Title
Computational Modeling of Protein Interactions at Multiple Lengthscales

Author
Yap, Eng Hui

Publication Date
2010

Computational Modeling of Protein Interactions at Multiple Lengthscales

by

Eng Hui Yap

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Joint Doctor of Philosophy
with University of California, San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Teresa Head-Gordon, Chair
Professor C. Anthony Hunt
Professor Berend Smit

Spring 2010

Computational Modeling of Protein Interactions at Multiple Lengthscales

Abstract

Computational Modeling of Protein Interactions at Multiple Lengthscales

by

Eng Hui Yap

Joint Doctor of Philosophy

with University of California, San Francisco

in Bioengineering

University of California, Berkeley

Professor Teresa Head-Gordon, Chair

We developed theories and algorithms for two coarse-grained implicit solvent models that can be deployed within a multiscale framework to enable computational studies of large-scale protein-protein associations. The first model is a *residue level* alpha-carbon bead model intended for simulating proteins at close range during formation of encounter complexes. This model introduces a novel forcefield term to model directional backbone hydrogen bond semi-explicitly, as well as a fourth bead flavor in its sequence-dependence to better represent the spectrum of residue-residue attractive interactions. We showed that the introduction of the orientation-dependent hydrogen bonding term resulted in more stable and realistic α–helices and β–sheets. In addition, the addition of a fourth bead flavor reduces energetic frustrations and competition from misfolded states. The overall model showed increased folding cooperativity, and a greater structural faithfulness to experimentally solved structures. The computational efficiency of the model has also permitted us to develop molecular models of the Alzheimer's A$\beta_{1\text{-}40}$ fibril to study nucleation and elongation[1, 2], providing a good proof-of-concept and laying the foundation for applications to other protein-protein assembly processes. The second model is a *protein level* model intended for simulating proteins during diffusional search. It treats proteins as rigid bodies interacting solely through long-range electrostatics. We first described the theory and implementation of a novel method, Poisson-Boltzmann Semi-Analytical Method (PB-SAM), to model electrostatic interactions by efficiently solving the linearized Poisson-Boltzmann equation (PBE). This novel method combines advantages of analytical and boundary element methods by representing the macromolecular surface realistically as a collection of overlapping spheres, for which polarization charges can then be iteratively solved using analytical multipole method[3]. Unlike finite difference solvers, PB-SAM is not constrained spatially by the box size, making it suitable for simulating dynamics. We showed that this method realizes better accuracy at reduced cost relative to either finite difference or boundary element PBE solvers. We derived expressions for force and torque that account for mutual polarization in both the zero and first order derivative of the surface charges, and incorporated the complete PB-SAM method into a

protein level Brownian dynamics simulation algorithm. We demonstrated for the first time dynamic propagation of multiple Brownian particles with accurate accounting of mutual polarization effects for successive timesteps, using a model system of two monomers of brome mosaic virus (PDB code: 1YC6[4]). While PB-SAM enable us to model mutual polarization effects in systems of hitherto inaccessible spatial dimensions, we can further reduce the computation time through parallelization, faster linear algebra operations, optimizing convergence criteria and polarization cutoffs, and approximating mutual polarization effects from analytical models. Finally, we discussed multiscale strategies to connect the two models described above for large-scale protein assembly studies. The two models can be employed successively in a novel nested variant of the Northrup-Allison-McCammon[5] formalism to compute bi-molecular kinetics rates. The kinetic parameters can in turn be inputs to chemical master equations or stochastic simulations. Such multiscale modeling can be used to determine kinetics rates and the order of association, and help investigate how changing physical interactions can alter the association rates, and consequently control overall sequences of association.

Dedicated to my parents.

Acknowledgements

This thesis would not have been possible without the infinite support and guidance from my thesis adviser, Prof. Teresa Head-Gordon. I will always remember that one December afternoon in 2003 when I, fresh off a plane ride from Singapore, made a cold call at her office and asked if I could join her lab. From that very first moment, Prof. Head-Gordon has been a caring mentor and a firm believer in her students' capabilities. Throughout the numerous technical challenges encountered in the course of this thesis work, her positive and indefatigable spirit, coupled with her patience and belief in me, enable us to scale the challenges and bring this work to fruition. In many different ways, she will serve as a role model, both professionally and personally.

Two members of the lab provide the solid foundation upon which this thesis work is based: Dr. Nicholas Fawzi implemented the forcefield for anisotropic hydrogen bonds for the residue-level protein model, which I extended by incorporating helical hydrogen bond, a fourth bead flavor, and optimized model parameters. Dr. Fawzi helped initiate my work on the residue-level model, and was always ready with a smile for more ideas and discussions. Dr. Itay Lotan derived an elegant analytical solution to the Poisson-Boltzmann equation for spheres, which motivated my development of the Poisson-Boltzmann Semi-Analytical Model. Dr. Lotan's well written source code also continues to be my best resource in programming.

I want to thank the inhabitants of Stanley 260 for scintillating lunchtime discussions, and for all childcare duties performed, voluntarily or entrusted upon, whenever I bring my two children to the lab. Thanks to Dr. Matt Lin for all technical support and many afternoon discussions, scientific or otherwise; Dr. Alex Sodt and Dr. Jerome Nilemeier for tutorials on normal mode analysis; Dr. Jonathan Kohn for extending my horizon with expositions on every conceivable topic; and Dr. Maggie Johnson and Shachi Katira for being great office mates.

I thank my parents for all the sacrifices they made in order to provide my brother and I with the best education, and for supporting my decision to pursue a doctoral degree ten thousand miles away from home. I am grateful to my brother Anthony for taking up the sole responsibility of caring for my parents in my absence. Finally, I want to thank my husband Bill for being a rock and reminding me to see humor during even the most frenzied times; and my two most wonderful children, Dina and Nikolas, for being the greatest kids a mother could wish for.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Protein-protein association is central to all biological processes. It is integral to a diverse range of functions, from catalysis, transport, immune response, signal transduction, transcription regulation, to maintaining cytoskeletal structure. Not surprisingly, protein-protein interactions cover a similarly broad spectrum in spatial and temporal scales, as well as varying levels of complexity. The interactions range from simple bi-molecular enzyme-substrate catalysis, to spatially inhomogeneous complexes involving multiple proteins such as the T-cell receptor macromolecular signalosomes[7], and to structurally uniform, macroscopic microtubules that span up to 25 μm.

Our knowledge of protein-protein interactions has been accumulated principally from biochemical and genetic experiments, including the widely used yeast two-hybrid screening protocol[8]. While simple binary protein-protein interactions can be characterized experimentally using biochemical methods[9-11], assemblies with three or more components entail complex cross-dependence, making it difficult to intuit mechanistic insights from experimental data alone. Computational modeling can complement the experimental effort to provide important molecular insights into structure and energetics, giving high throughput predictions of docking geometries and binding affinities[12-14]. In particular, computer simulations of large scale, multi-component assembly *processes* will provide hitherto unavailable information about the kinetics rates and mechanistic sequence of these complexation events.

Critical to successful computational modeling is selecting the appropriate level of theory and resolution that is commensurate with the research question. While all inter- and intramolecular interactions could in principle be investigated using *ab initio* quantum mechanics, it is neither tractable nor necessary for most purposes. In applications where the Born-Oppenheimer approximation is valid so electronics motions can be safely ignored (e.g. when simulating intact molecules without breaking any covalent bond), we can express energies and forces as a function of nuclei positions using empirical forcefields, whose parameters are derived from experiments and electronic structure calculations. Such forcefields (AMBER[15], CHARMM[16], GROMACS[17], etc.) form the integral model core of *molecular dynamics* (MD) simulation methods.

While classical MD simulations represent a significant speedup over *ab initio* quantum calculations, they are still too prohibitively expensive for studying large-scale protein complexes. In a typical MD simulation, biomolecules are solvated by explicit water molecules and ions; the position of each atom is then propagated by integrating Newton's law of motion over femtosecond timesteps. Although MD simulations of large *assembled* biological complexes in explicit solvent (~1 million atoms) could be accomplished in *tour-de-force* calculations deployed on state-of-the-art supercomputers[18-20], the steep computational cost of atomistic simulation with explicit solvent necessarily limits the simulation time to tens of nanoseconds. Atomistic simulation of the assembly *process*, which occurs over much longer time period (seconds to minutes), remains computationally intractable.

To study the *mechanistic process* of large-scale complexation, we must enhance sampling to collect relevant statistics with less timesteps; and/or reduce the computational cost per timestep. Longer timesteps can be used if we constrain intramolecular bond lengths, thereby avoiding the need to simulate bond vibrations at femtosecond timesteps. We can also enhance sampling of activated processes involving barrier crossing between metastable states, which are plagued by long time intervals between rare barrier crossings. In such cases, one can using parallel tempering[21] and metadynamics[22] techniques to accelerate barrier crossings. If one is principally interested with finding transition pathways between known states, the transition path sampling and its variants (nudged elastic band, string method)[23-25] can be used.

Alternatively, we can make large-scale simulations tractable by judiciously reducing the computational complexity per timestep. Fortunately, we are justified in using simplified models by two observations. Firstly, since water molecules (solvent) relax to their equilibrium positions and momenta quickly, and we are principally interested in the behavior of the proteins (solutes), we can replace the explicit water molecules with an implicit solvation model of water. Secondly, as will be elaborated later, the association process is inherently multiscale, consisting of a diffusion phase characterized by long time- and lengthscales, followed by a docking phase characterized by shorter time- and lengthscales. It is hence possible to further simplify the solute-solute interaction by incorporating the appropriate level of coarse-graining in space, time, and force field model. We shall first discuss the concept of implicit solvation, followed by coarse-grained models of solute-solute interactions.

(A) Implicit Solvation

Explicit solvent typically dominates the atom count in a simulated system, and requires extended simulation to ensure solvent configurations are equilibrated before data sampling can be done. Since we are interested in the behavior of the solute and not the solvent, we can dramatically reduce computational complexity by replacing the explicit water molecules with an implicit solvation model, which averages over all solvent degrees of freedom to produce potential of mean forces that act on the solutes. The procedure is valid because the relaxation time of the solvent is much faster than the large macromolecular solute. The model must account for three important effects of aqueous solvent: (i) temperature-dependent random collisions of water molecules with solutes, and the associated frictional drag on the solutes' motion; (ii) hydrophobic interaction; and (iii) electrostatic polarization by water and mobile ions. Below we elaborate on the nature of these effects, and survey how they are treated in implicit solvation models.

(i) Dynamics for Implicit Solvation Models

A solute molecule in a solution experiences constant, random collisions from all sides with solvent molecules, as well as a frictional drag on its motion. The Langevin equation[26, 27] describe the position of the solute, $\boldsymbol{r}$,

$$m\ddot{\mathbf{r}}(t) = -\xi\dot{\mathbf{r}}(t) + \mathbf{F}_C(t) + \mathbf{F}_R(t) \tag{1.1}$$

where m is the solute mass, ξ is the frictional constant, $\boldsymbol{F}_C(t)$ is the conservative (or systematic) force, and $\boldsymbol{F}_R(t)$ is a random force that is usually assumed to be Gaussian with an infinitely short correlation time.

We could further simplify the equation of motion in cases where the solute size and mass is much larger than that of the solvent molecules. In such cases the large number of solvent collisions with the solute averages out, allowing the solute momentum to relax quickly to its equilibrium distribution of $<m\mathbf{v}^2> = 3k_BT$. If we choose a timestep Δt to be within the diffusive regime, such that Δt is larger than the momentum relaxation time ($\tau = m/\xi$), yet smaller enough to ensure that $\boldsymbol{F}_c(t)$ is essentially constant, we could describe the solute position using Brownian dynamics. In a three-dimensional system with N solute molecules, the solute coordinate $r_i(t)$, where $1 \leq i,j \leq 3N$, is given by[28, 29]

$$r_i(t+\Delta t) = r_i(t) + \sum_j \frac{D_{ij}(t)F_{c,j}(t)}{k_BT}\Delta t + R_i(\Delta t) \tag{1.2}$$

where $D_{ij}(t)$ is the configuration-dependent diffusion tensor, and $R_i(\Delta t)$ is a Gaussian-distributed random displacement with zero mean and variance $2D_i(t)\Delta t$. The diffusion tensor could be implemented as Oseen or Rotne-Prager tensors[29], or simplified in isotropic cases (no hydrodynamic interaction between solutes) to a coefficient $D_{ij}(t) = D.\delta_{ii}$.

(ii) Hydrophobic Effect and Interaction

Experimentally, the hydrophobic *effect* refers to the fact that transferring a nonpolar solute molecule from gas to aqueous phase is an energetically uphill process, with a positive Gibbs free energy difference ($\Delta G_{transfer}$)[30].At room temperature, the hydrophobic effect is entropically driven: water molecules surrounding a nonpolar solute re-orientate themselves to maximize hydrogen bonds with other water molecules since they cannot form hydrogen bonds with the nonpolar solute[31, 32], resulting in a 'cage-like' water structure around the solute with decreased entropy. Hydrophobic *interaction* refers to the propensity for multiple nonpolar molecules or functional groups to aggregate with each other in water. It is the dominant driving force behind protein folding[33]. When multiple nonpolar solutes or intramolecular groups are present, the hydrophobic interaction is driven by the free energy difference between the entropically dominated solvation free energy of small molecules and the enthalpically dominated solvation free energy of their clustering into assemblies with large surface areas.

Implicit solvent forcefields represent the hydrophobic effect as an energetic penalty that is proportional to the "solvent accessible surface area (SASA)", i.e. the amount of hydrophobic surface area exposed to the solvent[34-37], although it has been argued that for small solutes that do not interrupt the hydrogen network, volume, not surface area, correlates better with hydrophobic effect[38]. Implicit solvent force fields can also represent the hydrophobic interaction for small hydrophobic groups as a potential mean force (PMF) that stabilize both an aggregated and solvent-separated configuration of two solutes species[39].

(iii) Electrostatics

In implicit solvation models, water is treated as a continuum, so the collective dielectric response of water (ε_w) includes contribution from each water molecule. A molecule responds to an external electric field through three physical processes: (i) electronic polarization, (ii) conformational change, and (iii) reorientation of permanent dipoles[40].A water molecule's strong permanent dipole (1.85 D), polarizability (α=1.415-1.528 Å^3)[41], and high number density, coupled with it's readiness to re-orientate cooperatively through the extensive hydrogen bond network, result in a high dielectric constant, ε_w ~ 78-80[40]. In contrast, the interior of a protein has a much lower dielectric constant, ε_p ~ 2-4 since large-scale reorientation of groups or domains are atypical[42]. Lastly, implicit solvation models account for mobile ions in bulk electrolytes using a mean field theory, where the distribution of each ion species is assumed to obey Boltzmann's statistics.

The electrostatic potential $\Phi(\mathbf{r})$ of the above continuum system is fully described by the *nonlinear* Poisson Boltzmann equation (PBE), in e.s.u-c.g.s. convention,

$$-\nabla\left[\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})\right]-4\pi\sum_i \bar{n}_i Z_i \exp(-\frac{eZ_i\Phi(\mathbf{r})}{k_BT}) = 4\pi\rho_{fixed}(\mathbf{r}) \qquad (1.3)$$

where ε is the relative dielectric function, ρ_{fixed} is the charge density due to the fixed protein partial charges, $\bar{n}_i$ and Z_i are the bulk concentration and valence of ion species i respectively, e is the fundamental electronic charge, k_B the Boltzmann constant, and T the absolute temperature. The PB theory inherently assumes that (i) ions are dimensionless, (ii) the potential of mean force experienced by each ion is equal to the mean electrostatic potential.

