature [13, 178, 179]. This part of my work has been published in the Journal of Computational Chemistry (see Ref. [174]).

We systematically implemented the dispersion term using the Caillet-Claverie [57, 58] approach, and found it to offer good compromise between accuracy and computational overhead. Free parameters are determined by comparison to experimental data as well as high-level quantum mechanical reference (PCM) calculations. Our study shows that the Caillet-

Claverie dispersion coefficients should be multiplied by a scaling factor of 0.70 in order to achieve close matching with the experimental solvation free energies [12]. The model is tested on various chemical substances and found to yield good quality estimates of the solvation free energy without obvious indication of any introduced bias. We find that when substituting the Caillet-Claverie formalism with the corresponding classical Lennard-Jones term using AMBER [18] parameters, a readjustment of scaling factors (0.76 for water) is required [12]. Either approach is competitive and comes with its own merits. Caillet-Claverie coefficients are more general and specific to chemical elements only. On the other hand, employment of AMBER parameters appears to be straightforward in the present context since the geometry of the boundary is already based on AMBER van der Waals radii.

After determining appropriate scaling factors for different solvents (e.g., water, methanol, ethanol, n-octanol, cyclohexane etc.), we applied our model to a series of proteins of increasing size and analyzed the relative contribution of the individual term as a function of system size. Moreover, we have carried out semi-empirical calculations on the same series of proteins, and compared effects resulting from different charge assignments to each other. Our investigations show that the replacement of static AMBER partial charges with semi-empirical PM5 charges introduces a rise in solvation free energy by about 20% of the classic results regardless of the size or total charge state of the systems [12]. The polarization term constiutes the major contribution, but apolar terms are far from negligible (see Fig. 5.5). This work on the dispersion free energy has been published in the Journal of Physical Chemistry B (see Ref. [12]).

Our study suggests that slightly larger BEs (0.45 Å^2) in comparison to BEs used in calculation of the polarization term (0.31 Å^2) could be used for the computation of the dispersion term without loosing any accuracy [15]. The best performance is achieved when the DIIS (Direct Inversion of the Iterative Subspace) matrix is dimensioned 7×7. We also find that the number of iterations necessary to achieve the convergence does not depend on the system size. These results have been published in a conference proceeding (see Ref. [15]).

Once optimized, the solvation model is employed to estimate the electrostatic potential (ESP) map of an anti-fungal protein (PDB code: 1P9G). It becomes clear that the major appearance of the protein in an aqueous solution is that of a macroscopic particle of largely positive ESP; hence the tendency to migrate to the membrane can be explained straightforwardly [183]. We compared ESP maps based on classic charge assignments using AMBER parameters [146] with ESP maps resulting from semi-empirical charges computed with program LocalSCF [19] at several levels of semi-empirical theory, ie AM1, MNDO, PM3 and PM5. Our investigation reveals the fact that semi-empirical charge assignments deliver a consistent picture of significant differences seen for the charged residues. However, individual semi-empirical models differ considerably amongst each other. These findings are published in a conference proceeding (see Ref. [20]).

Development and implementation of a new variant of the regular replica exchange method (REMD) is described in this dissertation. The new sampling method is called as Microcanonical Replica Exchange Molecular Dynamics (MREMD). We study the folding thermodynamics of a Trp-cage mini-protein in an implicit solvent using MREMD simulation protocol. Although this exchange scheme is rejection free, it leads to slower sampling compared to the regular REMD simulation. To circumvent this problem we give up the rejection-free scheme, and do importance sampling with the following weight function.

$$P(E_{pot};E) \propto \Omega_{pot}(E_{pot})\,(E-E_{pot})^{\frac{3N-f-2}{2}} \tag{8.2}$$

We call this variant of MREMD as Weighted Microcanonical Replica Exchange Molecular Dynamics (WMREMD). At lowest energy shell, the WMREMD method leads to the same frequency (90%) of native structure as the canonical REMD simulation. We show that the WMREMD performs three times more round trips compared to the canonical REMD simulation. This suggests that the WMREMD method samples faster than the regular canonical REMD, and yields better statistics compared to its canonical equivalent. Using this new sampling technique we could not only find the native structure of the Trp-cage protein within 2 Å rmsd, but also show that the folding thermodynamics of this protein is surprisingly rich, with not only a folding transition but also indications for a partial cold unfolding. This part of my work has been published in the journal Physical Review E (see Ref. [16]).

Our enhanced implicit solvation model has broad impacts in several areas of biomolecular simulations, such as (i) simulation of diffusional processes to determine ligand-protein and protein-protein binding kinetics [202, 203], (ii) molecular dynamics simulations of biomolecules in an implicit solvent [16, 199, 204, 205], (iii) titration studies of biomolecules

[206, 207], (iv) determining ligand-protein and protein-protein equilibrium binding constants required for rational drug design [202, 184, 185], (v) simulation of large biomolecules with increasing demands on high performance solutions [208], and (vi) simulation of complex environments of central biological importance [209]. Our newly developed algorithm samples faster compared to the canonical replica exchange molecular dynamics method. This means that the new sampling algorithm (WMREMD) will enable us to reliably study the folding of proteins of relatively larger size in an explicit solvent, which is currently prohibited due to the high computational demand.

Currently our enhanced implicit solvation model is valid for the solvation of proteins and organic molecules. Further work needs to be done to extend our model to study the solvation of nucleic acids (DNA, RNA). However, there are no fundamental restrictions that would preclude such an extension. In our model we solve the Poisson equation to obtain the electrostatic component of the solvation free energy. The effects of ions and salt are not captured implicitly in our approach. The Boltzmann term needs to be included into our model if we want to account for the charged background. Further parameterization as described for organic substances can be extended to nucleic acids in a straightforward way.

Most of the existing implicit solvation models have been developed for room temperature only. However, many simulation methods that optionally apply implicit models (e.g., Monte Carlo, molecular dynamics, parallel tempering, simulated annealing, etc) treat temperature as an adjustable system parameter. How such solvation terms will change with varying temperature needs to be addressed. One can think of using temperature-dependent dielectric constant $\varepsilon\ (T)$ in order to capture the temperature dependence in the polarization free energy term. The temperature dependence in the dispersion term can be introduced if we replace AMBER classic attractive 6-term of the Lennard-Jones potential with the London equation of dispersion [210]

$$E_{i,j}^{London,disp} = -\frac{3}{4}\alpha_i\alpha_j\frac{I_iI_j}{I_i+I_j}R_{i,j}^{-6} \tag{8.3}$$

where I is the ionization potential and α is the dipole polarizability. Here α is sensitive to temperature [211, 212, 213]. The cavitation free energy is obtained from the equation

$$\Delta G^{cav} = k_0 + k_1 r + k_2 r^2 \tag{8.4}$$

where r is the effective radius (Å) which can be derived from solvent excluded volume ($V^{exl.vol}$).

$$r = \left(\frac{3V^{exl.vol}}{4\pi}\right)^{1/3} \tag{8.5}$$

Mahajan et.al [56] have determined coefficients k_0, k_1, and k_2 for several discrete temperatures that may allow us to incorporate T-dependent cavitation term. Linear interpolation can be made to derive the appropriate coefficients for intermediate temperatures.

In WMREMD, all the exchange moves are not accepted. This limits the size of the protein that can be studied using our algorithm. Although the WMREMD samples faster than the canonical REMD, we still need to optimize the flow of replicas along the temperature/energy ladder for achieving faster equilibration, and reliably simulate larger proteins in an explicit solvent.

Appendix A

List of Publications

Following papers are included in this dissertation.

1. P. Kar, Y. Wei, U. H. E. Hansmann, S. Höfinger, Systematic Study of the Boundary Composition in Poisson Boltzmann Calculations, J. Comput. Chem., 28(16): 2538-2544 (2007)

2. P. Kar, M. Seel, U. H. E. Hansmann, S. Höfinger; Dispersion Terms and Analysis of Size- and Charge- Dependence in an Enhanced Poisson-Boltzmann Approach, J. Phys. Chem. B, 111 (2007) 8910

3. P. Kar, W. Nadler, U. H. E. Hansmann; Microcanonical Replica Exchange Molecular Dynamic Simulation of Proteins, Phys. Rev. E 80, 056703 (2009).

4. P. Kar, M. Seel, U. H. E. Hansmann, S. Höfinger. Comparing Semiempirical versus Classical Charge Assignments in Biomolecules and Their Effect on Electrostatic Potentials. NIC Publication Series, Vol. 36, 155-158 (2007)

5. P. Kar, M. Seel, U. H. E. Hansmann, S. Höfinger. Algorithmic refinements to an Enhanced Poisson-Boltzmann Approach Used in Biomolecular Simulations. NIC Publication Series, Vol. 36, 173-176(2007)

6. P. Kar, Y. Wei, U. H. E. Hansmann, S. Höfinger. The Influence of Molecular Surface Composition on the Outcome of Poisson-Boltzmann Calculations Performed to Obtain Solvation Free Energies. NIC Publication Series, Vol. 34, 205-209 (2006)

Appendix B

Hardware Used in My Research

I have used two clusters for my research purpose. They are NICole (jon von Neumann Institute for Computing, Forschungszentrum Jülich, Germany) and Hal (http://hal.phy.mtu.edu). The architecture of both clusters are discussed below.