Further assumptions can be made to simplify the PB theory. The nonlinear PB equation can be linearized in cases where the salt is monovalent and $q\Phi/k_BT << 1$, to yield the *linearized* PB equation:

$$-\nabla\left[\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})\right]+\kappa^2\Phi(\mathbf{r}) = 4\pi\rho_{fixed}(\mathbf{r}) \qquad (1.4)$$

where $\kappa = \sqrt{8\pi\bar{n}e^2/\varepsilon_w k_B T}$ is the inverse Debye length.

Solution of the PBE constitutes the most computationally intensive part of simulating protein association. Approaches to the PBE can be broadly categorized into analytical and numerical methods. Analytical solutions can be quickly computed, but are only available for certain idealized geometries (sphere, cylinder, and infinite plane). In contrast, numerical methods - finite difference (FD), finite element (FE) and boundary element methods (BEM) can handle realistic dielectric boundaries but are more computationally intensive. Reference [19] presents an excellent survey of current PBE methods, and below we highlight some of the most salient approaches that relate to this thesis.

FD methods, such as DelPhi[43] and APBS[44], are most commonly used due to its ease of implementation, and the large body of computational tools developed for solving sparse matrix linear algebra problems. However, FD methods do not impose continuity in electric displacement ($\varepsilon\boldsymbol{E}$) across the dielectric boundary, affecting their accuracy and convergence. More importantly, FD and FE methods are spatially constrained to the grid or mesh, making them unsuitable for multi-molecular simulations in which molecules could be separated by large distances.

Boundary element methods[45-51] formulate integral equations on the surface of the molecules and solve for the potential and field on the surface. They impose both potential and electric displacement continuity by construction, and have reduced number of unknowns since the unknowns are on the molecular surface, not the volume. More importantly, they are not constrained to grid points and hence suitable for dynamic simulations. Historically, BEM methods have been plagued by expensive memory requirement and dense interaction matrices that scale with $O(N^2)$, where N is the number of boundary elements. Recent implementations of BEM[46, 48] that employs adaptive Fast Multipole Method (FMM) have achieved significant speed up, making them comparable to FD methods in timing and memory requirements. However, simulation of multiple proteins has not been demonstrated, except for simple test cases involving two spheres with monopole charge in vacuum.

The Head-Gordon group has recently derived an analytical method using multipoles to solve the linearized PBE for $N>2$ spherical molecules[3]. This method forms the basis of a semi-analytical approach, PB-SAM, to solve the linearized PBE by representing the macromolecular surface as a collection of overlapping spheres, for which polarization charges can then be iteratively solved using analytical multipole methods. Unlike finite difference solvers, PB-SAM is not constrained spatially by the box size, making it suitable for dynamics. This method realizes better accuracy at reduced cost relative to either finite difference or boundary element PBE solvers.

(B) Coarse-Graining Solute-Solute Interactions

A coarse-graining strategy can be motivated from the changing nature of the intermolecular interactions as two proteins approach each other. The association is comprised of two steps[52]: a diffusional search to form a mostly solvated encounter complex; followed by structural rearrangement and desolvation to form a docked complex (Figure 1.1). During the diffusion phase, intermolecular forces are dominated by long-range electrostatics[53]. In addition, conformational fluctuations of the macromolecules are insignificant compared to the lengthscale of their separation. Hence we can adopt a ***protein-level*** model, in which each molecule is represented as a rigid body interacting through electrostatic forces. During the formation of the encounter complex, short-range interactions such as hydrophobic interaction become significant. In addition, conformational changes are now comparable to the separation lengthscale, so a ***residue-level model*** becomes necessary to ensure correct sampling of the conformations. Finally, at short separation distances during docking, we must fall back to atomistic representations of the proteins. Simulations at atomistic resolutions can be performed using available MD software such as AMBER and CHARMM. This thesis work focuses on the development of the residue-level and protein-level models, to enable simulations across various resolutions.

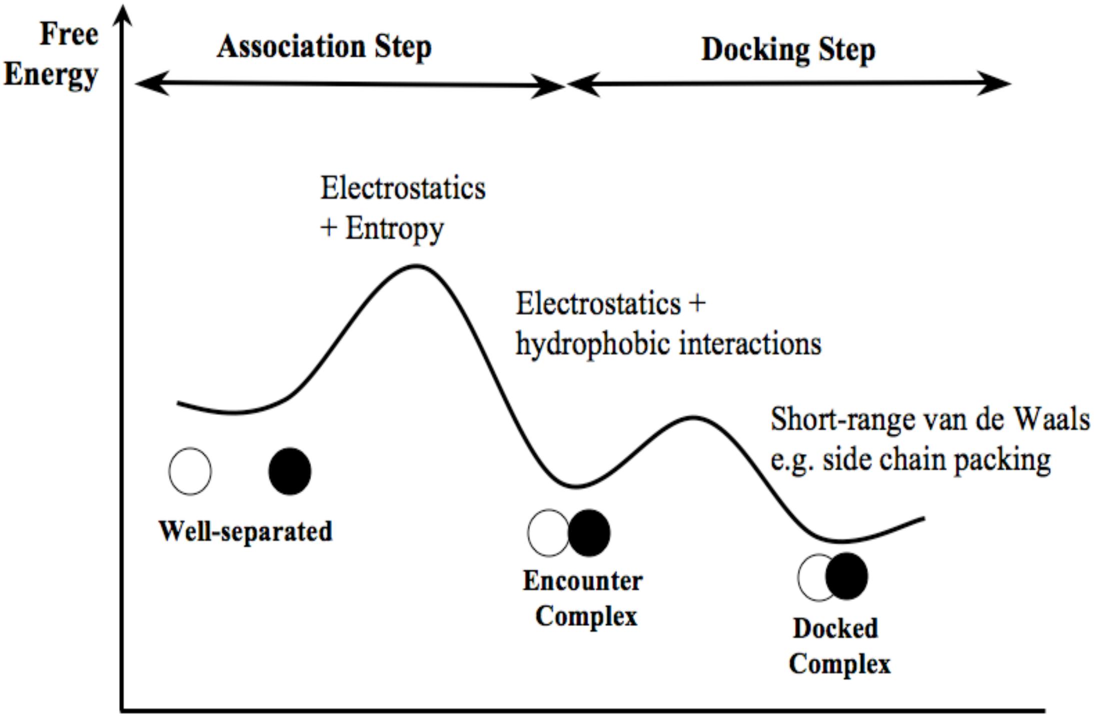

Figure 1.1 Schematic of a typical protein association pathway. Figure adapted from reference [52].

Residue-level coarse-grained models provide a cheap way to introduce residue-residue interactions and backbone flexibility. Depending on the specific models, each residue can be represented by one or more interacting centers. In the original Gō model for protein folding studies[54], the protein is represented as a chain of one-bead amino acids having attractive interaction between native contacts, and repulsive interactions between non-native contacts. The folding rate is then primarily correlated with the topological complexity of the native state, from which the folding pathways and the thermodynamics (and kinetics) of folding can be reasonably inferred[55].

However, completely unfrustrated Gō models fail to account for intermediate metastable folding states and different folding mechanisms amongst proteins with similar topologies. Head-Gordon *et*. *al*.[56, 57] introduced sequence specificity into their alpha-carbon only model through non-bonded terms describing hydrophobicity with three flavors (hydrophobic, neutral, polar). The model is able to discriminate the different folding behavior of proteins L and G, which have the same native topology. While the addition of sequence-specificity made the Head-Gordon model more realistic, its limited flavors did not reflect the graded spectrum of hydrophobicity amongst the 20 naturally occurring amino acids. This introduces large frustrations in the folding funnel, resulting in many degenerate misfolds that compete with the native conformation. In addition, one-bead models lack the anisotropy to stabilize secondary structures and account for cooperativity in their formation.

Chapter 2 describes a sequence-based $\alpha-$carbon model that we developed to incorporate a mean field estimate of the orientation dependence of the polypeptide chain that gives rise to specific backbone hydrogen bond pairing to stabilize $\alpha-$helices and $\beta-$sheets. Compared to a previous 3-flavor model without hydrogen bond developed in the Head-Gordon group[57], the new model shows greater folding cooperativity and improvements in designability of protein sequences, as well as predicting correct trends for kinetic rates and mechanism for immunoglobulin proteins L and G. This residue-level model has been applied to study protein-protein interactions in Aβ 1-40 peptide aggregation[1, 2].

Chapter 3 describes the theory and implementation of a new approach for solving the linearized PBE – the Poisson-Boltzmann Semi-Analytical Method (PB-SAM). This method represents the macromolecular surface as a collection of overlapping spheres, for which polarization charges can then be iteratively solved using analytical multipole method[3]. Unlike finite difference solvers, PB-SAM is not constrained spatially by the box size, making it suitable for dynamics. This method realizes better accuracy at reduced cost relative to either finite difference or boundary element PBE solvers. We illustrate the strength of the PB-SAM approach by computing the potential profile of an array of 60 T1-particle forming monomers of the bromine mosaic virus.

Chapter 4 incorporates the PB-SAM method within the framework of a Brownian dynamics simulation algorithm, where molecules are treated as rigid bodies. We describe the variational formalism for force and torque, and the method to solve for gradient of the interaction energy. Timing data for force and torque calculations were then reported for a system with two bromine mosaic virus monomers.

Finally, chapter 5 surveys multiscale strategies currently used to connect simulations across varying resolutions. We propose ways in which the coarse-grained residue level model (Chapter 2) and the protein-level model (Chapter 3 and 4) that we developed can be deployed in a multiscale framework to study protein-protein association, by using either a serial, nested framework or feeding kinetic parameters from our coarse-grained simulations into stochastic simulations.

Chapter 2

A Coarse-Grained alpha-Carbon Protein Model with Anisotropic Hydrogen-Bonding[*]

Introduction

Understanding the general energetic principles of protein self-assembly is a long-standing problem in biophysical chemistry. Recently, the framework of energy landscape theory has provided direction in the design of protein folding models that should exhibit correct folding thermodynamics by optimization of a funneled free energy surface[59-61]. The spatial resolution of the models do not have to be at full atomic detail since it is well known that models with sufficient topological features (correct sequence distribution of local and non-local spatial contacts) are sufficient for reproducing trends in thermodynamic and even kinetic folding data.[55]

Inspired by early efforts of Thirumalai and co-workers[62-65], we have developed a "minimalist" protein bead model that uses an α-carbon (C_α) trace to represent the protein backbone, in which structural details of the amino acids and aqueous solvent are integrated out and replaced with effective bead-bead interactions. These physics-based potentials are formulated so that there is still a connection between bead type and amino acid sequence in a reduced letter code, and hence stand distinct from Go-based potentials.[54] We have successfully used the coarse-grained protein model to study the folding mechanism and kinetics of several proteins of the ubiquitin α/β topology, and to analyze folding simulation protocols[56, 57, 66-69], for competition between folding and aggregation in which we correlate differences in aggregation

[*] Reproduced with permission from reference [58].

kinetic rates to differences in structural populations of unfolded ensembles[70] and most recently in disease aggregation processes relevant for the Aβ peptide indicted in Alzheimer's disease[1, 2].

When the experimental folding and aggregation data to be understood is of higher spatial or timescale resolution, then isotropic interactions used in protein bead models may break down. One example is the study of early molecular origins of amyloid fiber formation for the Aβ peptide, in which the mature amyloid aggregate has a precise morphology of unbranched fibers composed of parallel intermolecular β-sheets.[71] To understand these more complex protein assembly or co-assembly problems, it is important to both retain the efficiency of a single bead C_α model while incorporating some of the orientation dependent properties of amino acids in protein structures. Several models formulated in this spirit include the extension of bead Go-potentials with orientation-dependent statistical potentials[72], or amino acid specific residue-residue distances[73].

More closely related to this work are formulation of backbone hydrogen bond potentials in the context of off-lattice bead models[60, 74-76]. Onuchic and Cheung incorporated an implicit hydrogen bond in terms of a pseudo-dihedral angle between four C_α centers straddling two separate beta-strands potential within their Go model that uses two centers per residue[77]. However, their formulation incorrectly assumes that the strands' C_α centers and hydrogen bonds lie in the same plane, when in fact hydrogen bonds are roughly perpendicular to the planes described by the C_α centers. Brooks and co-workers (private communications) use a three bead per residue model in which the C_α centers are straddled by additional centers embedded with a point dipole to represent the carbonyl and amide peptide linker. The work by Klimov and Thirumalai[74, 75] approximates virtual positions of CO and NH moieties based on C_α positions, which are then used to determine whether the strands are well oriented to form hydrogen bonds. However, their implementation only takes into account hydrogen bond directionality and not hydrogen bond distance, and as a result the folding transition does not exhibit great cooperativity, with folding transitions occurring over a broad temperature range. Furthermore, their model is only effective for α−helical and anti-parallel β−sheet structures, but could not adequately describe parallel β−sheets. The protein model of Smith and Hall[76] uses a four center residue in which hydrogen-bonds are described as pseudo-bonds between residues to restrict both distance and orientation to realize α−helical and β−sheet structure. In all of these coarse-grained models, the additional centers per residue, scales up the computational cost by $\sim(cN)^2$, where c is the number of centers per residue.

In this work we propose a reformulation of a one-site α−carbon model to introduce a fourth bead flavor, new dihedral angle potentials, and a potential of mean force hydrogen bonding term that encourages the cooperative formation of protein-like secondary structures. The orientation-dependent hydrogen bonding term is based on a similar *functional form* developed by Ben-Naim[78] and later adopted by Silverstein and co-workers[79] to characterize hydrogen-bonding in a model of bulk water. Our protein model now incorporates a mean field estimate of the orientation dependence of the polypeptide chain that give rise to specific hydrogen bond pairing to stabilize α−helices and β−sheets. The model is first parameterized for protein G (PDB code: 2GB1)[80], and then validated using folding studies of protein L (PDB code: 2PTL)[81]. As we show in the Results, the model shows improvements in designability and greater folding cooperativity, and kinetic rates and mechanistic outcomes consistent with experiment.

Models and Methods

Energy Function

The modified minimalist model potential energy function is given by

$$E = \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} \begin{bmatrix} A[1+\cos(\phi+\phi_o)] + B[1-\cos(\phi+\phi_o)] + \\ C[1+\cos 3(\phi+\phi_o)] + D[1+\cos(\phi+\phi_o+\frac{\pi}{4})] \end{bmatrix} \tag{2.1}$$
$$+ \sum_{i,j\geq i+3} 4\varepsilon_H S_1 \left[(\frac{\sigma}{r_{ij}})^{12} - S_2 (\frac{\sigma}{r_{ij}})^6 \right] + \sum_{Hbonds} U_{HB}$$

where θ is the bond angle defined by three consecutive C_α beads, ϕ is the dihedral angle defined by four consecutive C_α beads, and r_{ij} is the distance between beads i and j. The hydrophobic strength ε_H sets the energy scale. The bo0nd angle term is a stiff harmonic potential with force constant $k_\theta = 20\ \varepsilon_H$ /rad^2. The optimal bond angle θ_0 for bead i set to 95° if bead i-1 has helical dihedral propensity, and 105° otherwise. Each dihedral angle in the chain is designated to be one of the following types: helical (H), extended (E), or turns (T, P, U, or Q). The third term in Eq. 2.1 represents nonlocal interactions, and is determined according to the bead flavors: strong attraction (B), weak attraction (V), weak repulsion (N), and strong repulsion (L). The last term represents a new distance and orientation dependent potential that models backbone hydrogen bond explicitly. We describe these new features in more detail below.

Our model has been extended to now include new dihedral types in the turn region. As C_α-only models lack chirality, we introduced -/+90° turns (designated Q and P respectively) to distinguish the native topology from its mirror image decoys, and 0° dihedral (designated U) to impose some rigidity in hairpin turns. The parameters A, B, C, D, and ϕ_0 are chosen to produce the desired minima (Table 2.1). In accordance with the flexible nature of turn regions, these new dihedral types are weaker in strength than their helical and extended counterparts. While all dihedral types encourage formation of the assigned secondary structures, they also allow access to other competing local secondary minima through manageable (~1 - 2.8ε_H) barriers.

Table 2.1. Parameters for various Dihedral Types

Dihedral Type	A (ε_H)	B (ε_H)	C (ε_H)	D (ε_H)	k	ϕ_0 (rad)	Local minima (global minima in **bold**)
H (Helical)	0	1.2	1.2	1.2	1	+0.17	-65°, **+50°**, 165°
E (Extended)	0.45	0	0.6	0	1	-0.35	**-160°**, -45°, +85°
T (Turn)	0.2	0.2	0.2	0.2	1	0	**-60°, +60°, +180°**
P (+90°)	0.36	0	0.48	0	1	+1.57	-155°, -25°, **+90°**
Q (-90°)	0.36	0	0.48	0	1	-1.57	**-90°**, +25°, +155°
U (0°)	0.36	0	0.48	0	1	+3.14	-120°, **+0°**, 120°

We have also increased the number of bead flavors from three of our original model to four in our new model by adding a weak attractive bead (denoted V). The amino acid sequence of a protein can be mapped to its four-flavor sequence using the mapping rule shown in Table 2.2, and the bead types determine the type of non-bonded interaction between two beads (Figure 2.1). The attractive interactions B–B, B-V and V-V all have $S_2 = -1$, while S_1 = 1.4, 0.7, and 0.35 respectively. For repulsive interactions, $S_1 = 1/3$ and $S_2 = -1$ for L–L, L-V, and L–B interactions; and $S_1 = 1$ and $S_2 = 0$ for all N–X interactions. The sum of van de Waals radii σ is set at 1.16 to mimic the large exclusion volume due to side chains.

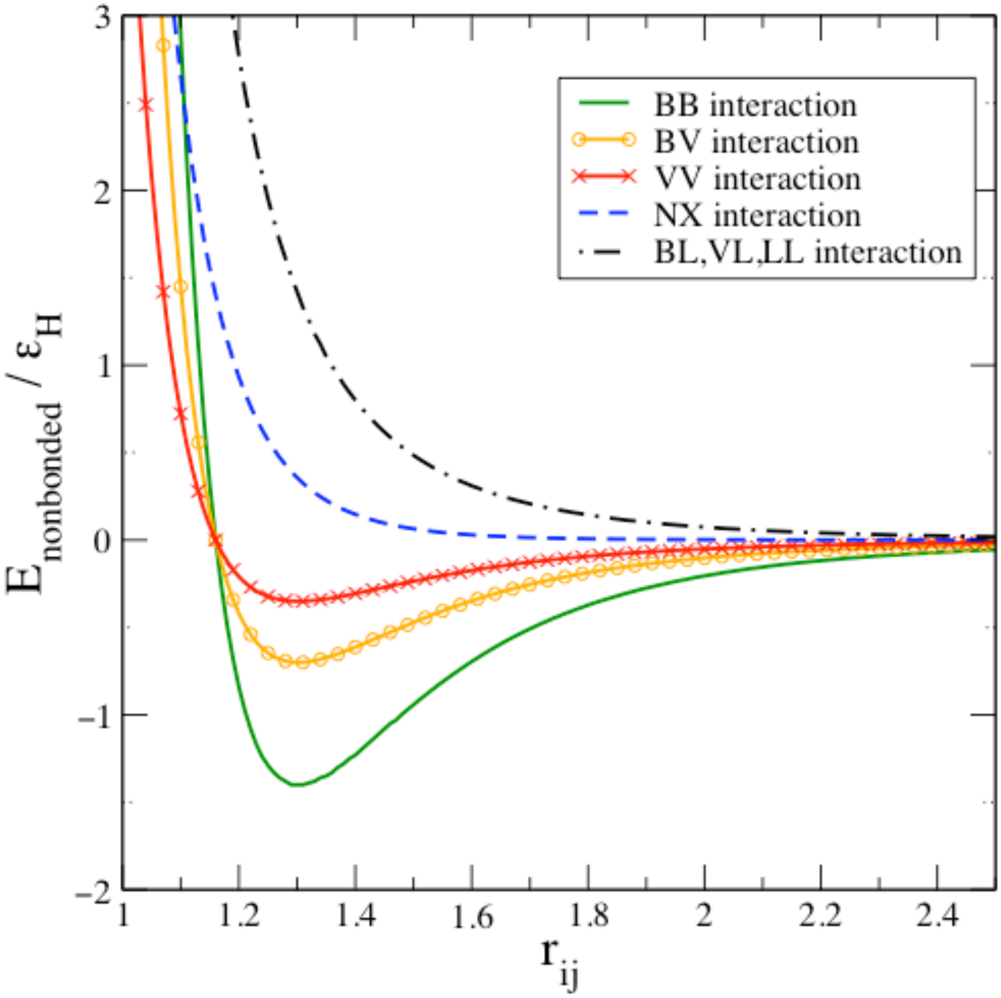

Figure 2.1. ***Non-bonded hydrophobic interaction energy as a function of pair-wise distance between bead i and j.*** Interactions BB, BV, and VV have attractive minima at $r_{ij} = 1.3$ while NX and BL/VL/LL interactions are purely repulsive.

Table 2.2. Mapping 20-letter (20) amino acid code to four-letter code (4)

(20)	(4)	(20)	(4)	(20)	(4)	(20)	(4)
Trp	B	Met	B	Gly	N	Asn	L
Cys	B	Val	B	Ser	N	His	L
Leu	B	Ala	V	Thr	N	Gln	L
Ile	B	Tyr	V	Glu	L	Lys	L
Phe	B	Pro	N	Asp	L	Arg	L

Lastly, we have added a new term to the Hamiltonian to describe a pair-wise mean force hydrogen bond interaction U_{HB}, inspired by the Mercedes Benz (MB) model of water first introduced by Ben-Naim[78] and further developed by Silverstein and co-workers[79]. In the original MB model, water molecules are represented as two-dimensional discs with three symmetrically arranged arms, separated by an angle of 120°. Water molecules interact through a standard Lennard-Jones term and an explicit hydrogen-bonding (HB) interaction that is favorable when the arm of one molecule aligns with the arm of another. We have adapted the functional form of the hydrogen bonding interaction to our three-dimensional minimalist protein model. The hydrogen bond potential between two beads *i* and *j* is given by:

$$U_{HB} = \varepsilon_{HB} F\left(r_{ij} - r_{HB}\right) G\left(\left|\mathbf{t}_{HB,i} \cdot \hat{\mathbf{r}}_{ij}\right| - 1\right) H\left(\left|\mathbf{t}_{HB,j} \cdot \hat{\mathbf{r}}_{ij}\right| - 1\right) \tag{2.2}$$

where

$$\begin{gathered} F(r_{ij} - r_{HB}) = \exp\left[-\left(r_{ij} - r_{HB}\right)^2 / \sigma_{HBdist}{}^2\right], \\ G\left(\left|\mathbf{t}_{HB,i} \cdot \hat{\mathbf{r}}_{ij}\right| - 1\right) = \exp\left[\left(\left|\mathbf{t}_{HB,i} \cdot \hat{\mathbf{r}}_{ij}\right| - 1\right) / \sigma_{HB}{}^2\right], \\ H\left(\left|\mathbf{t}_{HB,j} \cdot \hat{\mathbf{r}}_{ij}\right| - 1\right) = \exp\left[\left(\left|\mathbf{t}_{HB,j} \cdot \hat{\mathbf{r}}_{ij}\right| - 1\right) / \sigma_{HB}{}^2\right] \end{gathered} \tag{2.3}$$

where r_{ij} is the distance and $\hat{\mathbf{r}}_{ij}$ the unit vector between beads *i* and *j* respectively. The distance dependent term *F* is a Gaussian function centered at the ideal hydrogen bond distance r_{HB}. For the direction dependent terms *G* and *H*, we use an exponential instead of a Gaussian function to ensure a smoother potential energy surface. The vectors $\boldsymbol{t}_{HB,i}$ and $\boldsymbol{t}_{HB,j}$ are unit vectors normal to the planes described by bead centers (i-1, i, i+1) and (j-1, j, j+1) respectively. The ideal hydrogen bond distance r_{HB} is set to 1.35 length units for α–helices and 1.25 length units for β–sheets in accordance with a survey of secondary structures in the PDB database. All other hydrogen bond parameters are identical for α–helices and β–sheets, with the width of functions *F*, *G* and *H* set by $\sigma_{\mathrm{HBdist}} = \sigma_{\mathrm{HB}} = 0.5$.

The hydrogen bond potential is evaluated for all *i-j* bead-pairs capable of forming hydrogen bonds. Depending on its dihedral propensity, each bead is assigned a hydrogen bond forming capability from three possible types: sheet (designated D), helical (designated A), or none (designated N). For a bead assigned D, the hydrogen bond potential is evaluated between itself and all D-beads situated within a cutoff distance of 3.0 length units. For a bead assigned A, helical hydrogen bond potential is evaluated if its +3 neighbor is similarly assigned A. We find that the helical hydrogen bond is better modeled in a C_{α}-only model as an interaction between (i,i+3) bead pairs, rather than (i,i+4). From a survey of helices in the PDB, the distribution of

$r_{i,i+3}$ has both a smaller mean and variance than $r_{i,i+4}$. Hence a potential using (i,i+3) bead pairs is more stringent in discriminating between helical and non-helical geometry.

The strength of the hydrogen bond is modulated by ε_{HB}, which is set to $0.7\varepsilon_H$ if the bead pair is B-B, B-V or V-V. For L-X and N-X pairs, a higher ε_{HB} of $0.98\varepsilon_H$ is required to compensate for the non-bonded repulsion. This provides anisotropy in our C_α-only model: L and N residues could maintain closer contact with their hydrogen bonding partners, while remaining repulsive to beads in all other directions.

Protein Model

The structural, thermodynamic, and kinetic properties of protein L and G have been well characterized experimentally[82-91]. Both proteins consist of an N-terminus hairpin, made up by β-strands 1 and 2, followed by a helix, and lastly a C-terminus hairpin made up by β-strands 3 and 4. Despite their similar topologies, L and G share only 15% sequence identity, and fold via different mechanisms[92]. Experimental studies have shown that while the transition state of protein L consists of partially formed β-hairpin 1[84, 91], that of protein G comprises of partially formed β-hairpin 2[86, 93]. Our existing sequence-based model has been shown capable of predicting the mechanistic differences in L and G folding[57], something not possible with Go potentials.

Here we show that our new model preserves this sequence-based feature, and can thus replicate the different folding mechanisms of L and G. In developing the model we optimized the potential energy parameters for protein G in order to reliably reach a global minimum corresponding to the native state topology using simulated annealing, as well as yield reasonable thermodynamics such as sharp cooperative melting curves and heat capacities. We then fixed those parameters to validate the model by characterizing the kinetic mechanism of protein G, as well as the thermodynamics and kinetic mechanism of protein L.

The resulting amino acid sequences of proteins L and G were mapped to reduced minimalist code as per Table 2.2. The dihedral angle propensities were assigned according to their respective PDB structures, with the hairpin turns described using P, U, and Q to encourage the correct chirality. Since we wish to focus on whether differences in the folding behaviors are due to sequence, we assign identical dihedral propensities to hairpins in both L and G. However, the first hairpin turn in protein L (Phe, Ala, Asn, Gly, Ser) is one residue longer than that of protein G (Gly, Lys, Thr, Leu). To address this we use a modified sequence for protein L in which the 11th residue (Asn) is omitted. Dihedral propensities in the hairpins in both proteins can now be similarly assigned for fair comparison. The hydrogen bond forming capability (A, B, or N) follows the dihedral specification above. The mapped sequence, dihedral propensity and hydrogen bond assignments are listed in Table 2.3.

The initial mapping of the primary sequence from the 20-amino acid code to the 4-letter minimalist code contains some ambiguity. For instance, lysine has both a long hydrocarbon chain and a charged amine group, and could be treated as either hydrophilic or hydrophobic. The initial energy landscape contains many competing local minima due in part to such ambiguity. Sequence design based on the minimal frustration principle is done to smooth the potential energy surface and improve foldability. Our sequence design strategy is based on the theoretical criterion[26-28] that a foldable heteropolymer sequence has a significant energy gap ΔE between its native-state energy E_{native} and average misfold energy $\langle E_{misfold} \rangle$. Using our initial mapping sequence, we generate a library of misfolded (non-native) structures from simulated annealing.

To obtain a better folding sequence, we generated sequences with various single mutations, threaded them to structures in the misfold library, and select the mutant sequence that maximizes the energy gap ΔE. To minimize drift from the original sequence, we allow only single mutations of types B↔V, V↔N, or N↔L, or dihedral mutations. The mutation process is repeated until we obtain a foldable sequence that finds the native state reliably 50% of the time using simulated annealing.

Table 2.3. Sequence, dihedral, and hydrogen bond assignments for proteins L and G

Protein L	
1° 2PTL	VTIKANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEYTVDVADKGYTLNIKFAG
1° 2PTL (without *Asn-11)*	VTIKANLIFAGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEYTVDVADKGYTLNIKFAG
1° model L (mapped):	BNBLVLBBBVNNNLNVLBLNNBLLVNNLVVVVVLNBLLLLNLVNBLBVLLNVNBLBLBVN
1° model L (optimized):	**N**NBLV**N**B**NVN**NNNLNVL**V**LNNBLLVNNLVVVV**BN**N**V**LLLLNLVN**V**L**V**VLLNVNBLBLB**N**N
2° model L:	EEEEEEEQUPEEEEEEEET**P**THHHHHHHHHHHHHHHHT**PU**EEEEEEEEEEPUQEEEEEEEE
Hbond model L:	DDDDDDDDNNDDDDDDDDDNNAAAAAAAAAAAAAAAANNNNNDDDDDDDDNNNDDDDDDDDD
Protein G	
1° 2GB1	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
1°model G (mapped):	BNVLBBBLNLNBLNLNNNLVBLVVNVLLBBLLVVLLLNBLNLBNVLLVNLNBNBNL
1°model G (optimized):	**V**NVLB**N**BLNLN**V**LNLNNNLVBLV**N**N**N**LL**V**BLLVVLLLN**V**LNL**V**NVL**N**VN**N**NBNBN**N**
2°model G:	EEEEEEEQUPEEEEEEEETTT**Q**HHHHHHHHHHHHHHHTTTTTEEEEEPUQEEEEE
Hbond model G:	DDDDDDDDNNDDDDDDDDDNNNAAAAAAAAAAAAAAAANNNNDDDDDDDNNDDDDDD

Simulation Protocol

All simulations are performed in reduced units with mass *m*, energy ε_H, length σ_0, and k_B set to unity. The bond length between adjacent C_α beads serves as the unit of length σ_0, and is held rigid by using the RATTLE algorithm.[94] Reduced temperature and time are given by $T^*=\varepsilon_H/k_B$ and $\tau = (m\sigma_0^2/\varepsilon_H)^{1/2}$ respectively. We use constant-temperature Langevin dynamics with a friction coefficient of $0.05\tau^{-1}$, and a timestep of 0.005τ to perform simulations for characterizing the thermodynamics and kinetics of folding.

For each simulated annealing run we launch 50 trajectories at a high temperature (T* = 1.6) and evolve them for 1250τ to generate uncorrelated, unfolded conformations, then gradually cool these trajectories to T* = 0.1 for 7500τ. The trajectories are then annealed at T* =0.45 for 50τ, and cooled for 5000τ to T* = 0.1, and the anneal-cool cycle repeated once more before the resulting structure is quenched from T* = 0.1 to T* =0.