NICole:

• Number of Processors: 384

• Overall peak performance: 1.6 Teraflops

• Operating System: SuSE Linux 10.1

• Cluster management: ParaStation

• Operating mode: batch (TORQUE/Maui)

• Main memory: 72 × 8 GB (aggregate 576 GB)

- Processortype : AMD Opteron, 2.6 GHz

- Network: Infiniband

- Disc capacity: 4 TB

Hal:

- Number of Processors: 64

- Number of nodes: 8

- Processortype: Intel Xeon E5405, 2.00 GHz

- Disc capacity: 0.5 TB

- Network: ethernet

- Cluster management: Sun Grid Engine (SGE)

- Memory: 8GB per node

Appendix C

List of Abbreviations and Symbols

† ACE: Analytical Continuum Electrostatics

† AGB: Analytical Generalized Born

† AMBER: Assisted Model Building with Energy Refinement

† BE: Boundary Element

† BEM: Boundary Element Method

† BF: Bernal-Fowler (an explicit water model)

† BNS: Ben-Naim Stillinger

† CHARMM: Chemistry at HARvard Molecular Mechanics

† CREMD: Canonical Replica Exchange Molecular Dynamics

† DFT: Density Functional Theory

† DNA: Deoxyribonucleic Acid

† ECEPP: Empirical Conformational Energy Program for Peptides

† ESP: Electrostatic Potential

† FDPB: Finite Difference Poisson-Boltzmann

† GB: Generalized Born

† GB/MV: Generalized Born/Molecular Volume

† GB/SA: Generalized Born/Surface Area

† GROMACS: GROningen MAchine for Chemical Simulations

† LPB: Linearized Poisson-Boltzmann

† MC: Monte Carlo

† MD: Molecular Dynamic

† MREMD: Microcanonical Replica Exchange Molecular Dynamic

† mRNA: messenger Ribonucleic Acid

† MSROLL: Molecular Surface ROLL

† NPB: Nonlinear Poisson-Boltzmann

† OPLS: Optimized Potential for Liquid Simulations

† PB: Poisson Boltzmann

† PB/BEM : Poisson Boltzmann/Boundary Element Method

† PCM: Polarizable Continuum Model

† PDB: Protein Data Bank

† REM: Replica Exchange Method

† REMD: Replica Exchange Molecular Dynamics

† RMREMD: Random Microcanonical Replica Exchange Molecular Dynamics

† RMSD: Root Mean Square Deviation

† RNA: Ribonucleic Acid

† SA: Simulated Annealing

† SASA: Solvent Accessible Surface Area

† SIMS: Smooth Invariant Molecular Surface

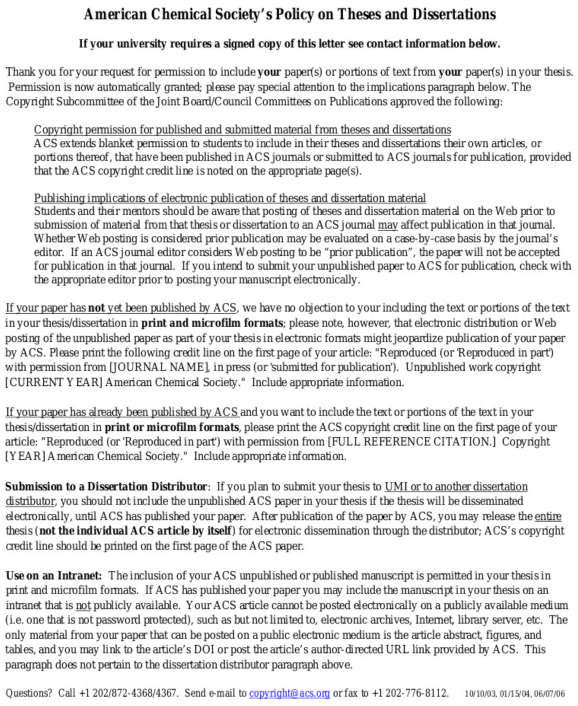
† SMMP: Simple Molecular Mechanics for Proteins

† VMD: Visual Molecular Dynamics

† WMREMD: Weighted Microcanonical Replica Exchange Molecular Dynamics

Appendix D

Copyright



American Chemical Society's Policy on Theses and Dissertations

If your university requires a signed copy of this letter see contact information below.

Thank you for your request for permission to include **your** paper(s) or portions of text from **your** paper(s) in your thesis. Permission is now automatically granted; please pay special attention to the implications paragraph below. The Copyright Subcommittee of the Joint Board/Council Committees on Publications approved the following:

> Copyright permission for published and submitted material from theses and dissertations
> ACS extends blanket permission to students to include in their theses and dissertations their own articles, or portions thereof, that have been published in ACS journals or submitted to ACS journals for publication, provided that the ACS copyright credit line is noted on the appropriate page(s).
>
> Publishing implications of electronic publication of theses and dissertation material
> Students and their mentors should be aware that posting of theses and dissertation material on the Web prior to submission of material from that thesis or dissertation to an ACS journal may affect publication in that journal. Whether Web posting is considered prior publication may be evaluated on a case-by-case basis by the journal's editor. If an ACS journal editor considers Web posting to be "prior publication", the paper will not be accepted for publication in that journal. If you intend to submit your unpublished paper to ACS for publication, check with the appropriate editor prior to posting your manuscript electronically.

If your paper has **not** yet been published by ACS, we have no objection to your including the text or portions of the text in your thesis/dissertation in **print and microfilm formats**; please note, however, that electronic distribution or Web posting of the unpublished paper as part of your thesis in electronic formats might jeopardize publication of your paper by ACS. Please print the following credit line on the first page of your article: "Reproduced (or 'Reproduced in part') with permission from [JOURNAL NAME], in press (or 'submitted for publication'). Unpublished work copyright [CURRENT YEAR] American Chemical Society." Include appropriate information.

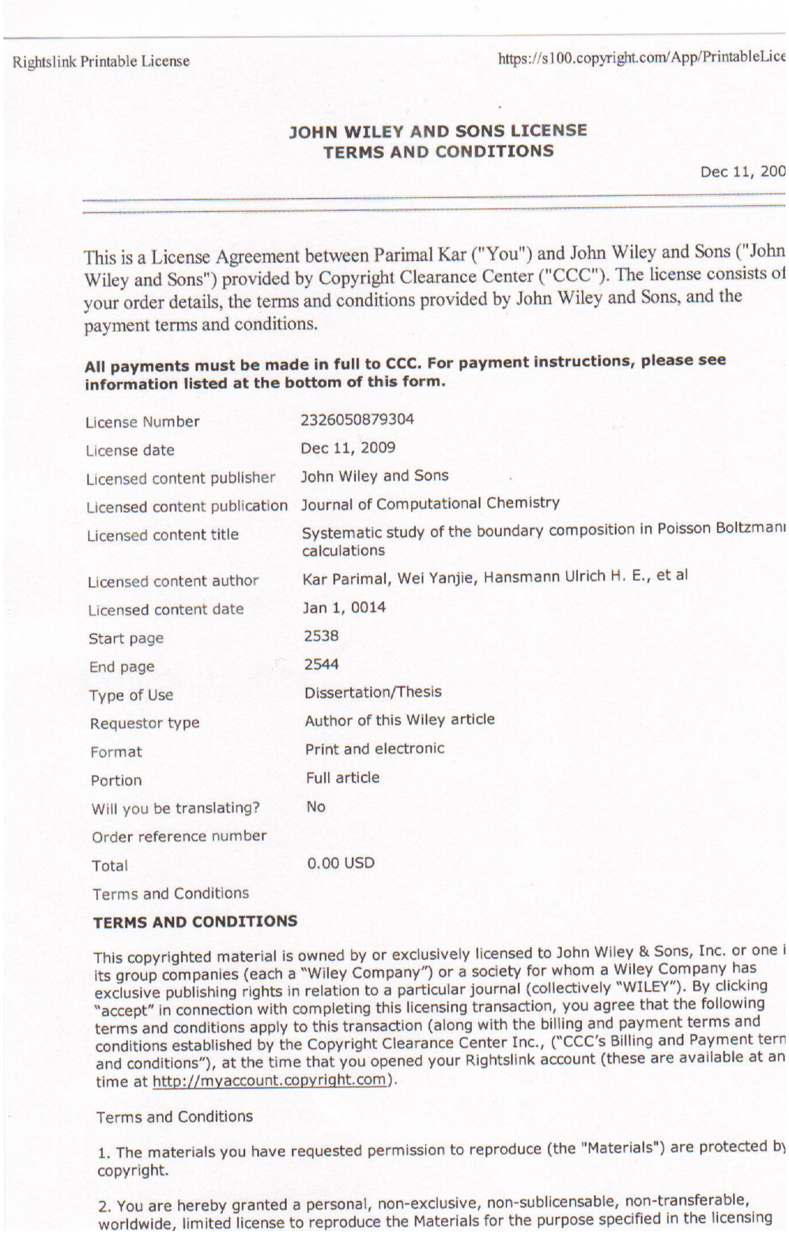
If your paper has already been published by ACS and you want to include the text or portions of the text in your thesis/dissertation in **print or microfilm formats**, please print the ACS copyright credit line on the first page of your article: "Reproduced (or 'Reproduced in part') with permission from [FULL REFERENCE CITATION.] Copyright [YEAR] American Chemical Society." Include appropriate information.

Submission to a Dissertation Distributor: If you plan to submit your thesis to UMI or to another dissertation distributor, you should not include the unpublished ACS paper in your thesis if the thesis will be disseminated electronically, until ACS has published your paper. After publication of the paper by ACS, you may release the entire thesis (**not the individual ACS article by itself**) for electronic dissemination through the distributor; ACS's copyright credit line should be printed on the first page of the ACS paper.

Use on an Intranet: The inclusion of your ACS unpublished or published manuscript is permitted in your thesis in print and microfilm formats. If ACS has published your paper you may include the manuscript in your thesis on an intranet that is not publicly available. Your ACS article cannot be posted electronically on a publicly available medium (i.e. one that is not password protected), such as but not limited to, electronic archives, Internet, library server, etc. The only material from your paper that can be posted on a public electronic medium is the article abstract, figures, and tables, and you may link to the article's DOI or post the article's author-directed URL link provided by ACS. This paragraph does not pertain to the dissertation distributor paragraph above.

Questions? Call +1 202/872-4368/4367. Send e-mail to copyright@acs.org or fax to +1 202-776-8112. 10/10/03, 01/15/04, 06/07/06

Figure D.1: ACS's copyright policy on theses and dissertation



Rightslink Printable License https://s100.copyright.com/App/PrintableLice

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Dec 11, 200

This is a License Agreement between Parimal Kar ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists ot your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2326050879304
License date	Dec 11, 2009
Licensed content publisher	John Wiley and Sons
Licensed content publication	Journal of Computational Chemistry
Licensed content title	Systematic study of the boundary composition in Poisson Boltzmani calculations
Licensed content author	Kar Parimal, Wei Yanjie, Hansmann Ulrich H. E., et al
Licensed content date	Jan 1, 0014
Start page	2538
End page	2544
Type of Use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Total	0.00 USD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one i its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment tern and conditions"), at the time that you opened your Rightslink account (these are available at an time at http://myaccount.copyright.com).

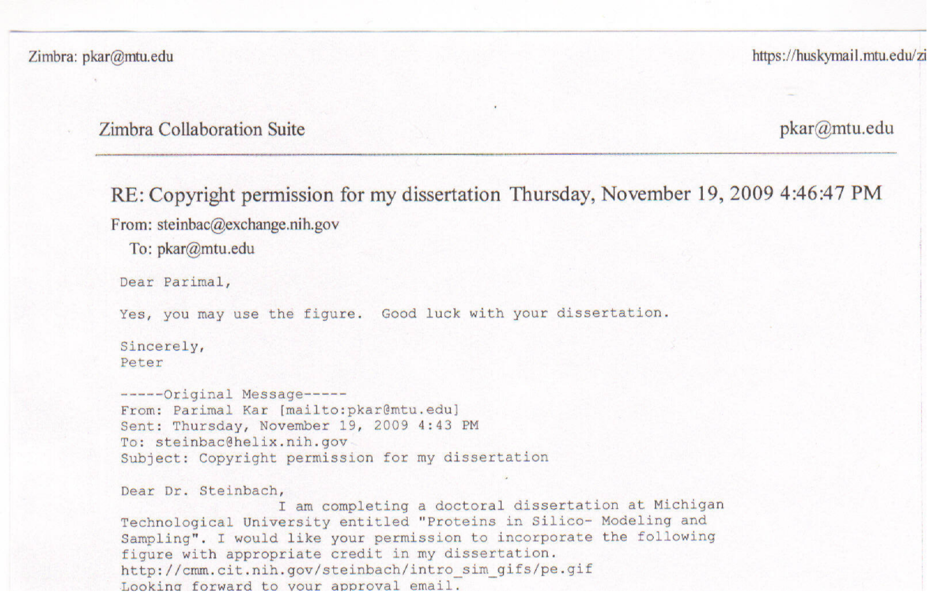
Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2. You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing

Figure D.2: Copy right permission letter from Wiley for Chapter 4.

Copyright notice for Fig 1.1, Fig 1.2, Fig 1.3, Fig 2.1, Fig 3.1: "I, the copyright holder of this work, hereby publish it under the following licenses: Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entittled "GNU Free Documentation License" [1, 2, 3, 6, 8].



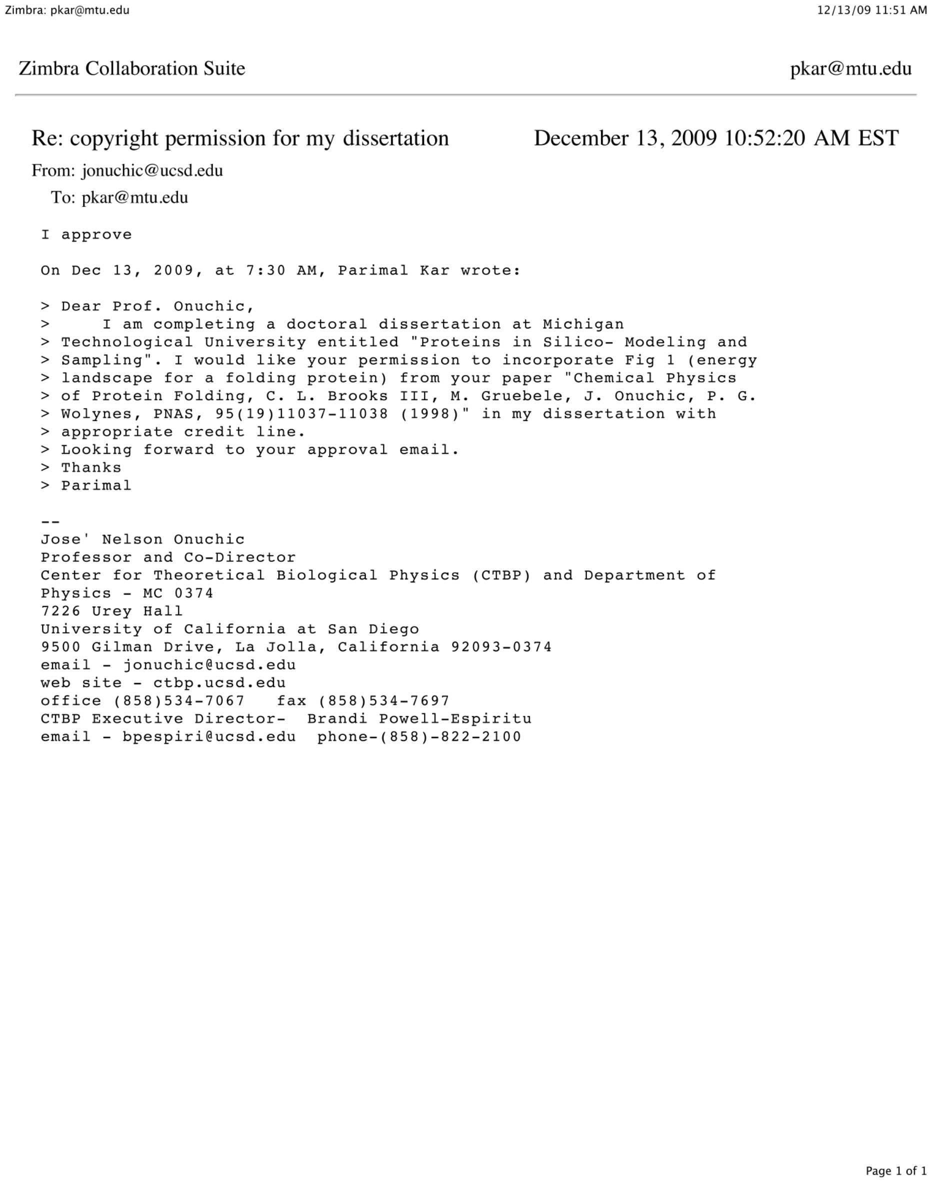
Zimbra: pkar@mtu.edu https://huskymail.mtu.edu/zi

Zimbra Collaboration Suite pkar@mtu.edu

RE: Copyright permission for my dissertation Thursday, November 19, 2009 4:46:47 PM
From: steinbac@exchange.nih.gov
To: pkar@mtu.edu

```
Dear Parimal,

Yes, you may use the figure.  Good luck with your dissertation.

Sincerely,
Peter

-----Original Message-----
From: Parimal Kar [mailto:pkar@mtu.edu]
Sent: Thursday, November 19, 2009 4:43 PM
To: steinbac@helix.nih.gov
Subject: Copyright permission for my dissertation

Dear Dr. Steinbach,
                  I am completing a doctoral dissertation at Michigan
Technological University entitled "Proteins in Silico- Modeling and
Sampling". I would like your permission to incorporate the following
figure with appropriate credit in my dissertation.
http://cmm.cit.nih.gov/steinbach/intro_sim_gifs/pe.gif
Looking forward to your approval email.
```

Figure D.3: Permission letter for Fig 2.2.



Zimbra: pkar@mtu.edu 12/13/09 11:51 AM

Zimbra Collaboration Suite pkar@mtu.edu

Re: copyright permission for my dissertation December 13, 2009 10:52:20 AM EST

From: jonuchic@ucsd.edu
To: pkar@mtu.edu

```
I approve

On Dec 13, 2009, at 7:30 AM, Parimal Kar wrote:

> Dear Prof. Onuchic,
>     I am completing a doctoral dissertation at Michigan
> Technological University entitled "Proteins in Silico- Modeling and
> Sampling". I would like your permission to incorporate Fig 1 (energy
> landscape for a folding protein) from your paper "Chemical Physics
> of Protein Folding, C. L. Brooks III, M. Gruebele, J. Onuchic, P. G.
> Wolynes, PNAS, 95(19)11037-11038 (1998)" in my dissertation with
> appropriate credit line.
> Looking forward to your approval email.
> Thanks
> Parimal

--
Jose' Nelson Onuchic
Professor and Co-Director
Center for Theoretical Biological Physics (CTBP) and Department of
Physics - MC 0374
7226 Urey Hall
University of California at San Diego
9500 Gilman Drive, La Jolla, California 92093-0374
email - jonuchic@ucsd.edu
web site - ctbp.ucsd.edu
office (858)534-7067   fax (858)534-7697
CTBP Executive Director-  Brandi Powell-Espiritu
email - bpespiri@ucsd.edu  phone-(858)-822-2100
```

Page 1 of 1

Figure D.4: Permission letter for Fig 1.5

Appendix E

Dispersion

In this chapter, some additional plots and tables from our investigations on dispersion term are reproduced from our paper– P. Kar, M. Seel. U. H. E. Hansmann and S. Höfinger, Dispersion Terms and Analysis of Size- and Charge- Dependence in an Enhanced Poisson-Boltzmann Approach, J. Phys. Chem. B, 111 (2007) 8910. Copyright ©2007, American Chemical Society. All the figures are tables are reproduced without any changes of our original paper.

Table E.1: Comparison of average PB/BEM solvation free energies ΔG^{solv} of homo-dipeptides in water to corresponding data obtained from PCM reference calculations [12].