The free energy landscape is characterized with the multidimensional histogram technique.[95, 96] We collect multiple nine-dimensional histograms over energy E, radius of gyration Rg, number of native contacts formed Q, number of native contacts formed between strand 1 and strand 2 ($Q_{\beta1}$), number of native contacts formed between strand 3 and 4 ($Q_{\beta2}$), and native-state similarity parameters χ, χ_α, $\chi_{\beta1}$, and $\chi_{\beta2}$, where χ is given by

$$\chi = \frac{1}{M}\sum_{i,j\geq i+4}^{N} h\left(\varepsilon - \left|r_{ij} - r_{ij}^{native}\right|\right) \tag{2.4}$$

The double sum is over beads on the chain, and r_{ij} and r_{ij}^{native} are the distances between beads i and j in the state of interest and the native state, respectively; *h* is the Heaviside step function, with $\varepsilon = 0.2$ to account for thermal fluctuations away from the native-state structure. M is a normalizing constant to ensure that $\chi= 1$ when the chain is identical to the native state and $\chi \approx 0$ in the random coil state. The remaining χ parameters are specific to their respective elements of secondary structure. That is, χ_α involves summation over beads in the helix, and $\chi_{\beta 1}$ and $\chi_{\beta 2}$ involve summation over beads in the first and second β-sheet regions, respectively.

From the histogram method, we get the density of states as a function of nine order parameters, Ω(E, Rg, Q, $Q_{\beta1}$, $Q_{\beta2}$, χ, χ_α, $\chi_{\beta 1}$, $\chi_{\beta 2}$), which can be used to calculate thermodynamic quantities. In constructing the free energy surfaces, we collect histograms at 14 different temperatures: 1.30, 1.00, 0.80, 0.60, 0.50, 0.40, 0.38, 0.36, 0.34, 0.32, 0.30, 0.25, 0.20, and 0.15. We run five to eight independent trajectories at each temperature and collect 4,000 data points per trajectory. The potential of mean force w along reaction coorindate Q is given by

$$w(Q') = -kT \ln \left[\begin{array}{c} \int dE dR_g dQ dQ_{\beta 1} dQ_{\beta 2} d\chi d\chi_H d\chi_{\beta 1} d\chi_{\beta 2} \delta(Q - Q') \\ \Omega\left(E, R_g, Q, Q_{\beta 1}, Q_{\beta 2}, \chi, \chi_H, \chi_{\beta 1}, \chi_{\beta 2}\right) e^{-E/kT} \end{array} \right] \tag{2.5}$$

where the δ-function is approximated using Gaussian functions.

The folding kinetics is studied using mean first passage time (MFPT) based on a native state cut-off. With the MFPT method, we decorrelate 2000 independent trajectories at T* = 1.6 for 1,250τ, jump to the temperature of interest, and continue evolving the trajectories. We recorded the time τ_i that each trajectory took to enter the native basin of attraction, defined as Q>0.8. The fraction of trajectories folded at time *t* is then calculated by $P_{nat}(t)$ = (no. of trajectories with τ_i < t)/N. Analysis of the $P_{Nat}(t)$ kinetic data are detailed in Results and Discussions.

Studies of transition state (TS) ensembles are performed using the P_{fold} analysis method[97]. Noting that true transition states should (a) be the highest point along the minimum free energy path, and (b) sit on saddle points on the multi-dimensional landscape, we first identify putative transition states from various projections of order parameters onto the free energy surface. Because we are vetting the new model against a known mechanism, we focused our free energy projections for protein L and G along the order parameters Q and/or $\chi_{\beta 1}$ and Q and/or $\chi_{\beta 2}$, respectively, in order to collect putative TS structures. P_{fold} analysis is then performed: for each putative TS structure, we launch 100 trajectories at the folding temperature, evolve them for 1000τ, and evaluate the probability (P_{fold}) that these trajectories fall into the folded basin (defined as Q>0.8). Structures with $0.4 \leq P_{fold} \leq 0.6$ are considered to be part of the TS ensemble.

Results and Discussions

Sequence Design and Native Structures

We obtained an optimized sequence for protein L after twelve sequence mutations and three dihedral mutations, while the optimized sequence for protein G consists of nine sequence mutations and one dihedral mutation. Table 2.3 compares the optimized sequences to their original mapping. We find that the original mapping is robust since 50% of the sequence mutations involved ambiguous definitions of valine (B or V) or alanine (V or N), and thus could be explained by these amino acids being 'borderline' on the hydrophobic scale. We find a trend that valines and alanines in the core tend to be retained as B and V (more strongly hydrophobic), while those on the periphery are mutated to V and N (less hydrophobic).

We performed simulated annealing using these optimized sequences to obtain the lowest energy structures (Figure 2.2). First we compare the structural similarity of the native state of our protein L and G models with the experimental structures using the Combinatorial Extension (CE) method[98]. The CE algorithm excludes loop α−carbon positions to align the model and solution structures despite the different lengths of the loop regions. Using the CE method the new model gave RMSDs of 2.6Å for Protein L and 3.0Å for protein G, compared to the old model RMSDs of 4.4Å for Protein L and 5.3Å for protein G.[57] We also calculated the root mean square distance (RMSD) of C_α atoms between these simulated native structures and their NMR counterparts using the *rms.pl* script from the MMTSB toolbox.[99] To ensure a stringent comparison, this time we do not allow gaps or deletions in our alignments, although we modified the 2PTL coordinate file to omit Asn-11 to allow a bead-to-bead comparison with our 60-bead model of protein L. The calculated RMSDs of our simulated native structures are 4.4Å for Protein L and 3.0Å for protein G using the alignments with no gaps.

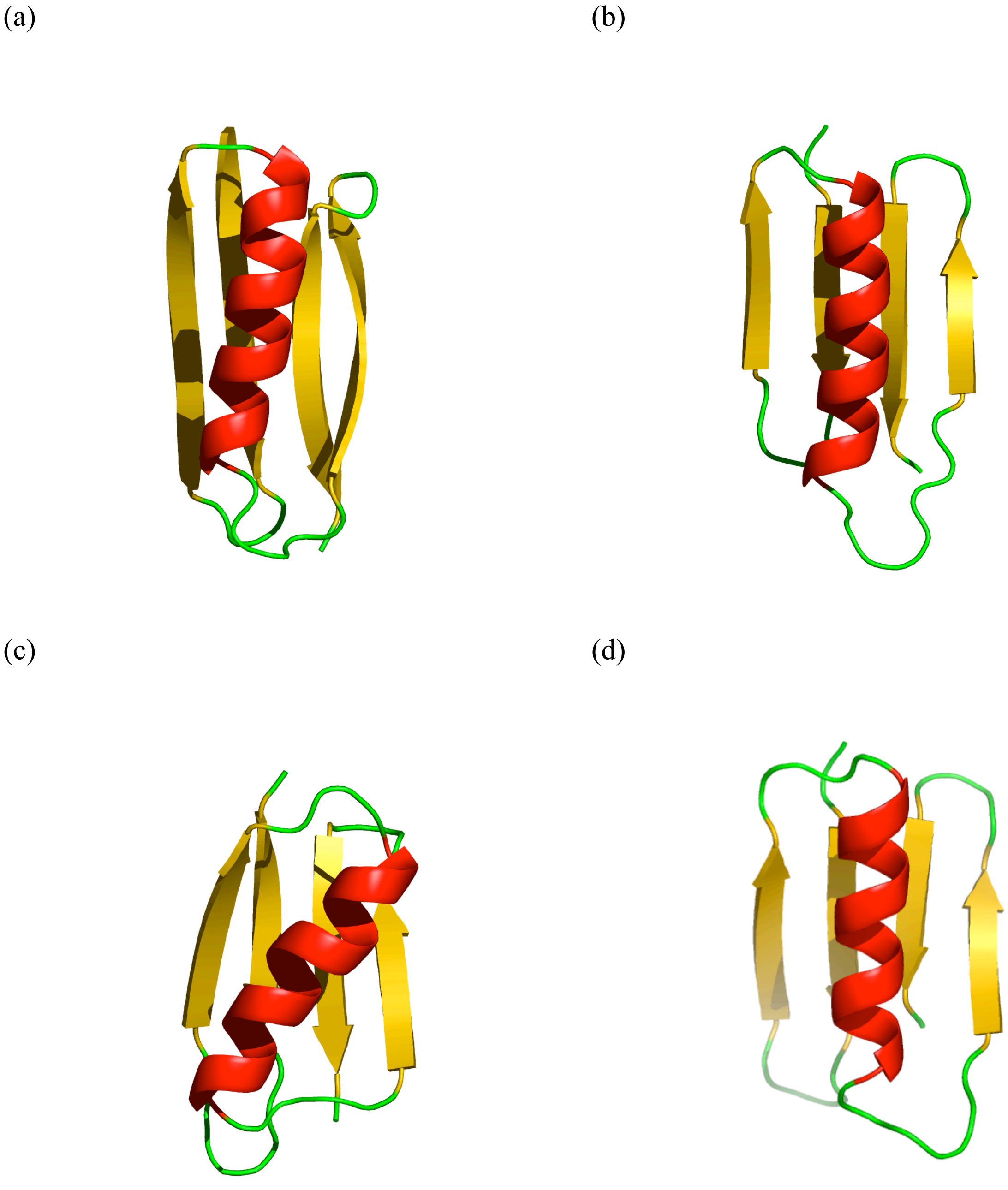

Figure 2.2. ***Simulated Annealing Results for Protein L and G.*** (a) PDB structure of Protein L (2PTL) with N-terminus loop region (residue 1-17) omitted. (b) Lowest energy structure from simulated annealing of 60-residue optimized sequence of Protein L. RMSD between2PTL and our model protein L is 4.4Å (c) PDB structure of Protein G (2GB1). (d) Lowest energy structure from simulated annealing of 56-residue optimized sequence of Protein G. RMSD between 2GB1 and our model protein G is 3.5Å.

Thermodynamics

Figure 2.3 plots the thermodynamic averages of percentage folded P_{Nat} (Figure 2.3a), heat capacity C_v (Figure 2.3b), and radius of gyration R_g (Figure 2.3c) against temperature for Protein L and G. Compared to results from our old model without the hydrogen bond[57], the new model demonstrates improved folding cooperativity. The folding temperature T_f, defined as the temperature at which $P_{Nat} = 0.5$, is 0.36 for protein L and 0.325 for protein G. The thermal stability plots show sharp transitions about T_f, a sign of greater folding cooperativity. The heat capacity and radius of gyration plots likewise show distinct transitions. The collapse temperatures are T_θ=0.36 for protein L and T_θ=0.335 for protein G, indicating that folding (T_f) is almost concomitant with collapse (T_θ).

The thermal stability P_{Nat} plot suggests that Protein L is more stable than protein G at any given temperature. This disagrees with experimental findings that protein G is marginally more stable than protein L under various denaturant conditions.[53, 58] It has been suggested that protein L's instability arises in part from torsional strain in the second hairpin.[84] Since we have adopted identical dihedral propensities for hairpins in our model L and G to focus on sequence effects, our models do not take into account this torsional destabilization. This could explain why our model protein L appears more stable than protein G. The heat capacity peak for protein L has a larger magnitude than that of protein G, which could be explained by protein L forming more hydrophobic contacts and hydrogen bonds in its native state than protein G.

To examine the free energy landscape, we project the potential mean force W along various order parameters. Figures 2.4a and 2.4b show the projections along Q for protein L and G at different temperatures. At their respective folding temperatures, proteins L and G each have two minima (denatured and native), suggesting a two-state folding mechanism. Figure 2.4c and 4d show the two-dimensional (2-D) projections along $\chi_{\beta 1}$ and $\chi_{\beta 2}$ for L and G at their folding temperatures. For Protein L, the minimum-energy path proceeds through a transition state in which hairpin 1 is partially formed while hairpin 2 is structureless, before reaching the native state. Protein G, on the other hand, has a minimum energy path that involves formation of a native-like hairpin 2, before crossing the transition state to reach the native state. The 2-D projections are in agreement with experimental evidence that the denatured state ensemble (DSE) and transition state ensemble (TSE) of protein L consist of partially formed β−hairpin 1[58, 59], while those of protein G involve a partially buried β−hairpin 2[53, 60]. However, P-fold analysis is needed to determine whether transition state ensembles obtained from the free energy projections are meaningful with respect to folding mechanism.

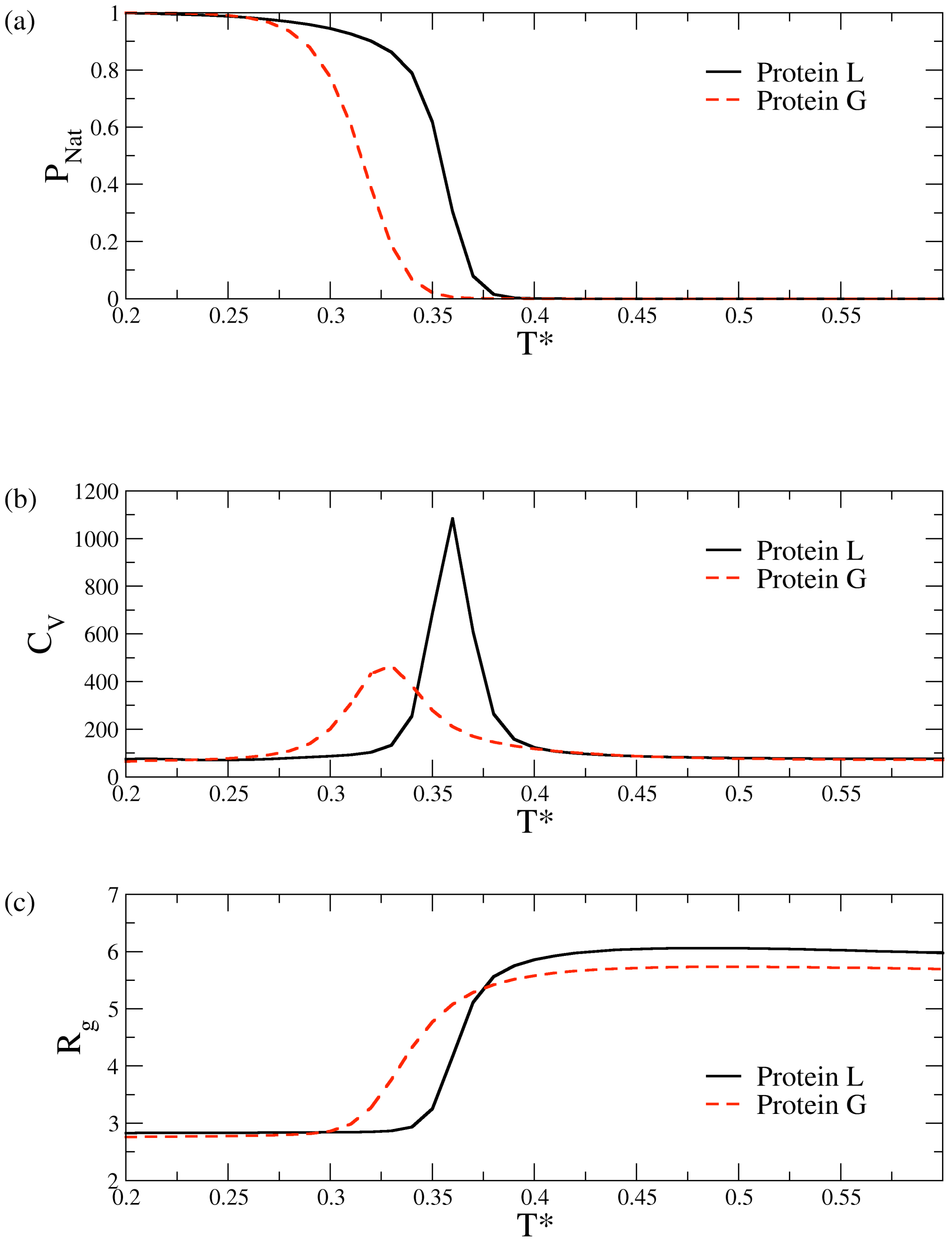

Figure 2.3. ***Thermodynamics averages for proteins L and G as functions of temperature.*** (a) percentage folded P_{Nat}, (b) heat capacity C_v, and (c) radius of gyration R_g.