Dipeptide	Mean $\Delta G^{solv,PB/BEM}$	Mean $\Delta G^{solv,PCM}$	Mean $\Delta\Delta G^{solv}$	Number of
Type		PCM Reference	Deviation	References
	[kcal/mol]	[kcal/mol]	[kcal/mol]	
AA	-85.01 (9.33)	-73.84 (9.72)	11.17	9
CC	-101.37 (12.31)	-89.29 (13.08)	12.08	9
DD	-292.81 (18.18)	-275.36 (18.64)	17.45	9
EE	-260.66 (15.05)	-247.40 (13.87)	13.26	9
GG	-92.83 (10.13)	-82.85 (11.70)	9.98	9
II	-75.41 (7.57)	-61.80 (9.45)	13.62	9
KK	-241.39 (18.12)	-223.55 (20.59)	17.84	9
LL	-77.14 (7.48)	-52.58 (7.78)	24.56	9
MM	-81.66 (7.94)	-67.93 (8.89)	13.73	9
NN	-96.78 (7.87)	-92.56 (11.58)	5.24	9
QQ	-111.53 (11.25)	-103.86 (13.11)	7.66	9
RR	-227.26 (18.48)	-215.16 (21.51)	12.10	6
SS	-108.11 (14.60)	-97.42 (14.28)	10.69	9
TT	-100.48 (11.81)	-90.90 (13.60)	9.58	9
VV	-78.04 (7.38)	-63.94 (8.57)	14.10	9
YY	-85.18 (4.75)	-76.66 (7.83)	8.51	5
FF	-79.28 (4.59)	-67.95 (6.61)	11.33	6
HH	-229.71 (22.32)	-225.59 (21.34)	4.14	9
PP	-57.30 (5.65)	-74.07 (8.86)	16.77	9
WW	-96.14 (9.14)	-81.19 (17.29)	14.95	3

Table E.2: Comparison of the average PB/BEM ΔG^{disp} contribution, to ΔG^{solv}, of homo-dipeptides in water to corresponding data obtained from PCM calculation [12].

Dipeptide Type	Mean $\Delta G^{disp,PB/BEM}$ [kcal/mol]	Mean $\Delta G^{disp+rep,PCM}$ PCM Reference [kcal/mol]	Mean $\Delta\Delta G^{disp}$ Deviation [kcal/mol]	Number of References
AA	-7.82 (0.21)	-15.05 (0.24)	7.23	9
CC	-10.83 (0.26)	-20.59 (0.28)	9.76	9
DD	-9.80 (0.22)	-19.92 (0.26)	10.12	9
EE	-10.91 (0.25)	-22.84 (0.60)	11.94	9
GG	-7.11 (0.08)	-13.51 (0.28)	6.41	9
II	-10.72 (0.24)	-21.57 (0.51)	10.85	9
KK	-14.05 (0.34)	-27.85 (0.68)	13.80	9
LL	-10.83 (0.24)	-22.61 (0.50)	11.78	9
MM	-12.57 (0.30)	-24.85 (0.66)	12.28	9
NN	-10.59 (0.17)	-21.18 (0.20)	10.60	9
QQ	-11.88 (0.35)	-24.18 (0.79)	12.30	9
RR	-16.04 (0.43)	-31.89 (0.59)	15.85	6
SS	-9.17 (0.19)	-17.88 (0.27)	8.71	9
TT	-9.80 (0.29)	-19.50 (0.31)	9.69	9
VV	-9.76 (0.19)	-19.02 (0.27)	9.26	9
YY	-13.53 (0.17)	-28.49 (0.50)	14.96	5
FF	-12.48 (0.13)	-26.81 (0.40)	14.34	6
HH	-12.74 (0.17)	-25.72 (0.58)	12.98	9
PP	-9.10 (0.19)	-19.71 (0.22)	10.62	9
WW	-14.75 (0.34)	-31.29 (0.93)	16.54	3

Table E.3: Methanol: Comparison of average molecular surfaces based on scaled AMBER vdW radii used in PB/BEM ($r_{probe}^{CH_3OH} = 1.9$Å) with data from PCM calculations ($r_{probe}^{CH_3OH} = 1.855$Å) [12].

Dipeptide Type	Mean Surface AMBER Scaled (1.06) [Å^2]	Mean Surface PCM Reference [Å^2]	Deviation [Å^2]
AA	203.53 (3.68)	211.08 (5.04)	7.55
CC	225.90 (5.14)	224.37 (6.03)	1.69
DD	241.22 (8.15)	239.96 (7.07)	1.54
EE	286.49 (7.84)	282.40 (6.45)	4.08
GG	163.77 (3.89)	167.53 (3.37)	3.76
II	293.76 (11.97)	301.17 (12.63)	7.41
KK	336.59 (8.08)	339.73 (8.01)	3.14
LL	290.53 (12.57)	292.25 (10.66)	1.84
MM	314.12 (8.51)	328.82 (9.08)	14.70
NN	244.26 (7.63)	245.88 (7.25)	1.62
QQ	291.05 (7.54)	291.91 (6.55)	1.17
RR	375.38 (7.21)	379.73 (7.37)	4.35
SS	208.32 (3.94)	211.61 (4.96)	3.29
TT	236.26 (9.96)	238.25 (10.50)	2.07
VV	264.63 (7.55)	275.52 (9.31)	10.89
YY	355.92 (15.10)	344.65 (14.65)	11.27
FF	341.45 (14.99)	325.22 (14.00)	16.23
HH	296.70 (11.78)	295.41 (12.01)	1.65
PP	234.58 (10.08)	232.42 (9.81)	2.16
WW	369.30 (26.98)	359.16 (24.50)	10.14

Table E.4: Methanol: Comparison of average molecular volumes based on scaled AMBER vdW radii used in PB/BEM ($r_{probe}^{CH_3OH} = 1.9$Å) with data from PCM calculations ($r_{probe}^{CH_3OH} = 1.855$Å) [12].

Dipeptide Type	Mean Volume AMBER Scaled (1.06) [Å^3]	Mean Volume PCM Reference [Å^3]	Deviation [Å^3]
AA	218.52 (2.65)	230.50 (3.13)	11.98
CC	252.11 (5.05)	245.04 (4.05)	7.07
DD	273.54 (5.66)	263.20 (5.74)	10.34
EE	331.37 (6.13)	314.95 (5.58)	16.41
GG	159.71 (2.78)	165.34 (1.94)	5.64
II	362.81 (9.49)	369.93 (7.06)	7.12
KK	393.01 (6.45)	392.01 (6.08)	1.67
LL	352.63 (9.14)	347.51 (7.76)	5.12
MM	365.52 (5.39)	381.78 (5.78)	16.26
NN	278.91 (8.78)	273.94 (5.77)	4.97
QQ	339.98 (6.49)	329.91 (7.30)	10.06
RR	433.26 (5.96)	428.44 (5.41)	4.82
SS	226.09 (3.61)	227.08 (2.82)	1.02
TT	274.80 (9.52)	275.55 (7.01)	1.51
VV	316.00 (5.48)	332.94 (8.52)	16.94
YY	439.88 (8.00)	412.28 (6.71)	27.61
FF	421.05 (8.64)	392.43 (6.81)	28.61
HH	356.34 (6.00)	343.87 (5.54)	12.47
PP	268.97 (9.02)	264.79 (10.71)	4.18
WW	478.49 (18.15)	453.04 (16.78)	25.45

Table E.5: Ethanol: Comparison of average molecular surfaces based on scaled AMBER vdW radii used in PB/BEM ($r_{probe}^{C_2H_5OH} = 2.2$Å) with data from PCM calculations ($r_{probe}^{C_2H_5OH} = 2.180$Å) [12].

Dipeptide Type	Mean Surface AMBER Scaled (1.06) [Å^2]	Mean Surface PCM Reference [Å^2]	Deviation [Å^2]
AA	204.97 (4.93)	210.87 (4.77)	5.90
CC	226.79 (5.20)	224.03 (5.83)	2.76
DD	240.79 (8.29)	239.53 (6.89)	1.47
EE	285.56 (9.04)	281.88 (6.44)	3.68
GG	164.30 (4.51)	167.49 (3.46)	3.20
II	293.16 (15.47)	300.40 (12.81)	7.25
KK	337.35 (9.04)	339.50 (7.99)	2.36
LL	289.47 (10.05)	291.74 (10.96)	2.40
MM	314.80 (8.80)	328.59 (9.29)	13.80
NN	244.51 (7.59)	245.49 (7.11)	1.02
QQ	291.32 (8.92)	291.08 (7.39)	1.39
RR	375.21 (7.42)	379.43 (7.12)	4.22
SS	209.14 (4.35)	211.31 (5.05)	2.16
TT	236.06 (9.37)	237.89 (10.49)	1.83
VV	262.57 (9.85)	275.17 (9.22)	12.60
YY	354.25 (14.33)	344.01 (13.96)	10.23
FF	339.32 (13.94)	324.40 (14.52)	14.92
HH	296.78 (11.99)	294.80 (11.73)	2.03
PP	233.88 (11.20)	232.11 (9.56)	1.77
WW	371.10 (26.73)	358.23 (24.44)	12.87

Table E.6: Ethanol: Comparison of average molecular volumes based on scaled AMBER vdW radii used in PB/BEM ($r_{probe}^{C_2H_5OH} = 2.2$Å) with data from PCM calculations ($r_{probe}^{C_2H_5OH} = 2.180$Å) [12].

Dipeptide Type	Mean Volume AMBER Scaled (1.06) [Å^3]	Mean Volume PCM Reference [Å^3]	Deviation [Å^3]
AA	222.32 (3.54)	231.95 (3.67)	9.62
CC	255.42 (4.94)	246.50 (4.51)	8.92
DD	275.65 (6.27)	264.60 (5.45)	11.05
EE	334.55 (6.38)	317.10 (7.04)	17.45
GG	161.17 (4.03)	166.07 (2.13)	4.89
II	365.48 (10.21)	372.08 (7.14)	6.60
KK	400.13 (6.74)	394.36 (6.63)	5.78
LL	355.61 (8.64)	349.47 (8.09)	6.14
MM	371.57 (8.47)	384.59 (6.88)	13.02
NN	281.96 (7.47)	275.44 (6.19)	6.52
QQ	344.23 (7.09)	332.30 (7.72)	11.93
RR	438.11 (9.96)	431.38 (7.89)	6.74
SS	229.28 (3.55)	228.46 (2.92)	1.04
TT	277.64 (7.60)	276.95 (7.37)	1.91
VV	316.82 (9.30)	334.66 (8.97)	17.84
YY	442.09 (8.74)	415.33 (6.62)	26.76
FF	422.93 (9.19)	395.15 (7.77)	27.78
HH	360.78 (6.09)	346.26 (5.67)	14.52
PP	270.54 (9.96)	266.02 (10.44)	4.52
WW	486.78 (18.77)	456.45 (18.69)	30.34

Table E.7: n-Octanol: Comparison of average molecular surfaces based on scaled AMBER van der Waals radii used in PB/BEM ($r_{probe}^{C_8H_{17}OH} = 2.945$Å) with data from PCM reference calculations ($r_{probe}^{C_8H_{17}OH} = 2.945$Å) [12].

Dipeptide Type	Mean Surface AMBER Scaled (1.05) [Å^2]	Mean Surface PCM Reference [Å^2]	Deviation [Å^2]
AA	202.11 (5.74)	210.82 (4.73)	8.70
CC	226.01 (5.49)	223.80 (6.29)	2.21
DD	239.04 (6.31)	239.28 (7.36)	1.28
EE	282.23 (7.88)	281.70 (6.73)	1.17
GG	162.96 (4.44)	167.52 (3.64)	4.56
II	290.48 (10.88)	299.73 (13.15)	9.25
KK	334.37 (9.81)	340.07 (12.74)	5.70
LL	287.06 (11.27)	291.21 (10.81)	4.23
MM	314.52 (8.80)	328.24 (9.34)	13.73
NN	242.92 (8.11)	245.07 (7.66)	2.15
QQ	289.02 (9.69)	290.89 (7.47)	2.50
RR	371.52 (8.05)	379.87 (10.01)	8.35
SS	207.15 (5.90)	211.18 (4.98)	4.03
TT	234.20 (10.29)	237.46 (10.34)	3.25
VV	260.13 (9.22)	274.78 (9.04)	14.65
YY	347.57 (14.32)	343.46 (14.35)	4.14
FF	335.63 (14.76)	323.66 (14.54)	11.97
HH	295.00 (11.34)	294.17 (11.88)	1.77
PP	232.87 (13.19)	231.68 (9.71)	1.31
WW	368.60 (23.31)	356.92 (24.09)	11.68

Table E.8: n-Octanol: Comparison of average molecular volumes based on scaled AMBER van der Waals radii used in PB/BEM ($r_{probe}^{C_8H_{17}OH} = 2.945$Å) with data from PCM reference calculations ($r_{probe}^{C_8H_{17}OH} = 2.945$Å) [12].

Dipeptide Type	Mean Volume AMBER Scaled (1.05) [Å^3]	Mean Volume PCM Reference [Å^3]	Deviation [Å^3]
AA	221.75 (5.74)	234.29 (4.06)	12.54
CC	257.60 (5.08)	249.12 (4.82)	8.47
DD	276.64 (6.03)	267.32 (5.21)	9.31
EE	335.59 (7.64)	322.45 (8.78)	13.14
GG	160.76 (2.82)	167.45 (2.29)	6.69
II	367.89 (7.76)	376.29 (7.76)	8.40
KK	403.23 (7.18)	401.12 (6.86)	2.74
LL	356.68 (9.22)	353.55 (8.66)	3.13
MM	377.61 (6.62)	391.07 (8.27)	13.46
NN	283.89 (6.03)	278.66 (6.71)	5.23
QQ	346.36 (7.96)	338.62 (8.84)	7.74
RR	442.46 (12.86)	439.20 (12.55)	3.25
SS	229.66 (4.77)	230.90 (3.92)	1.25
TT	278.95 (7.47)	279.20 (7.82)	1.77
VV	318.05 (9.29)	338.23 (9.80)	20.18
YY	440.24 (11.50)	421.48 (9.26)	18.76
FF	424.07 (13.18)	400.56 (11.15)	23.51
HH	362.43 (7.62)	350.96 (7.18)	11.46
PP	272.69 (10.68)	268.18 (10.22)	4.51
WW	490.61 (19.88)	463.21 (21.20)	27.40

Table E.9: Ethanol: Effect on total solvation free energies as PB/BEM-computed with AMBER style of dispersion (λ=0.94) versus Caillet-Claverie style of dispersion (λ=0.82) and comparison to the experimental value [12].

Species	$\Delta G^{solv}_{Caillet-Claverie}$ $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{AMBER} $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{Exp} $\left[\frac{kcal}{mol}\right]$
n-octane	-0.70	-1.56	-4.23
toluene	-3.30	-4.53	-4.57
dioxane	-6.03	-6.74	-4.68
butanone	-4.83	-3.88	-4.32
chlorobenzene	-3.52	-4.16	-3.30

Table E.10: n-Octanol: Effect on total solvation free energies as PB/BEM-computed with AMBER style of dispersion (λ=2.60) versus Caillet-Claverie style of dispersion (λ=0.74) and comparison to the experimental value [12].

Species	$\Delta G^{solv}_{Caillet-Claverie}$ $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{AMBER} $\left[\frac{kcal}{mol}\right]$	ΔG^{solv}_{Exp} $\left[\frac{kcal}{mol}\right]$
acetone	-5.28	-4.35	-3.15
anisole	-4.80	-6.74	-5.47
benzaldehyde	-6.16	-6.25	-6.13
benzene	-3.87	-5.54	-3.72
bromobenzene	-3.75	-4.93	-7.47
butanal	-5.02	-4.19	-4.62
butanoic acid$^{(a)}$	-8.74	-8.14	-7.58
cyclohexane	-0.64	-2.02	-3.46
acetic acid$^{(a)}$	-8.96	-7.84	-6.35
ethylbenzene	-2.94	-4.84	-5.08
hexanoic acid$^{(a)}$	-8.89	-8.84	-8.82
propanal	-4.71	-3.63	-4.13
propionic acid$^{(a)}$	-8.75	-8.03	-6.86
propene	-1.61	-2.44	-1.14
propyne	-2.81	-3.86	-1.59
bromoethane	-2.69	-2.58	-2.90

$^{(a)}$ protonated form

Table E.11: Comparison of computed versus experimental total solvation free energies, ΔG^{solv}, of amino acid side-chain analogues in water.[12].

Species	$\Delta G^{solv,PB/BEM+SASA\gamma^{(a)}}$ $\left[\frac{kcal}{mol}\right]$	$\Delta G^{solv,Exp}$ $\left[\frac{kcal}{mol}\right]$	Deviation $\left[\frac{kcal}{mol}\right]$
acetamide	-10.55	-9.68	0.87(+)
butane	0.84	2.15	1.31(-)
ethanol	-4.51	-4.88	0.37(-)
isobutane	0.57	2.28	1.71(-)
methane	0.62	1.94	1.32(-)
methanethiol	-2.53	-1.24	1.29(+)
methanol	-5.91	-5.06	0.85(+)
methyl-ethyl-sulfide	-0.55	-1.48	0.93(+)
methylindole	-5.14	-5.88	0.74(+)
p-cresol	—	-6.11	—
propane	1.00	1.99	0.99(-)
propionamide	—	-9.38	—
toluene	—	-0.76	—

Table E.12: Performance evaluation of the components involved in the calculation of the dispersion term, ΔG^{disp}, according to eq. 3 (AMBER/TIP3P) [12].

PDB	Number of Residues	Number of Atoms	CPU Time Mol. Surf. [sec]	CPU Time [sec]
1P9GA	41	517	5 (10 %)	1 (2 %)
2B97	70	981	30 (21 %)	1 (1 %)
1LNI	96	1443	50 (14 %)	2 (1 %)
1NKI	134	2082	67 (9 %)	5 (1 %)
1EB6	177	2570	108 (18 %)	5 (1 %)
1G66	207	2777	127 (20 %)	5 (1 %)
1P1X	250	3813	185 (15 %)	10 (1 %)
1RTQ	291	4287	214 (17 %)	11 (1 %)
1YQS	345	5147	247 (14 %)	16 (1 %)
1GPI	430	6164	200 (8 %)	21 (1 %)

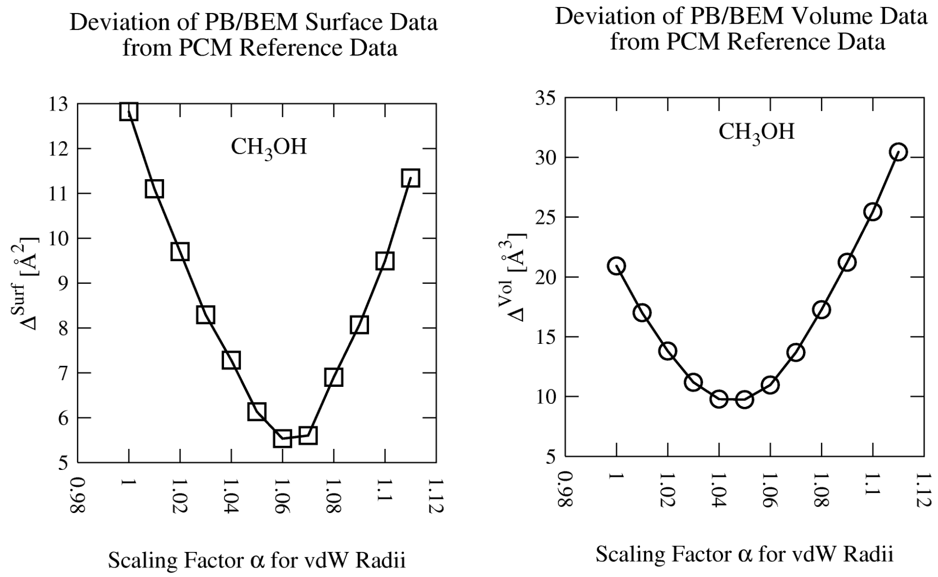


Figure E.1: Methanol: Comparison of employed molecular surfaces (L) and Molecular volumes (R) in the PB/BEM series based on scaling the AMBER default van der Waals radii by a factor α to the reference data obtained from PCM calculations [12].