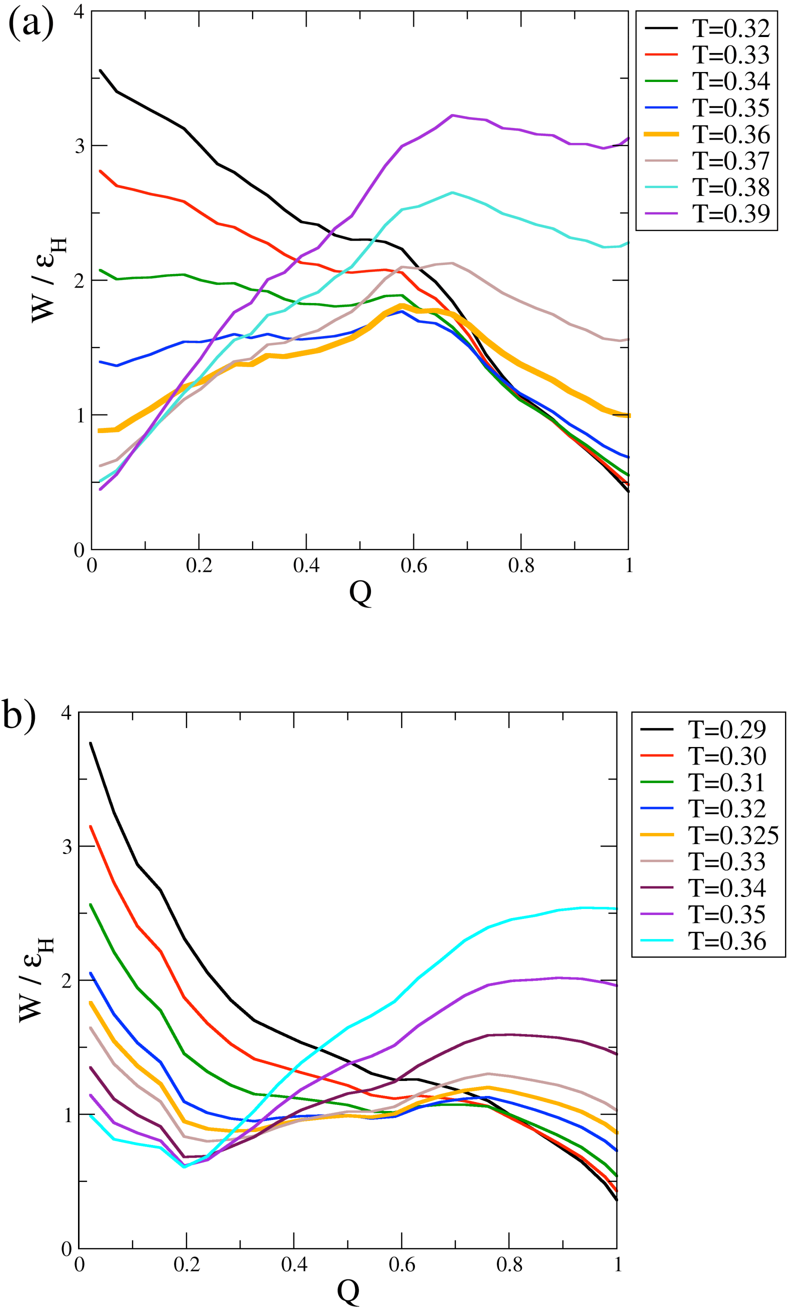

Figure 2.4. ***Free energy surface projections onto different reaction coordinates.*** (a) Projection of protein L's free energy along reaction coordinate Q over temperature range of 0.32<T<0.39. (b) Projection of protein G's free energy along reaction coordinate Q over temperature range of 0.29<T<0.36.

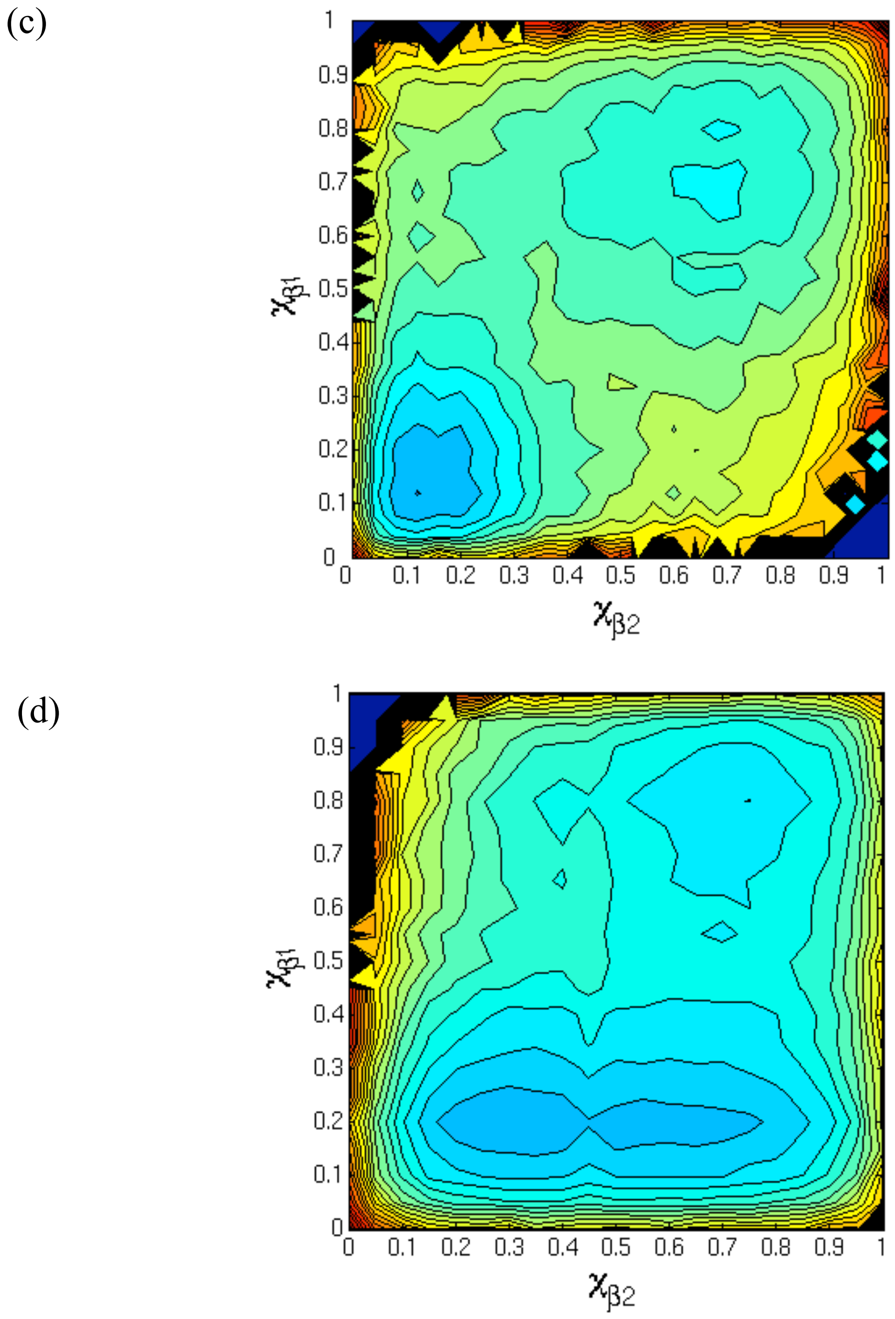

Figure 2.4. ***Free energy surface projections onto different reaction coordinates (continued)***
(c) Projection of protein L’s free energy surface onto $\chi_{\beta 1}$ and $\chi_{\beta 2}$ at T_f=0.36.
(d) Projection of protein G’s free energy surface onto $\chi_{\beta 1}$ and $\chi_{\beta 2}$ at T_f=0.325. Contours for (c) and (d) are spaced 0.5kT apart.

Transition States Analysis

The 2D free energy projections along $\chi_{\beta 1}$ and $\chi_{\beta 2}$ (Figure 2.4c and 2.4d) suggest different minimum free energy paths for the folding of L and G. From these projections, highest energy state for protein L appears to have a partially formed β−hairpin 1, while that of protein G has a partially formed β−hairpin 2. Noting that true transition states should be the highest point along the minimum free energy path, and correspond to saddle points on the multi-dimensional landscape, the relevant transition state ensemble (TSE) may be of higher dimension than suggested by simpler reaction coordinates $\chi_{\beta 1}$ or $\chi_{\beta 2}$. In fact these simpler reaction coordinates proved not to be saddle points on the multi-dimensional energy landscape according to P_{fold}, and therefore we needed to collect putative transition states for more complicated reaction coordinates. We found that the collective Q coordinate combined with $\chi_{\beta 1}$ and $\chi_{\beta 2}$ for proteins L and G respectively were sufficient to determine the TSE. According to the Q-$\chi_{\beta 1}$ projection for protein L, the putative TSE structures are collected for structures with $0.4 < Q < 0.6$ and $0.5 < \chi_{\beta 1} < 0.7$ (Figure 2.5a). According to the Q-$\chi_{\beta 2}$ projection for protein G, putative TSE structures are collected for structures with $0.6 < Q < 0.8$ and $0.35 < \chi_{\beta 2} < 0.8$ (Figure 2.5b). P_{fold} analysis was performed (see methods) and we identified the true transition state ensembles for proteins L and G (Figure 2.5c and 2.5d, respectively). Comparing the transition state contacts (red contours) for protein L and G, it is evident that the TSE of protein L consists of more native-like contacts in hairpin 1, while the TSE of protein G has more native-like contacts in hairpin 2. This is consistent with experimental studies using ϕ-value analysis[86, 93]. Both TSE contours indicate well-formed helices for L and G, while mutagenesis studies have suggested helices are relatively disrupted in TSEs. The contact maps also show some contacts between strand 1 and 4, which are consistent with experiments.

To explore how our simulated TSE correlates with mutagenesis experiments at a residue level, we perform single mutations on the optimized sequence of protein L and monitor how its transition state is perturbed by each mutation. From the mutations done by Kim et. al[84], we performed sixteen single mutations which can be represented by our four-flavor code. Table 2.4 lists the actual experimental mutations, and implementation in our model. Note that the residue indices are different for the experiment and simulation.

We then compute $1 - N_{TSE(MUT,i)}/N_{TSE}$ to quantify how much the transition state is perturbed. N_{TSE} refers to the number of conformations in the transition state of the optimized sequence. $N_{TSE(MUT.i)}$ refers to the number of conformations collected when we performed P_{fold} analysis on conformations in TSE with the mutation *i*. To compare against ϕ-values, we define a parameter R_i

$$R_i = \begin{cases} 0.2; & 0 \leq 1 - N_{TSE(MUT,i)}/N_{TSE} \leq 0.33 \\ 0.4; & 0.33 < 1 - N_{TSE(MUT,i)}/N_{TSE} \leq 0.5 \\ 0.8; & 0.5 < 1 - N_{TSE(MUT,i)}/N_{TSE} \leq 1.0 \end{cases}$$

Figure 2.6 shows the correlation between the experimental ϕ-values and R_i. While there are some outliers (namely N11S, N26S and N41S), the general trend is consistent with the experimental findings that residues in hairpin 1 are more important in the transition state then those in hairpin 2.

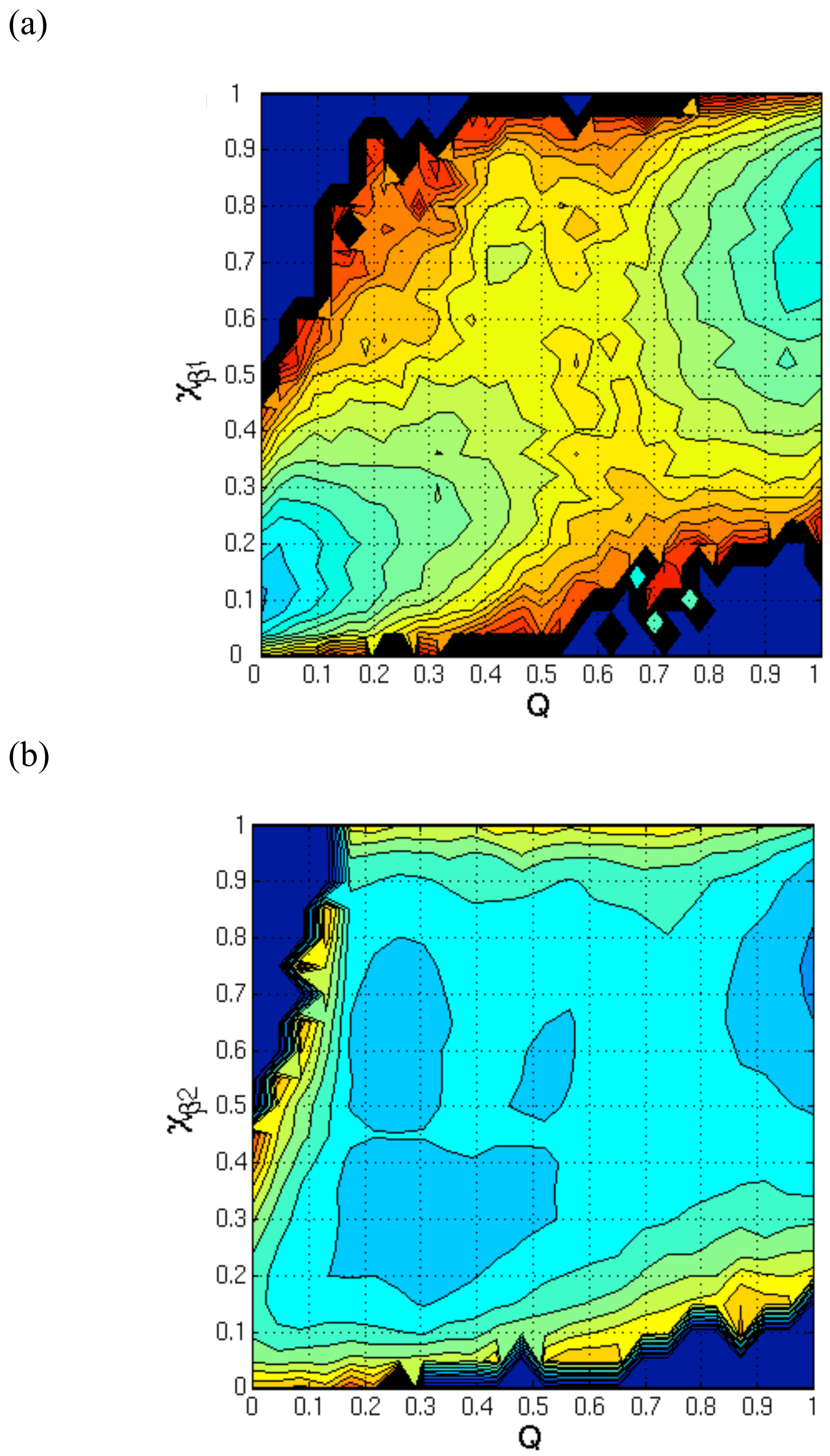

Figure 2.5. ***P_{fold} analysis of proteins L and G.*** Putative transition state ensembles are identified from free energy projections along (a) Q-$\chi_{\beta 1}$ and (b) Q-$\chi_{\beta 2}$ for proteins L and G, respectively.