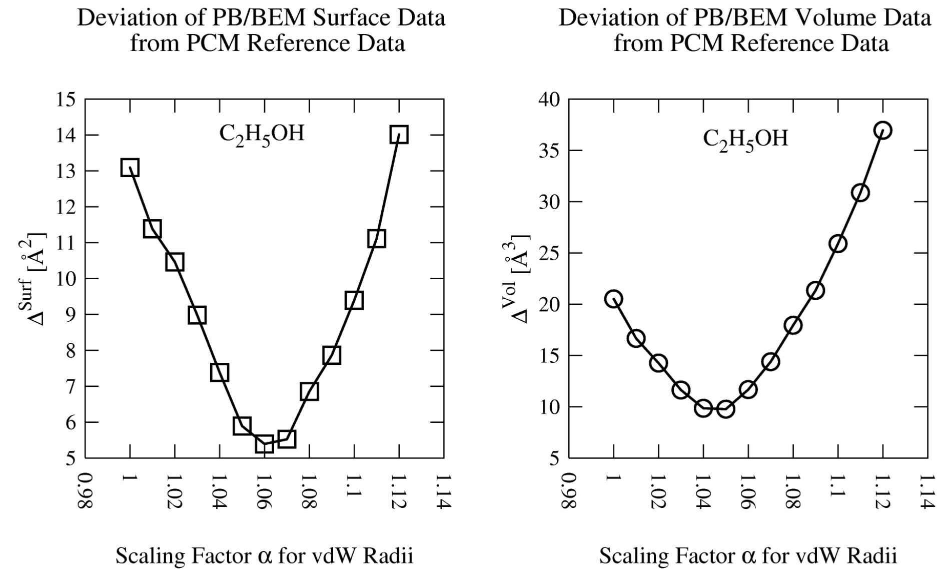


Figure E.2: Ethanol: Comparison of employed molecular surfaces (L) and molecular volumes (R) in the PB/BEM series based on scaling the AMBER default van der Waals radii by a factor α to the reference data obtained from PCM calculations [12].

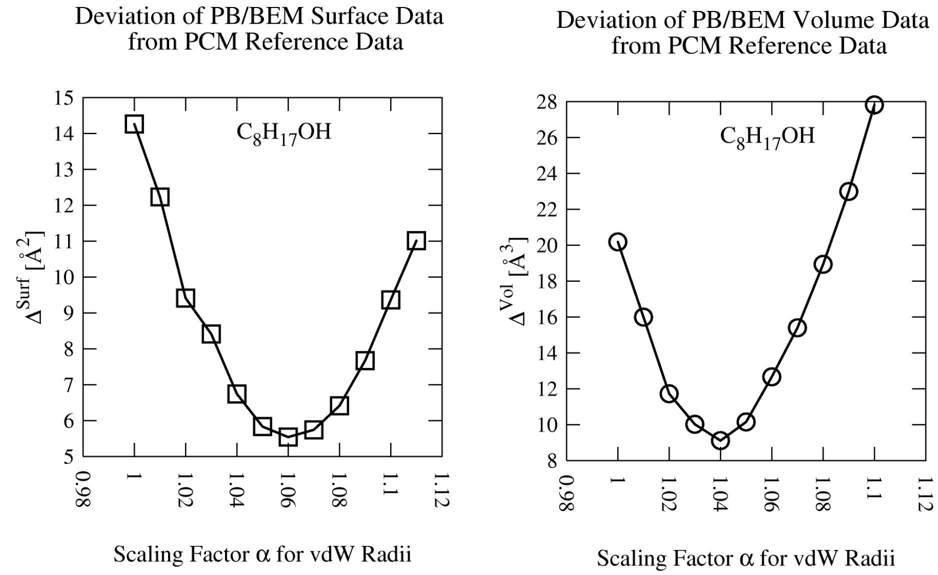


Figure E.3: n-Octanol: Comparison of employed molecular surfaces (L) and molecular volumes (R) in the PB/BEM series based on scaling the AMBER default van der Waals radii by a factor α to the reference data obtained from PCM calculations [12].

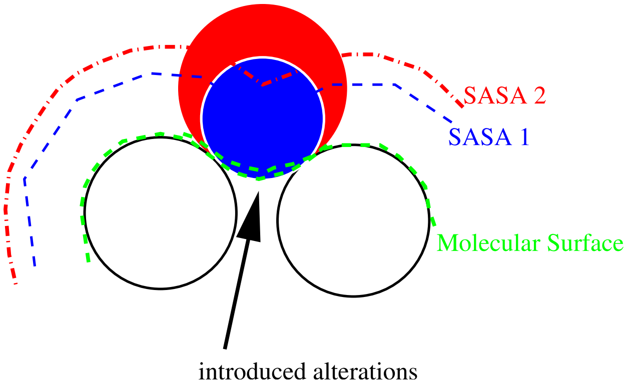


Figure E.4: Graphical representation of introduced changes when switching from a small probe sphere (blue) to a larger probe sphere (red) [12].

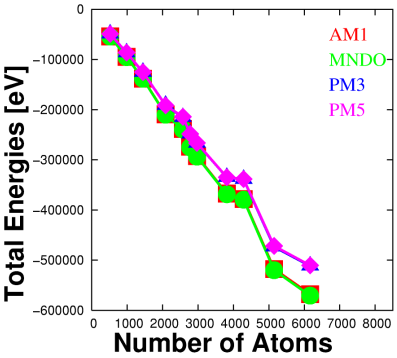


Figure E.5: Graphical representation of the total energies determined at different levels of semiempirical theory using the program LocalSCF [12].

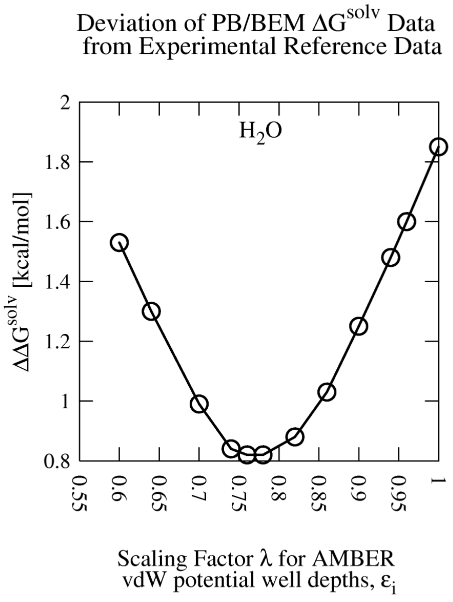


Figure E.6: Deviation of the PB/BEM solvation free energies ΔG^{solv} from experimental values as a function of λ, a scaling factor uniformly applied to all AMBER vdW potential well depths ε_i [12].

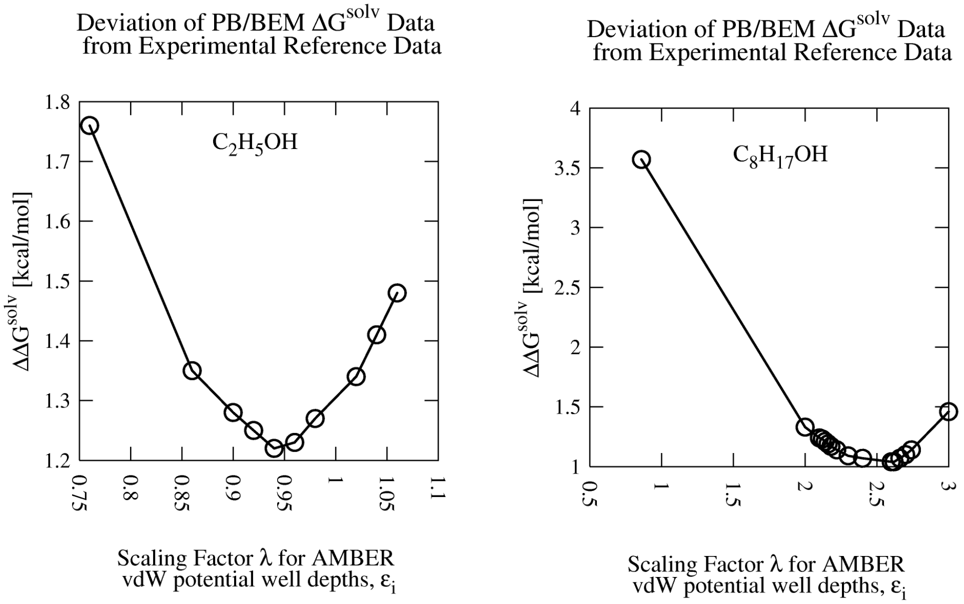


Figure E.7: Ethanol (L) & n-Octanol (R) : Deviation of the PB/BEM solvation free energies ΔG^{solv} from experimental values as a function of λ, a scaling factor uniformly applied to all AMBER vdW potential well depths ε_i [12].

References

[1] Matanya. *http://commons.wikimedia.org/wiki/File:Amino_acids_2.png;* 2005.

[2] Mrabet, Y. *http://en.wikipedia.org/wiki/File:Peptidformationball.svg;* 2007.

[3] Mrabet, Y. *http://en.wikipedia.org/wiki/File:AminoAcidball.svg;* 2007.

[4] *http://www.genome.gov//Pages/Hyperion/DIR/VIP/Glossary/Illustration/protein.cfm;* NIH.

[5] Leopold, P. E.; Onuchic, J. N. *Proc. Natil. Acad. Scie. USA* **1992**, *89*, 8721.

[6] Somoza. *http://en.wikipedia.org/wiki/File:Morse-potential.png;* 2006.

[7] Steinbach. *http://cmm.cit.nih.gov/steinbach/intro_sim_gifs/pe.gif;* 2005.

[8] *http://en.wikipedia.org/wiki/Water_model;* 2007.

[9] Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118.

[10] Vorobjev, Y. N.; Hermans, J. *Biophys. J.* **1997**, *73*, 722.

[11] Kar, P.; Wei, Y.; Hansmann, U. H. E.; Höfinger, S. *J. Comput. Chem.* **2007**, *28*, 2538–2544.

[12] Kar, P.; Seel, M.; Hansmann, U. H. E.; Höfinger, S. *J. Phys. Chem. B* **2007**, *111*, 8910–8918.

[13] Chang, J.; Lenhoff, A. M.; Sandler, S. I. *J. Phys. Chem. B* **2007**, *111*, 2098–2106.

[14] Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1999**, *103*, 9–63.

[15] Kar, P.; Seel, M.; Hansmann, U. H. E.; Höfinger, S. *NIC Publication Series* **2007**, *36*, 173–176.

[16] Kar, P.; Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2009**, *80*, 056703.

[17] Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E.* **2007**, *76*, 057102.

[18] Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Jr., K. M. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[19] Anikin, N. A.; Anisimov, V. M.; Bugaenko, V. L.; Bobrikov, V. V. *J. Chem. Phys.* **2005**, *121*, 1266–1270.

[20] Kar, P.; Seel, M.; Hansmann, U. H. E.; Höfinger, S. *NIC Publication Series* **2007**, *36*, 155–158.

[21] Lesk, A. M. *Introduction to Protein Architecture;* Oxford University Press: New York, USA, 2001.

[22] Garrett, R. H.; Grisham, C. M. *Biochemistry;* Cengage Learning: USA, 2006.

[23] Nelson, D. L.; Cox, M. M. *Lehninger's Principles of Biochemistry;* W. H. Freeman and Co., 2005.

[24] Mckee, T.; Mckee, J. R. *Biochemistry: An Introduction;* McGraw-Hill: Beijing, 1999.

[25] Wei, Y. *On Side Chain and Backbone Ordering in Polypeptides;* 2007.

[26] Creighton, T. E. *Proteins;* W. H. Freeman and Company: New York, USA, 1993.

[27] Radzicka, A.; Wolfenden, R. *Science* **1995**, *267*, 90.

[28] Rojnuckarindagger, A.; Kimdagger, S.; Subramaniam, S. *Biophys. J.* **1998**, *95*, 4288.

[29] Sadiqi, M.; Fushman, D.; Muoz, V. *Nature* **2006**, *317*, 442.

[30] Pain, R. H. *Mechanism of Proteins Folding;* Oxford University Press: New York, USA, 2000.

[31] Anfinsen, C. B. *Science* **1973**, *181*, 223.

[32] Dobson, C. M.; Sali, A.; karplus, M. *Angew. Chem. Int. Ed. Eng.* **1998**, *37*, 868.

[33] Karplus, M. *Fold. Des.* **1997**, *2*.

[34] Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding;* Freeman, 1999.

[35] Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct. Funct. Genet.* **1995**, *21*, 167.

[36] Baldwin, R. L. *Curr. Opin. Struct. Biol.* **1993**, *3*, 84.

[37] Matouschek, A.; Jr., J. R. K.; Serrano, L.; Fersht, A. R. *Nature* **1989**, *340*, 122.

[38] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.

[39] Drenth, J. *Principles of Protein X-Ray Crystallography;* Springer-Verlag Inc., 1999.

[40] Hansmann, U. H. E. *Comp. Sci. Eng.* **2003**, *5*, 64.

[41] Karplus, M. *Biopolymers* **2003**, *68*, 350.

[42] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.

[43] Leach, A. R. *Molecular Modelling: Principles and Applications;* Prentice Hall, 2001.

[44] Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

[45] Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

[46] Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

[47] Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

[48] Rinaldi, D.; Ruiz-López, M. F.; Rivail, J. L. *J. Chem. Phys.* **1983**, *78*, 834.

[49] Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671–679.

[50] Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144–1149.

[51] Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244–1253.

[52] Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10037–10041.

[53] Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219–10225.

[54] Zauhar, R. J.; Morgan, R. S. *J. Mol. Biol.* **1985**, *186*, 815–820.

[55] Juffer, A. H.; Botta, E. F. F.; van Keulen, B. A. M.; van der Ploeg, A.; Berendsen, H. J. C. *J. Comput. Phys.* **1991**, *97*, 144–171.

[56] Mahajan, R.; Kranzlmüller, D.; Volkert, J.; Hansmann, U. H. E.; Höfinger, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5515–5521.

[57] Caillet, J.; Claverie, P. *Acta Cryst. A* **1975**, *31*, 448–461.

[58] Claverie, P. *Intermolecular Interactions: From Diatomic to Biopolymers;* Wiley: New York, 1978.

[59] Höfinger, S.; Zerbetto, F. *Chem. Soc. Rev.* **2005**, *34*, 1012–1020.

[60] Reiss, H.; Frisch, H. L.; Lebowitz, J. L. *J. Chem. Phys.* **1959**, *31*, 369–380.

[61] Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717–726.

[62] Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.

[63] Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118–1124.

[64] Morse, P. M. *Phys. Rev.* **1929**, *34*, 57–64.

[65] Verma, A. *Development and Application of a Free Energy Forccce Field for All Atom Protein Folding (PhD thesis);* Forschungszentrum Karlsruhe: Karlsruhe, DE, 2007.

[66] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

[67] Tironi, I.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

[68] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[69] Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

[70] Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933.

[71] Sippl, M. J.; Nemethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1984**, *88*, 6231.

[72] Eisenmenger, F.; Hansmann, U. H. E.; Hayryan, S.; Hu, C. K. *Comp. Phys. Comm.* **2001**, page 192.

[73] Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127.

[74] Allinger, N. L.; Yuh, Y. H.; Lii, J. H. **1989**.

[75] Hagler, A. T.; Ewig, C. S. *Comp. Phys. Comm.* **1994**, *84*, 131–155.

[76] Patel, S.; III, C. L. B. *Mol. Sim.* **2006**, *32*, 231–249.

[77] Rick, S. W. *J. Chem. Phys.* **2004**, *120*, 6085–6093.

[78] Garrod, C. *Statistical Mechanics and Thermodynamics;* Oxford University Press, 1995.

[79] Schwabl, F. *Statistical Mechanics;* Springer, 2002.

[80] Zimmermann, O.; Hansmann, U. H. E. *Biochimica et Biophysica Acta- Proteins and Proteomics* **2008**, *252*, 1784.

[81] Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903.

[82] Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E.* **2007**, *75*, 026109.

[83] Nadler, W.; Meinke, J. A.; Hansmann, U. H. E. *Phys. Rev. E.* **2008**, *78*, 061905.

[84] Gront, D.; Kolinski, A. *J. Phys. Condens. matter* **2007**, *19*, 036225.

[85] Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96–123.

[86] Born, M. *Z. Phys.* **1920**, *1*, 45–48.

[87] Kirkwood, J. G.; Westheimer, F. H. *J. Chem. Phys.* **1938**, *6*, 506.

[88] Westheimer, F. H.; Kirkwood, J. G. *J. Chem. Phys.* **1938**, *6*, 513.

[89] Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333.

[90] Tanford, C. *J. Am. Chem. Soc.* **1957**, *79*, 5340–5348.

[91] Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

[92] Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

[93] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces;* Reidel Publishing Company: Dordrecht, the Netherlands, 1981.

[94] Jr., A. D. M.; Bashford, D.; Bellott, R. L.; Jr., R. L. D.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; III, W. E. R.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem.* **1998**, *102*, 3586–3616.

[95] Bernal, J. D.; Fowler, R. H. *J. Chem. Phys.* **1933**, *1*, 515.

[96] Jorgensen, W. L. *J. Chem. Phys.* **1982**, *77*, 4156–4163.

[97] Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665–9678.

[98] Abascal, J. L. F.; Sanz, E.; Fernndez, R. G.; Vega, C. *J. Chem. Phys.* **2005**, *122*, 234511.

[99] Abascal, J. L. F.; Vega, C. *J. Chem. Phys.* **2005**, *123*, 234505.

[100] Stillinger, F. H.; Rahman, A. *J. Chem. Phys.* **1974**, *60*, 1545–1557.

[101] Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

[102] Nada, H.; van der Eerden, J. P. J. M. *J. Chem. Phys.* **2003**, *118*, 7401.

[103] McQuarrie, D. A. *Statistical Mechanics;* Harper and Row: New York, USA, 1976.

[104] Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.

[105] Simonson, T.; Brünger, A. T. *Biochemistry* **1992**, *31*, 8661–8674.

[106] Boresch, S.; Archontis, G. *Proteins* **1994**, *20*, 25.

[107] Im, W.; Beglov, D.; Roux, B. *Comp. Phys. Commun.* **1998**, *111*, 1–3.

[108] Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. *J. Phys. Chem.* **1993**, *97*, 3591–3600.

[109] Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, pages 128–137.

[110] Nicholls, A.; Sharp, K. A.; Honig, B. *Proteins* **1991**, pages 281–296.

[111] Bahsford, D. *Lecture Notes in Computer Science* **1997**, pages 233–240.

[112] Grant, J. A.; Pickup, B. T.; Nicholls, A. *J. Comput. Chem.* **2001**, pages 608–640.

[113] Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comp. Phys. Commun.* **1995**, pages 57–95.

[114] Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, pages 1570–1590.

[115] Cortis, C. M.; friesner, R. A. *J. Comput. Chem.* **1997**, pages 1591–1608.

[116] Jr., A. D. M.; Brooks, B.; III, C. L. B.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. *The Encyclopedia of Computational Chemistry;* John Wiley and Sons: Chichester, 1998.