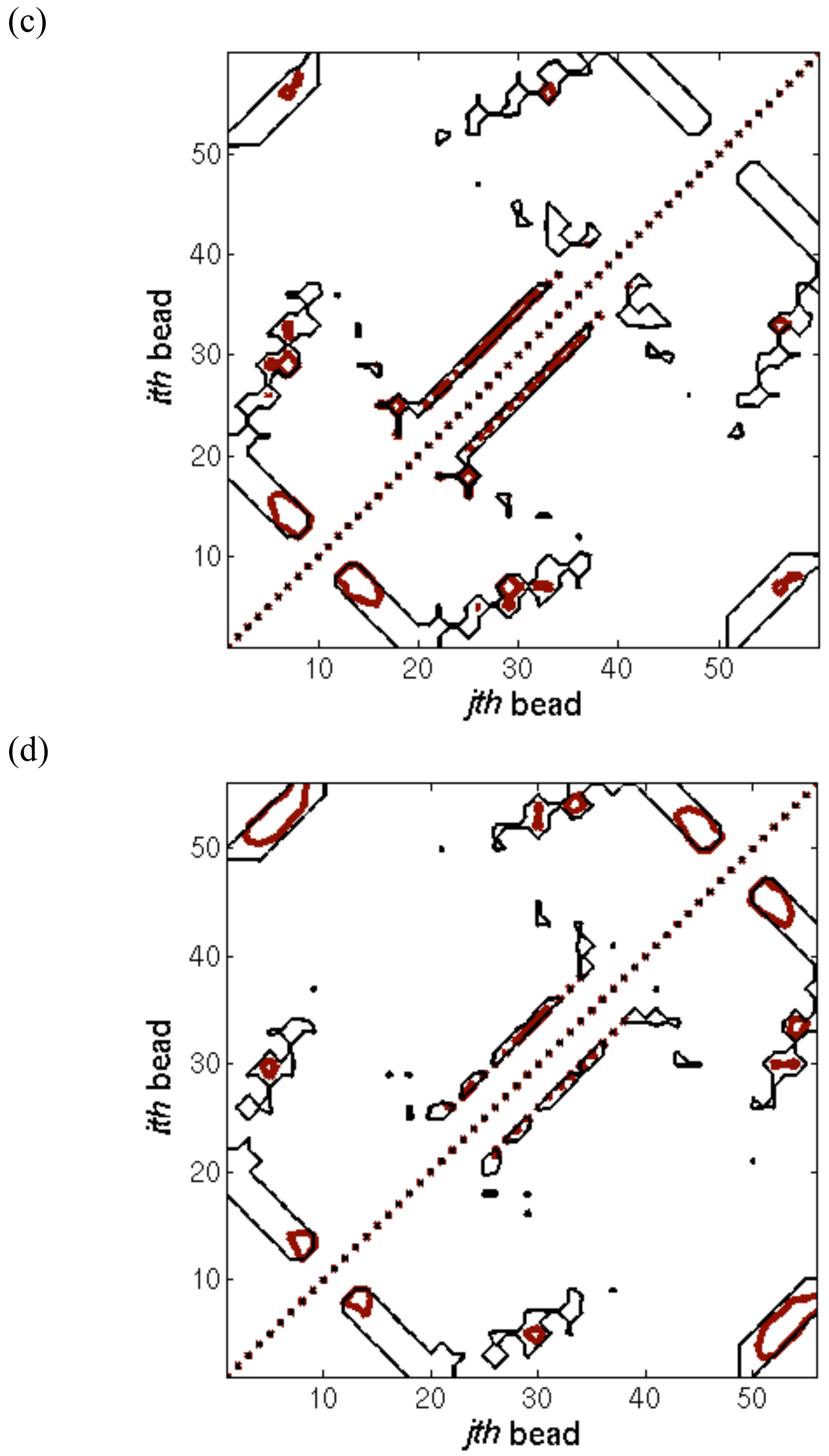

Figure 2.5. ***P_{fold}* analysis of proteins L and G (continued)** Contact maps of transition state ensembles from P_{fold} for (c) Protein L and (d) Protein G. Black contours denote native contacts. Red contours denote contacts made by 90% of structures in the transition state ensembles.

Table 2.4. Mutations performed on Protein L

Experimental Mutation[84]	Model Mutation	Experimental ϕ-values[84]	$1 - N_{TSE(MUT,i)} / N_{TSE}$	R
V4A	L4S	0.7	0.61	0.80
A8G	S5N, E5T*	0.43	0.39	0.4
G15A	N11S	0.86	0.24	0.20
T17A	N13S	0.42	0.36	0.40
T19A	N15S	0.17	0.27	0.20
E21A	L17S	1.08	0.61	0.80
K23A	L19S	0.57	0.39	0.40
G24A	N20S	0.2	0.33	0.20
T30A	N26S	0.14	0.88	0.80
N44A	L40S	0.08	0.27	0.20
G45A	N41S	-0.1	0.39	0.40
T48A	N44S	0.44	0.30	0.20
G55A	N51S	0.18	0.33	0.20
T57A	N53S	0.07	0.42	0.40
N59A	L55S	0.12	0.39	0.40
K61A	L57S	0.18	0.33	0.20

* Mutation in dihedral sequence

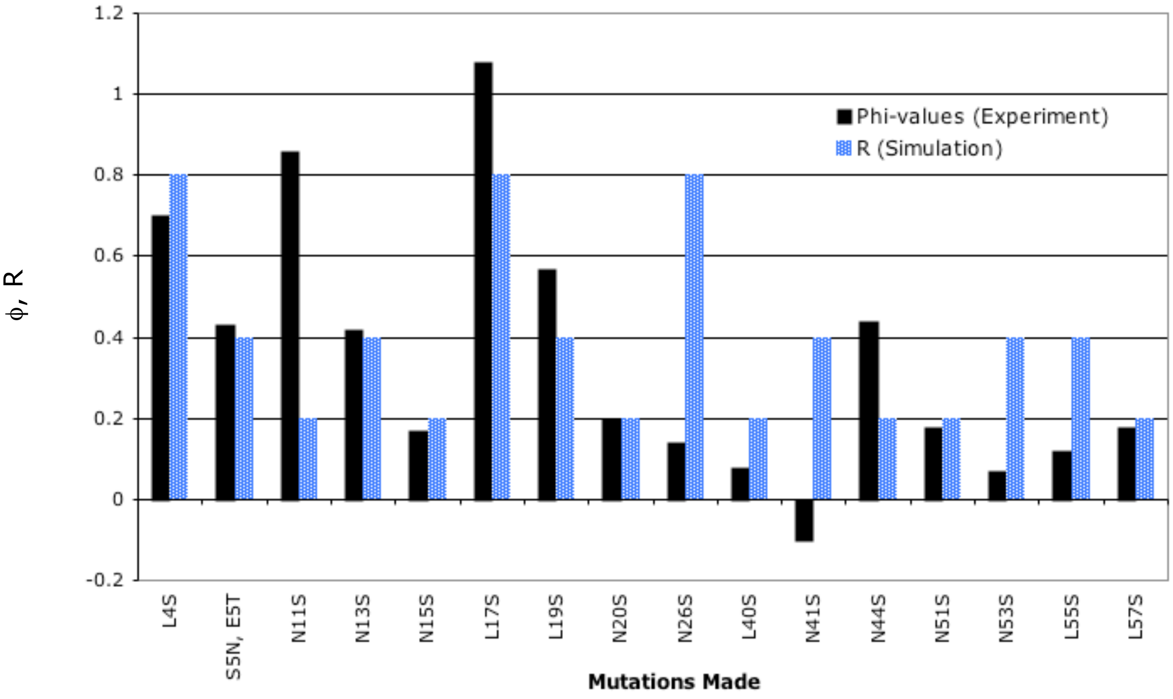

Figure 2.6. ***Correlation between experimental ϕ-values and perturbation to transistion state R for protein L.*** There is general agreement between experiment and R, with some outliers (N11S, N26S, N41S). Both experiment and our simulation indicate residues in Hairpin 1 are more important for the transition state than those of hairpin 2.

Kinetics

To rule out the possibility of glassiness, we evaluate the glass transition temperature, T_g, for our model. Wolynes and co-workers[59] have shown that a foldable, minimally-frustrated heteropolymer has a folding temperature well above its glass transition, so that a ratio of T_f to T_g should be greater than one. A working definition of the kinetic glass temperature T_g is the temperature at which average folding time $<\tau_f>$ is midway between τ_{min}, the fastest (minimum) folding time achievable, and τ_{max} the simulation cutoff time chosen to greatly exceed the observable folding times[6] (set to 100,000τ in this work). In Figure 2.7 we show that this occurs at $T_g = 0.14$, so that $T_f/T_g \sim 2.2$ for our model of Protein G, indicating that the energy landscapes is sufficiently smooth down to fairly low temperatures.

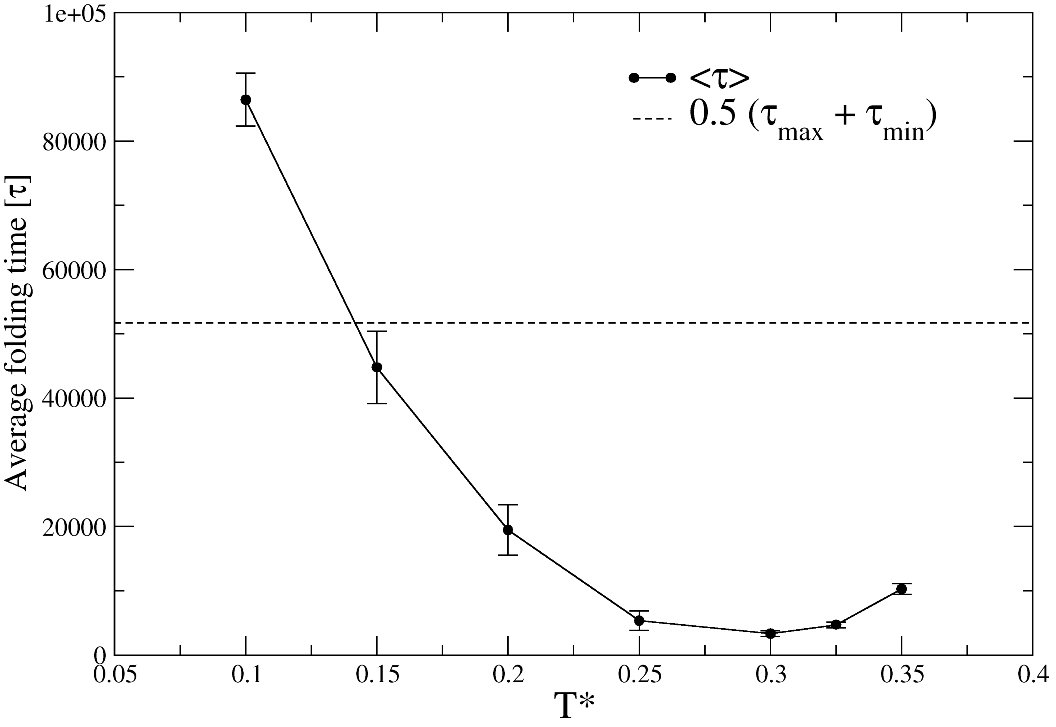

Figure 2.7. ***Determining the kinetic glass temperature T_g of protein G.***
The temperature at which average folding time $<\tau_f>$ is midway between τ_{min}, the fastest (minimum) folding time achievable, and τ_{max} the simulation cutoff time chosen to greatly exceed the observable folding times[6] (set to 100,000τ in this work). We determine that $T_g = 0.14$, so that $T_f/T_g \sim 2.2$ for our model of Protein G, indicating that the energy landscapes is sufficiently smooth down to fairly low temperatures.

Table 2.5. Kinetic Fit Parameters

		Sequential fit [100] (Gaussian relaxation followed by single exponential) *			
	Conditions	μ	σ	τ_0	χ^2
1.	L, T*=0.36	712	305	11,895	0.0408
2.	G, T*=0.325	741	450	3,963	0.0935

		Single exponential fit with deadtime †		
	Conditions	τ_D	τ_0	χ^2
3.	L, T*=0.36	694	11,928	0.0506
4.	G, T*=0.325	641	4,142	0.3436

The protein L and G models were next analyzed for the kinetic rates and mechanism of folding at their folding temperatures T_f=0.36 and T_f=0.325 respectively. During folding simulations, there is a finite equilibration time during which trajectories equilibrate from the initial free energy surfaces at T=1.6 to those at their target temperatures. The conventional treatment is to include a fitting parameter for dead time τ_D when fitting $P_{Nat}(t)$

$$P_{Nat}(t) = 1 - \sum_i A_i \exp\left[-\left(t - \tau_D\right) / \tau_i\right] \tag{2.6a}$$

where A_i is the population for average timescale process τ_i. The parameters used to fit the kinetic data for proteins L and G using Equation 2.6a are listed in Table 2.5.

We have shown in previous work[100] that, instead of using a constant deadtime, the initial equilibration to the new folding conditions could be better modelled as a relaxation process with Gaussian distributed probability. The overall kinetic data could hence be modelled as a sequential process with (a) initial Gaussian relaxation followed by (b) subsequent (multi)exponential kinetics

$$P_{Nat}(t) = \int_{u=0}^{t} \int_{s=0}^{t-u} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(u-\mu)^2}{2\sigma^2}\right)} \cdot \alpha e^{-\alpha s} ds du \tag{2.6b}$$

Integration of Equation 2.6b leads to

$$P_{Nat}(t) = \frac{1}{2}\left[1 + erf\left(\frac{t-\mu}{\sigma\sqrt{2}}\right)\right] - \sum_i \frac{A_i}{2}\left[1 + erf\left(\frac{t-B_i}{\sigma\sqrt{2}}\right)\right] e^{-\frac{t}{\tau_i}} e^{-\frac{D_i}{2\sigma^2}} \tag{2.6c}$$

where $B_i = (\mu+\alpha_i\sigma^2)$ and $D_i = \mu^2-(\mu+\alpha_i\sigma^2)^2$, t is the time over which the relaxation process happens, with mean μ and variance σ, and α_i is the kinetic folding rate for average timescale

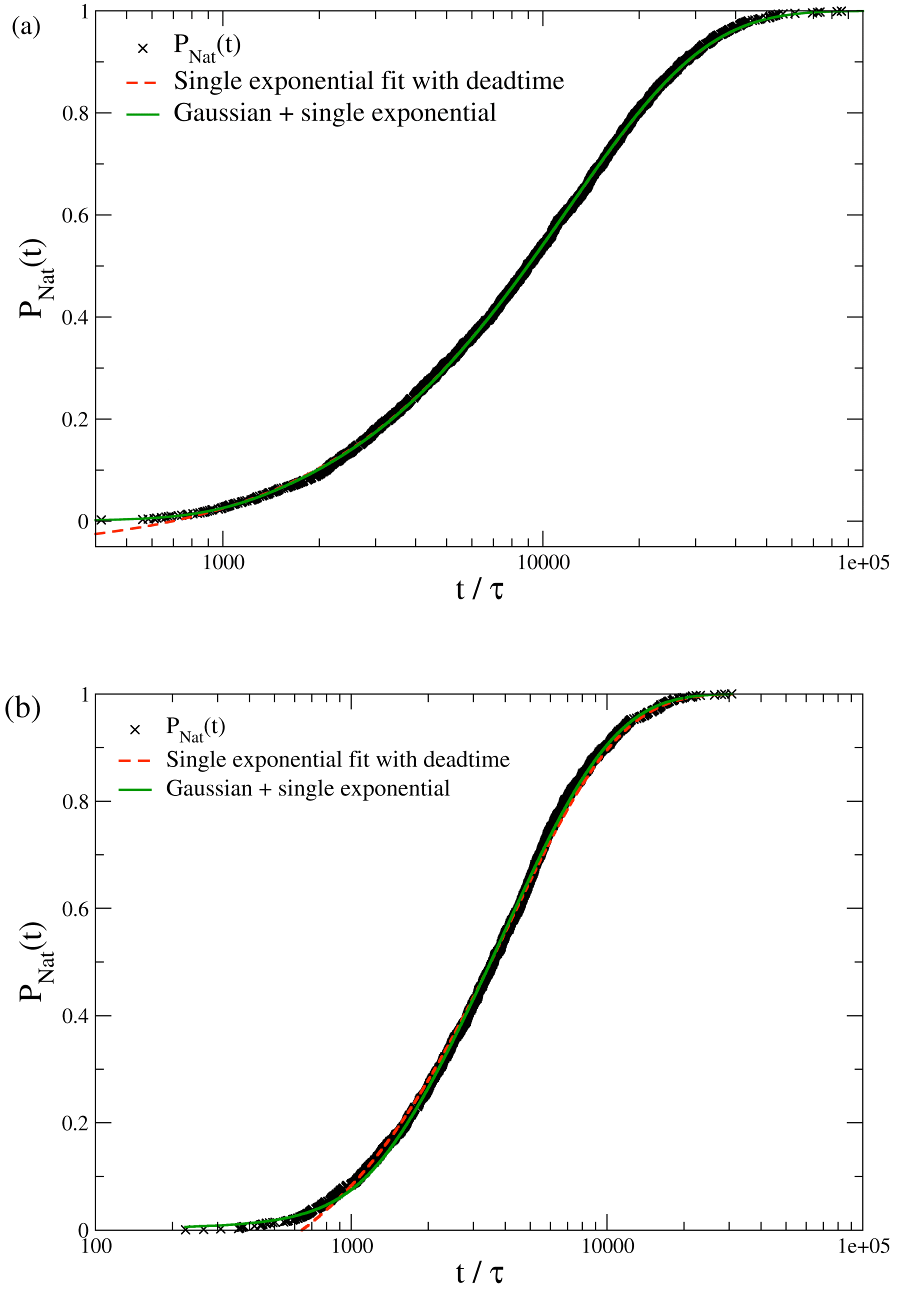

Figure 2.8. ***Kinetics data with fits for L and G at their respective folding temperatures using mean first passage time (MFPT) data.*** (a) Percentage of trajectories folded (P_{Nat}) as a function of time for protein L at $T_f = 0.36$. (b) Percentage of trajectories folded (P_{Nat}) as a function of time for protein G at $T_f = 0.325$. Both set of data are fitted to both a sequential and dead time model (see text). Fit parameters are listed in Table 2.5. The sequential process is seen to give a better fit to the kinetic data.

process τ_i. The fitting parameters using the sequential fit are also listed in Table 2.5. Comparing the fit quality, it is evident that the sequential mechanism provides a better fit than the dead time treatment, and Figure 2.8 shows the quality of fit for $P_{Nat}(t)$ for Protein L and G at their respective folding temperatures. Beyond the equilibration phase, the $P_{Nat}(t)$ data of Protein L (Figure 2.8a) fits to a single exponential, in agreement with experimental data[91]. The $P_{Nat}(t)$ data of Protein G at $T_f = 0.325$ also fits a single exponential (Figure 2.8b), agreeing with single exponential kinetics reported for protein G at its denaturant midpoint[88]. The folding time constants for L and G are 11,895τ and 3,963τ respectively. This is in qualitative agreement with experimental data[53, 58] that protein G folds faster than L.

Conclusions

We have presented an improved coarse-grained model capable of modeling directional hydrogen bonding. The model retains a strong connection between sequence and folding mechanism for proteins L and G, and shows increased folding cooperativity. The model native states also exhibit a greater structural faithfulness to experimentally solved structures. The addition of a fourth bead flavor (V) also provides an improvement over the old model by providing a more graded spectrum of attractive interaction energies (Figure 2.1). Overall the improvements to the original model, without introducing greater computational cost, translate to a smoother energy landscape and improved T_f/T_g ratios. The thermodynamic data presented demonstrate that our model assembles more cooperatively and preserves the sequence information that result in different free energy pathways for proteins L and G. This finding is further reinforced by kinetic P_{fold} analysis of their respective TSEs, which show good agreement with experimental mechanisms of protein L and G folding. The kinetics performed at their melting point ($T = T_f$) showed that both L and G fold via two-state mechanisms, consistent with experimental consensus under these midpoint denaturant conditions[88, 91].

We believe the model shows promise in application to other protein folding studies. One interesting outcome of the new model is our observation of kinetic complexity and burst phase kinetics under more strongly folding conditions for protein G that we hope to report in a future paper. The computational efficiency of the model has also permitted us to develop molecular models of the Alzheimer's $A\beta_{1\text{-}40}$ fibril in order to determine the critical nucleus, stability with chain size, and fibril elongation[1], opening opportunities for other protein-protein co-assembly processes.

Acknowledgements

THG gratefully acknowledges a Schlumberger Fellowship while on sabbatical at Cambridge University. Molecular graphics for this paper were created in PyMOL[101]. This work was supported in part by a grant from the NIH. NLF thanks the Whitaker Foundation for its graduate research fellowship.

Chapter 3

A New and Efficient Poisson-Boltzmann Solver for Interaction of Multiple Proteins

Introduction

The formation of protein complexes is ubiquitous in a crowded, salty cellular environment. Since electrostatic forces dominate the earliest of protein-protein recognition events in the cell, various analytical and numerical continuum theories of bulk electrolytes have been adapted for use to describe protein complexation mechanisms on the supramolecular scale.[102] One popular continuum mean-field theory is the Poisson-Boltzmann (PB) treatment, which forms the basis of Gouy-Chapman theory[103, 104] in electrochemistry, and under the low field linearized PB (LPB) approximation, the Debye-Hückel theory in solution chemistry[105] and Derjaguin-Landau-Verwey-Overbeek (DLVO) theory in colloid chemistry[106, 107]. Numerous techniques for solving the PB equation exist[108], including both analytical or numerical methods, and each has its drawbacks and its strengths.

Analytical methods typically allow rapid solution of the PB equation using multipole expansions under specialized geometries such as spheres or cylinders. A complete PB solution comprising one spherical macromolecule was developed by Kirkwood[109] more than 70 years ago, but generalization of this complete solution to two or more spherical macromolecules proved to be more difficult, and many different partial and approximate solutions have been proposed[110-113]. We have recently achieved a fundamental result in deriving an analytical PB solution for computing the screened electrostatic interaction between *arbitrary* numbers of spherical proteins of *arbitrarily* complex charge distributions, separated by *arbitrary* distance[3]. While such idealized protein geometries will typically be inappropriate for describing complexation on a supermolecular scale, this new analytical solution is a novel component of our new numerical PB solver for arbitrary protein shape. It also serves as a benchmark for the accuracy of the numerical solutions in certain idealized test cases.

By contrast, numerical methods (see reference [108] for a recent survey) such as finite-difference (FD) [43, 44, 114] and finite-element (FE)[115-117] methods can handle arbitrary dielectric boundaries by solving for the PB potential on a 3-D grid or mesh. However there are limitations of the FE or FD formulations, such as singularities in the potential solution due to point charges, that electric displacement continuity could not be enforced across dielectric boundaries (thereby reducing the solution accuracy and convergence rate), and forces must be estimated from finite-difference calculations[108]. But most importantly, the requirement that the solution be solved on a grid limits its practical application to spatial domains of either two to three typical macromolecules at reasonably high resolution (~0.2Å), or to larger numbers of macromolecules with greatly diminished resolution and thus solution accuracy. For example, the PBE solution for an assembled 50S ribosomal subunit has been evaluated at 0.45Å resolution[44], at the limit of machine memory, but to describe the preceding assembly process that occur over much larger spatial distances, the spatial resolution and consequently the solution accuracy would greatly deteriorate. As such, computational and memory cost in FD and FE methods are strictly functions of the number of grid points, and not the number of macromolecules described.

Boundary element (BE) methods[45-47, 50, 51] are an attractive alternative since they satisfy both the Dirichlet and von Neuman boundary conditions by construction, singular charges can be correctly treated, and most importantly the 2D solutions on the macromolecular surface removes spatial resolution limitations imposed by the 3D grid of the FD or FE solvers. However increasing the number of boundary surface element results in an increasingly large dense matrix to be solved with severe memory requirements, a problem which scales with the number of macromolecules. Acceleration of the BE approach[46, 49] incorporating fast multipole methods have rendered BE computational times comparable to state-of-the-art software packages like the Adaptive Poisson Boltzmann Solver (APBS)[44] based on FD solutions.

In this work we derive a new numerical approach to solving the PB equation by combining the advantages of both the boundary element and our analytical model[3] formalism. In particular, we replace the discretization of the molecule surface into a large number (tens of thousands) of boundary elements, by a discretization involving a smaller number (tens to hundreds) of spheres. The surface charges can then be iteratively solved using analytical multipole methods[3]. We show that our Poisson Boltzmann semi-analytical method, PB-SAM, converges to the analytical solution with better accuracy and at greatly reduced cost relative to the readily available public domain PB solver APBS.[44] Furthermore, we define a high quality benchmark using 140 poles to describe the electrostatic potential for two overlapping spheres that are models for the sharp features that are sometimes present in real protein geometries, in which we show that our PB-SAM solution converges to the correct solution with the same computational cost or better than the finite difference solution. Finally we illustrate the strength of the PB-SAM approach by computing the potential profile of an array of 60 T1-particle forming monomers of the bromine mosaic virus (PDB code 1YC6[4]).

Theory

Mathematical Preliminaries

Our theory makes extensive use of the spherical harmonics (SH) family of functions. The spherical harmonic function of order n and degree m, at polar angle θ and azimuthal angle ϕ, is defined per the convention from Gumerov and Duraiswami[118] as

$$Y_{nm}(\theta,\phi)=(-1)^m\sqrt{\frac{(n-|m|)!}{(n+|m|)!}}P_{n|m|}(\cos\theta)e^{im\phi} \tag{3.1}$$

where $P_{nm}(x)$ is the *associated Legendre polynomial.* Note that this definition of $Y_{nm}(\theta,\phi)$ differs from the common convention by a $\sqrt{(2n+1)/4\pi}$ factor. The complex conjugate of $Y_{nm}(\theta,\phi)$ will be denoted as $\overline{Y_{nm}}(\theta,\phi)$.

We shall utilize two important properties of spherical harmonics – their addition theorems and orthogonality. Let $\mathbf{r_1} = [r_1,\theta_1,\phi_1]$ and $\mathbf{r_2} = [r_2,\theta_2,\phi_2]$ be two points in 3D space specified by spherical coordinates, where $r_2 > r_1$. The Euclidean distance $|\mathbf{r_1}\text{-}\mathbf{r_2}|$ between them then obeys the addition theorems[46, 119]:

$$\frac{1}{|\mathbf{r}_1-\mathbf{r}_2|}=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{r_1^n}{r_2^{n+1}}\overline{Y_{nm}}(\theta_1,\phi_1)Y_{nm}(\theta_2,\phi_2) \tag{3.2a}$$

and for the screened Yukawa potential Eq. (3.2a) is modified to read as

$$\frac{e^{-\kappa|\mathbf{r}_1-\mathbf{r}_2|}}{|\mathbf{r}_1-\mathbf{r}_2|}=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{r_1^n}{r_2^{n+1}}\hat{i}_n(\kappa r_1)e^{-\kappa r_2}\hat{k}_n(\kappa r_2)\overline{Y_{nm}}(\theta_1,\phi_1)Y_{nm}(\theta_2,\phi_2) \tag{3.2b}$$

where κ is the inverse Debye Huckel screening length (described later), and $\hat{k}_n(z)$ and $\hat{i}_n(z)$ are *adapted modified spherical Bessel functions* defined as

$$\hat{k}_n(z)=\sqrt{\frac{2}{\pi}}\frac{e^z z^{n+1/2}}{(2n-1)!!}K_{n+1/2}(z) \tag{3.3a}$$

$$\hat{i}_n(z)=\sqrt{\frac{\pi}{2}}\frac{(2n+1)!!}{z^{n+1/2}}I_{n+1/2}(z) \tag{3.3b}$$

$I_n(z)$ and $K_n(z)$ are the *modified Bessel functions of the first and second kind* respectively. Detailed properties of $\hat{k}_n(z)$ and $\hat{i}_n(z)$ have been described in ref [3].

The spherical harmonic functions are also orthogonal over the surface of a unit sphere (S_1):

$$\int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi}Y_{ls}(\theta,\phi)\overline{Y_{nm}}(\theta,\phi)\sin\theta\partial\theta\partial\phi=\frac{4\pi}{2n+1}\delta_{nl}\delta_{ms} \tag{3.4a}$$

Hence a square-integrable function $g(\theta,\phi)$ on S_1 can be expanded using $\{Y_{nm}\}$ as the basis set:

$$g(\theta,\phi)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{2n+1}{4\pi}G_{nm}Y_{nm}(\theta,\phi) \tag{3.4b}$$

with the coefficients G_{nm} determined through the reciprocal transform

$$G_{nm}=\int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi}g(\theta',\phi')\overline{Y_{nm}}(\theta',\phi')\sin\theta'\partial\theta'\partial\phi' \tag{3.4c}$$

Setting up the boundary value problem

We seek to set up a boundary value problem for a system of N_{mol} macromolecules immersed in an implicit aqueous salty solvent. Figure 3.1 gives an example of the spatial domain for which we solve the linearized PB equation (LPBE). Each macromolecule I is embedded with $N_C^{(I)}$ fixed partial charge and represented as a collection of $N_S^{(I)}$ overlapping spheres with dielectric constant ε_{in}. For simplicity we consider in this paper the same ε_{in} for all molecules, but the model can handle different dielectric constants. The solvent is treated as a continuum with dielectric constant ε_{out}, with screening effects due to mobile ions captured via the inverse Debye length κ. The LPBE gives the potential Φ at any point **r** in space $\Re^3$ as

$$-\nabla\left[\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})\right]+\kappa^2\Phi(\mathbf{r})=4\pi\rho_{fixed}(\mathbf{r}) \tag{3.5}$$

where ε is the relative dielectric function, ρ_{fixed} is the charge density due to the fixed protein partial charges, and $\kappa=\sqrt{8\pi\bar{n}e^2/\varepsilon_{out}k_BT}$, where $\bar{n}$ is the bulk concentration of monovalent salt in the solution, e is the fundamental electronic charge, k_B the Boltzmann constant, and T the absolute temperature.