[117] Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

[118] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

[119] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1995**, *246*, 122–129.

[120] Lee, M. S.; Jr., F. R. S.; III, C. L. B. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

[121] Lee, M. S.; Feig, M.; Salsbury, F. R.; III, C. L. B. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

[122] Stillinger, F. *J. Solution Chem.* **1973**, *2*, 141–158.

[123] Tolman, R. C. *J. Chem. Phys.* **1948**, *16*, 758–774.

[124] Sharp, K.; Nicholls, A.; Fine, R.; Honig, B. *Science* **1991**, *252*, 106–109.

[125] Sharp, K.; Nicholls, A.; Friedman, R.; Honig, B. *Biochemistry 30*, 9696–9697.

[126] Eisenberg, D.; McClachlan, A. *Nature* **1986**, *319*, 199–203.

[127] Wesson, L.; Eisenberg, D. *protein Sci.* **1993**, *1*, 227–235.

[128] Fraternali, F.; van Gunsteren, W. F. *J. Mol. Biol.* **1996**, *256*, 939–948.

[129] Scheraga, H. A. *Acc. Chem. Res.* **1979**, *12*, 7–14.

[130] Kang, Y. K.; Gibson, K. D.; Nemethy, G.; Scheraga, H. *J. Phys. Chem.* **1988**, *92*, 4739–4742.

[131] Colonna-Cesari, F.; Sander, C. *Biophys. J.* **1990**, *57*, 1103–1107.

[132] Stouten, P.; Frömmel, C.; Nakamura, H.; Sander, C. *Mol. Sim.* **1993**, *10*, 97–120.

[133] Lazaridis, T.; Karplus, M. *Science* **1997**, *278*, 1928–1931.

[134] Wodak, S. J.; Janin, J. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 1736–1740.

[135] Curutchet, C.; Orozco, M.; Luque, F. J.; Mennucci, B.; Tomasi, J. *J. Comput. Chem.* **2006**, *27*, 1769–1780.

[136] Lii, J. H.; Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8576.

[137] Aqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021.

[138] Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *196*, 765.

[139] Halgren, T. *J. Am. Chem. Soc.* **1992**, *114*, 7827.

[140] Kollman, P. A.; Massaova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, Q.; Cieplak, P.; Srinivasan, J.; Case, D. A.; III, T. E. C. *Acc. Chem. Res.* **2000**, *33*, 889.

[141] Page, C. S.; Bates, P. A. *J. Comput. Chem.* **2006**, *27*, 1990.

[142] Becke, A. D. *J. Chem. Phys.* **1997**, *107*, 8554.

[143] Sadlej, A. J. *Theor. Chim. Acta* **1991**, *79*, 123.

[144] Akiyama, Y.; Onizuka, K.; Noguchi, T.; Ando, M. *Proc. 9th Genome Informatics Workshop (GIW'98);* Universal Academy Press, 1998.

[145] Schaftenaar, G.; Noordik, J. H. *J. Comput. Aid Mol. Des.* **2000**, *14*, 123.

[146] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

[147] Höfinger, S. *J. Comput. Chem.* **2005**, *26*, 1148–1154.

[148] Marshall, N. J.; Grail, B. M. *J. Pept. Sci.* **2000**, *6*, 186.

[149] Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

[150] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. A.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Menucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, A.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. **2004**.

[151] Ross, B. O.; Widmark, P. *European Summerschool in Quantum Chemistry, Book II;* 2000.

[152] Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

[153] Ooi, T.; Oobatake, M.; Nemethy, G.; Scherega, H. A. *Proc. Natl. Acad. Sci.* **1987**, *84*, 3086–3090.

[154] Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; III, C. L. B. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

[155] Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

[156] Mohan, V.; Davis, M. E.; McCammon, J. A.; Pettitt, B. M. *J. Phys. Chem.* **1992**, *96*, 6428–6431.

[157] Kar, P.; Wei, Y.; Hansmann, U. H. E.; Höfinger, S. *NIC Publication Series* **2006**, *34*, 161–164.

[158] Zacharias, M. *J. Phys. Chem. A* **2003**, *107*, 3000–3004.

[159] Pitera, J. W.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 3163–3164.

[160] Simonson, T.; Brunger, A. T. *J. Phys. Chem.* **1994**, *98*, 4683–4694.

[161] Su, Y.; Gallicchio, E. *Biophys. Chem.* **2004**, *109*, 251–260.

[162] Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 251–260.

[163] Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

[164] Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

[165] Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523–9530.

[166] Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.

[167] Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onuvriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

[168] Floris, F. M.; Tomasi, J.; Ahuir, J. L. P. *J. Comput. Chem.* **1991**, *12*, 784–791.

[169] Choudhury, N.; Pettitt, B. M. *J. Am. Chem. Soc.* **2005**, *127*, 3556–3567.

[170] Choudhury, N.; Pettitt, B. M. *Modelling Molecular Structure and Reactivity in Biological Systems;* RSC Publishing, 2006.

[171] Choudhury, N.; Pettitt, B. M. *J. Am. Chem. Soc.* **2007**, *129*, 4847–4852.

[172] Zhou, R.; Huang, X.; Margulis, C. J.; Berne, B. J. *Science* **2004**, *305*, 1605–1609.

[173] Klamt, A.; Schürmann, G. *J. Chem. Soc.* **1993**, *T2*, 799.

[174] Kar, P.; Seel, M.; Hansmann, U. H. E.; Höfinger, S. *J. Comput. Chem.* **2007**, *28*, 2538–2544.

[175] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. A.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Menucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, A.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *GAUSSIAN 98, Rev A.7;* Gaussian Inc.: Pittsburgh, PA, 1998.

[176] Stewart, J. J. P. *J. Mol. Model.* **2004**, *10*, 6–12.

[177] White, C. A.; Head-Gordon, M. *J. Chem. Phys.* **1994**, *101*, 6593.

[178] Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.

[179] MacCallum, J. L.; Tieleman, D. T. *J. Comput. Chem.* **2003**, *24*, 1930–1935.

[180] Höfinger, S.; Zerbetto, F. *Theor. Chem. Acc.* **2004**, *112*, 240.

[181] He, Y.; Xiao, Y.; Liwo, A.; Scheraga, H. A. *J. Comp. Chem.* **2009**, *30*, 2127.

[182] Pulay, P. *Chem. Phys. Lett.* **2005**, *26*, 1148–1154.

[183] Xiang, Y.; Huang, R. H.; Liu, X. Z.; Zhang, Y.; Wang, D. C. *J. Struct. Biol.* **2004**, *148*, 86–97.

[184] Kim, S. S.; Zhang, R. G.; Braunstein, S. E.; Joachimiak, A.; Cvekl, A.; Hegde, R. S. *Structure* **2002**, *10*, 787–795.

[185] Ye, Q.; Krug, R. M.; Tao, Y. J. *Nature* **2006**, *444*, 1078–1082.

[186] Yu, J. W.; Mendrola, J. M.; Audhya, A.; Singh, S.; Keleti, D.; deWald, D. B.; Murrary, D.; Emr, S. D.; Lemon, M. A. *Mol. Cell* **2004**, *13*, 677–688.

[187] Morikis, D.; Lambrish, J. D. *Trends Immunol.* **2004**, *25*, 700–707.

[188] Schleinkofer, K.; Weidemann, U.; Otte, L.; Wang, T.; Krause, G.; Oschkinat, H.; Wade, R. C. *J. Mol. Biol.* **2004**, *344*, 865–881.

[189] Mohanty, S.; Meinke, J. H.; Zimmermann, O.; Hansmann, U. H. E. *Proc. Natl. Acad. Sci.* **2008**, *105*, 8004.

[190] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids;* Oxford University Press: Oxford, UK, 1989.

[191] Frenkel, D.; Smit, B. *Understanding Molecular Simulation;* Academic Press: San Diego, USA, 2002.

[192] Kohtani, M.; Jones, T. C.; Schneider, J. E.; Jarrold, M. F. *J. Am. Chem. Soc.* **2004**, *126*, 7420.

[193] Neidigh, J. W.; Fesinmeyer, R. M.; Anderson, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425.

[194] Simmerling, C.; Strockbine, B.; Roitberg, A. *J. Am. Chem. Soc.* **2002**, *124*, 11258.

[195] Schug, A.; Herges, T.; Wenzel, W. *Phys. Rev. Lett.* **2003**, *30*, 158102.

[196] Geyer, C. J.; Thompson, A. *J. Am. Stat. Ass.* **1995**, *90*, 909.

[197] Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.

[198] Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comp. Chem.* **2000**, *21*, 1047.

[199] Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 7587.

[200] Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952.

[201] Eisenmenger, F.; Hansmann, U. H. E. *J. Phys. Chem. B* **1997**, *101*, 3304.

[202] Wang, T.; Wade, R. C. *Proteins* **2003**, *50*, 158–169.

[203] Guenot, J.; Kollman, P. A. *Proteins* **1992**, *9*, 1185–1205.

[204] Chen, J.; III, C. L. B.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

[205] Onufriev, A. *Annu. Rep. in Comput. Chem.* **2008**, *4*, 125–137.

[206] Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

[207] Lee, M. S.; Salsbury, F. R.; III, C. L. B. *Proteins* **2004**, *56*, 738–752.

[208] Narumi, T.; Yasuoka, K.; Taiji, K.; Höfinger, S. *J. Comput. Chem.* **2009**, *30*, 2351–2357.

[209] Kar, P.; Seel, M.; Weidemann, T.; Höfinger, S. *FEBS Lett.* **2009**, *583*, 1909–1915.

[210] London, F. *Z. Phys.* **1930**, *60*, 245.

[211] Hohm, U.; Trümper, U. *Chem. Phys.* **1994**, *189*, 443.

[212] Hohm, U.; Kerl, K. *Mol. Phys.* **1986**, *58*, 541.

[213] Höfinger, S.; Zerbetto, F. *Chem. Phys. Lett.* **2009**, *480*, 313–317.