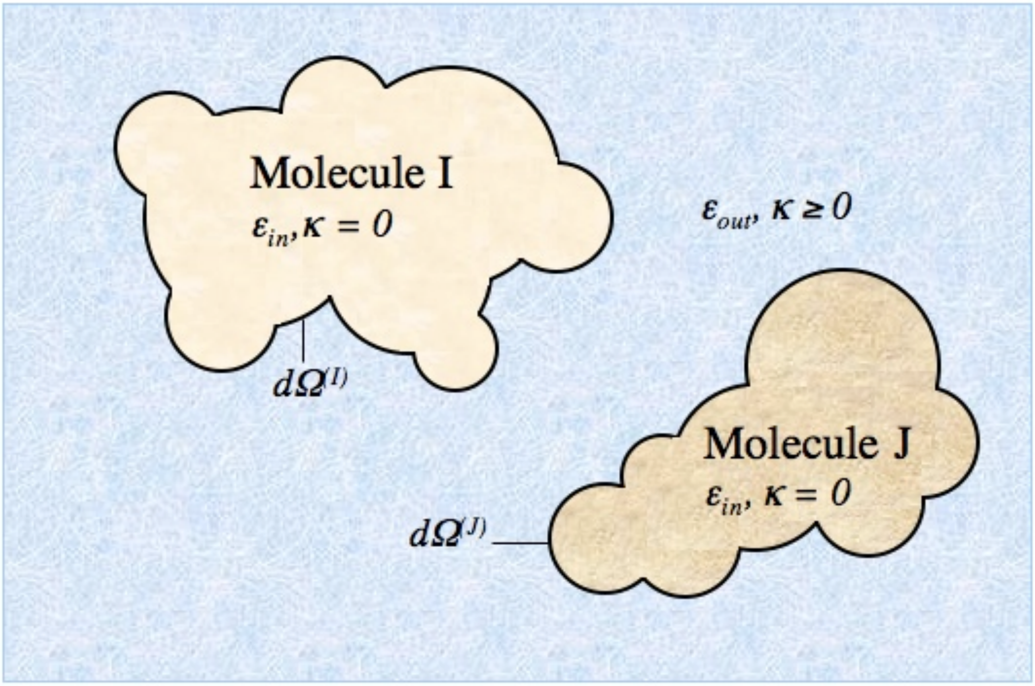
Molecule I
$\varepsilon_{in}, \kappa = 0$
$\varepsilon_{out}, \kappa \geq 0$
$d\Omega^{(I)}$
Molecule J
$\varepsilon_{in}, \kappa = 0$
$d\Omega^{(J)}$

Figure 3.1. ***Setting up the boundary value problem.*** The example system is comprised of two proteins with arbitrary charge distribution, each represented as a collection of overlapping spheres to describe an arbitrarily shaped dielectric boundary containing no salt, immersed in a high dielectric salty continuum solvent. Salt screening effects are captured via the Debye Huckel parameter κ.

Inside each macromolecule *I*, the potential $\Phi_{in}^{(I)}(\mathbf{r})$ satisfies the Poisson equation

$$-\nabla^2\Phi_{in}^{(I)}(\mathbf{r}) = \rho_{fixed}^{(I)}(\mathbf{r})/\varepsilon_{in} \tag{3.6a}$$

while in the region outside all macromolecules, the potential $\Phi_{out}(\mathbf{r})$ satisfies the Helmholtz equation

$$\nabla^2\Phi_{out}(r) - \kappa^2\Phi_{out}(r) = 0 \tag{3.6b}$$

We first express the potential $\Phi_{in}^{(I)}(\mathbf{r})$ anywhere inside molecule *I* as the sum of the potentials due to the embedded fixed charges and a single-layer of yet unknown reaction charges $f^{(I)}(\mathbf{r})$ on the surface $d\Omega^{(I)}$[45, 120]:

$$\Phi_{in}^{(I)}(\mathbf{r}) = \sum_{\alpha=1}^{N_C^{(I)}} \frac{1}{\left|\mathbf{r}-\mathbf{r}_\alpha^{(i)}\right|}\frac{q_\alpha^{(I)}}{\varepsilon_{in}} + \frac{1}{4\pi}\int_{d\Omega^{(I)}} \frac{1}{\left|\mathbf{r}-\mathbf{r}'\right|} f^{(I)}(\mathbf{r}')d\mathbf{r}' \tag{3.7}$$

In our new approach, the surface of molecule *I* is discretized into $N_S^{(I)}$ spheres. We consider each sphere *k* of molecule *I* of radius $a^{(I,k)}$ in turn, and all position vectors and coefficients are defined with the center of sphere *k* as the origin. We apply the first addition theorem (Eq. (3.2a)) to Eq. (3.7) to obtain

$$\begin{aligned}\Phi_{in}^{(I,k)}(\boldsymbol{r}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{E_{nm}^{(I,k)}}{r}\left(\frac{a^{(I,k)}}{r}\right)^n + \left(\frac{r}{a^{(I,k)}}\right)^n LE_{nm}^{(I,k)}\right)Y_{nm}^{(I,k)}(\theta,\phi) + \\ \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\left(\frac{r}{a^{(I,k)}}\right)^n \left(LF_{nm}^{(I,k)} + LFS_{nm}^{(I,k)}\right)\right)Y_{nm}^{(I,k)}(\theta,\phi)\end{aligned} \tag{3.8}$$

with the coefficients defined as

$$E_{nm}^{(I,k)} \equiv \sum_{\alpha=1}^{N_C^{(I,k)}} \frac{q_\alpha}{\varepsilon_{in}}\left(\frac{r_\alpha}{a^{(I,k)}}\right)^n \overline{Y_{nm}^{(I,k)}}(\theta_\alpha,\phi_\alpha) \tag{3.8a}$$

$$LE_{nm}^{(I,k)} \equiv \sum_{\alpha=1}^{\bar{N}_C^{(I,k)}} \frac{q_\alpha}{\varepsilon_{in}}\frac{1}{r_\alpha}\left(\frac{a^{(I,k)}}{r_\alpha}\right)^n \overline{Y_{nm}^{(I,k)}}(\theta_\alpha,\phi_\alpha) \tag{3.8b}$$

$$LF_{nm}^{(I,k)} \equiv \frac{1}{4\pi}\int_{\overline{d\Omega}^{(I,k)}} \frac{f^{(I,k)}(\mathbf{r}')}{r'}\left(\frac{a^{(I,k)}}{r'}\right)^n \overline{Y_{nm}^{(I,k)}}(\theta',\phi')d\mathbf{r}' \tag{3.8c}$$

$$LFS_{nm}^{(I,k)} \equiv \frac{1}{4\pi}\int_{d\Omega^{(I,k)}} \frac{f^{(I,k)}(\mathbf{r}')}{r'}\left(\frac{a^{(I,k)}}{r'}\right)^n \overline{Y_{nm}^{(I,k)}}(\theta',\phi')d\mathbf{r}' \tag{3.8d}$$

Notice that we have scaled the terms with r_α^n and r_α^{n+1} dependence by $\left(a^{(I,k)}\right)^n$ and $\left(a^{(I,k)}\right)^{-n}$ respectively. This is to avoid machine imprecision as *n* becomes large. Coefficients with $\left(r_\alpha / a^{(I,k)}\right)^n$ dependence, such as $E_{nm}^{(I,k)}$, are known as multipole (external) coefficients, while those with $a^{(I,k)n} / r_\alpha{}^{n+1}$ dependence ($LE_{nm}^{(I,k)}$, $LF_{nm}^{(I,k)}$ and $LFS_{nm}^{(I,k)}$) are known as Taylor (local) coefficients. The first sum in Eq. (3.8) represents the potential due to fixed charges, where $E_{nm}^{(i,k)}$ sums over $N_C^{(I,k)}$ fixed charges *inside* sphere *k* of molecule *I*, while $LE_{nm}^{(I,k)}$ sums over the remaining $\overline{N}_C^{(I,k)}$ fixed charges *outside* sphere *k*. The second sum in Eq. (3.8) gives the potential due to the unknown surface charge $f^{(I)}(\mathbf{r})$; $LFS_{nm}^{(I,k)}$ and $LF_{nm}^{(I,k)}$ account for represents reactive charges on sphere *k*, and on other spheres in molecule *I*, respectively.

In the solvent region outside the molecules, the potential $\Phi_{out}(\mathbf{r})$ can be represented as the sum of Yukawa potentials due to each molecule's yet unknown effective surface charges $h^{(I)}(\mathbf{r})$[45, 120]

$$\Phi_{out}(\mathbf{r}) = \sum_{I=1}^{N_{mol}}\left(\frac{1}{4\pi}\int_{d\Omega^{(I)}}\frac{e^{-\kappa|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}h^{(I)}(\mathbf{r}')d\mathbf{r}'\right) \tag{3.9}$$

The above equation valid for the *exposed* portion of sphere *k* of molecule *I*. Applying addition theorem 2 (Eq. (3.2b)) to Eq. (3.9), the potential on the exposed surface can be expressed as

$$\Phi_{out}^{(I,k)}(\mathbf{r}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{H_{nm}^{(I,k)}}{r}\left(\frac{a^{(I,k)}}{r}\right)^n e^{-\kappa r}\hat{k}_n(\kappa r) + \left(\frac{r}{a^{(I,k)}}\right)^n \hat{i}_n(\kappa r)\left(LH_{nm}^{(I,k)} + LHN_{nm}^{(I,k)}\right)\right)Y_{nm}^{(I,k)}(\theta,\phi) \tag{3.10}$$

where the coefficients are defined as

$$H_{nm}^{(I,k)} \equiv \frac{1}{4\pi}\int_{d\Omega^{(i,k)}} h^{(I,k)}(\mathbf{r}')\left(\frac{r'}{a^{(I,k)}}\right)^n \hat{i}_n(\kappa r')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')d\mathbf{r}' \tag{3.10a}$$

$$LH_{nm}^{(I,k)} \equiv \frac{1}{4\pi}\int_{d\Omega^{(I,k)}} \frac{h^{(I,k)}(\mathbf{r}')}{r'}\left(\frac{a^{(I,k)}}{r'}\right)^n e^{-\kappa r'}\hat{k}_n(\kappa r')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')d\mathbf{r}' \tag{3.10b}$$

$$LHN_{nm}^{(I,k)} \equiv \sum_{J\neq I}^{N_{mol}}\sum_{l=1}^{N_S^{(J)}}\left(\frac{1}{4\pi}\int_{d\Omega^{(J,l)}} \frac{h^{(J,l)}(\mathbf{r}')}{r'}\left(\frac{a^{(I,k)}}{r'}\right)^n e^{-\kappa r'}\hat{k}_n(\kappa r')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')d\mathbf{r}'\right) \tag{3.10c}$$

The multipole coefficient $H_{nm}^{(I,k)}$ represents effective polarization charges on sphere *k* of molecule *I*'s exposed surface. The local coefficients $LH_{nm}^{(I,k)}$ and $LHN_{nm}^{(I,k)}$ represent effective polarization charges on other spheres in molecule *I*, and on other molecules, respectively.

With equations (3.8) and (3.10) in hand, we can impose boundary conditions at the dielectric boundary surface $\mathbf{r}_E = \left(a^{(I,k)},\theta_E,\phi_E\right) \in d\Omega_E^{(I,k)}$ between each sphere *k* in molecule *I* exposed to solvent:

$$\Phi_{in}^{(I,k)}\left(\mathbf{r}_E\right) = \Phi_{out}^{(I,k)}\left(\mathbf{r}_E\right) \tag{3.11a}$$

$$\varepsilon \left.\frac{d\Phi_{in}^{(I,k)}}{dn}\right|_{\mathbf{r}_E} = \left.\frac{d\Phi_{out}^{(I,k)}}{dn}\right|_{\mathbf{r}_E}, \qquad \varepsilon = \varepsilon_{in}/\varepsilon_{out} \tag{3.11b}$$

The Dirichlet boundary condition (Eq. 3.11a) enforces potential continuity across the boundary

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(E_{nm}^{(I,k)} + a^{(I,k)}\left(LE_{nm}^{(I,k)} + LF_{nm}^{(I,k)} + LFS_{nm}^{(I,k)}\right)\right)Y_{nm}^{(I,k)}(\theta_E,\phi_E)\\ &\quad = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(H_{nm}^{(I,k)}e^{-\kappa a^{(I,k)}}\hat{k}_n(\kappa a^{(I,k)}) + a^{(I,k)}\hat{i}_n(\kappa a^{(I,k)})\left(LH_{nm}^{(I,k)} + LHN_{nm}^{(I,k)}\right)\right)Y_{nm}^{(I,k)}(\theta_E,\phi_E)\end{aligned} \tag{3.12a}$$

while the von Neumann boundary condition (Eq. 3.11b) enforces electric displacement continuity

$$\begin{aligned}&\varepsilon\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(-(n+1)E_{nm}^{(I,k)} + nF_{nm}^{(I,k)} + na^{(I,k)}\left(LE_{nm}^{(I,k)} + LF_{nm}^{(I,k)}\right)\right)Y_{nm}^{(I,k)}(\theta_E,\phi_E)\\ &\quad = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\begin{aligned}&H_{nm}^{(I,k)}e^{-\kappa a^{(I,k)}}\left[n\hat{k}_n(\kappa a^{(I,k)}) - (2n+1)\hat{k}_{n+1}(\kappa a^{(I,k)})\right] +\\ &a^{(I,k)}\left[n\hat{i}_n(\kappa a^{(I,k)}) + \frac{\left(\kappa a^{(I,k)}\right)^2\hat{i}_{n+1}(\kappa a^{(I,k)})}{2n+3}\right]\left(LH_{nm}^{(I,k)} + LHN_{nm}^{(I,k)}\right)\end{aligned}\right)Y_{nm}^{(I,k)}(\theta_E,\phi_E)\end{aligned} \tag{3.12b}$$

where we have introduced $F_{nm}^{(I,k)} \equiv a^{(I,k)}LFS_{nm}^{(I,k)}$. We continue to simplify Eqs. (3.12a) and (3.12b) by rearranging

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(-H_{nm}^{(I,k)}e^{-\kappa a^{(I,k)}}\hat{k}_n(\kappa a^{(I,k)}) + F_{nm}^{(I,k)} + XH_{nm}^{(I,k)}\right)Y_{nm}^{(I,k)}(\theta_E,\phi_E) = 0 \tag{3.13a}$$

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(e^{-\kappa a^{(I,k)}}\left[n\hat{k}_n(\kappa a^{(I,k)}) - (2n+1)\hat{k}_{n+1}(\kappa a^{(I,k)})\right]H_{nm}^{(I,k)} - n\varepsilon F_{nm}^{(I,k)} + XF_{nm}^{(I,k)}\right)Y_{nm}^{(I,k)}(\theta_E,\phi_E) = 0 \tag{3.13b}$$

where

$$XH_{nm}^{(I,k)} \equiv E_{nm}^{(I,k)} + a^{(I,k)}\left(LE_{nm}^{(I,k)} + LF_{nm}^{(I,k)}\right) - a^{(I,k)}\hat{i}_n(ka^{(I,k)})\left(LH_{nm}^{(I,k)} + LHN_{nm}^{(I,k)}\right) \tag{3.14a}$$

$$\begin{aligned}XF_{nm}^{(I,k)} \equiv\ & a^{(I,k)}\left[n\hat{i}_n(ka^{(I,k)}) + \frac{\left(\kappa a^{(I,k)}\right)^2\hat{i}_{n+1}(ka^{(I,k)})}{2n+3}\right]\left(LH_{nm}^{(I,k)} + LHN_{nm}^{(I,k)}\right) +\\ & (n+1)\varepsilon E_{nm}^{(I,k)} - n\varepsilon a^{(I,k)}\left(LE_{nm}^{(I,k)} + LF_{nm}^{(I,k)}\right)\end{aligned} \tag{3.14b}$$

The boundary equations above are valid on the solvent-exposed surfaces of sphere k on molecule I. We need another set of boundary equations on the buried surface $\mathbf{r}_B = \left[a^{(I,k)},\theta_B,\phi_B\right] \in d\Omega_B^{(I,k)}$. We shall utilize the fact that there is no polarization charge on the buried surface, i.e. $f^{(I,k)}(\mathbf{r}_B) = h^{(I,k)}(\mathbf{r}_B) = 0$, since there is no dielectric discontinuity. It follows that scaled versions of the charge distributions, $\tilde{f}^{(I,k)}(\theta,\phi) \equiv \left(a^{(I,k)}\right)^2 f^{(I,k)}\left(a^{(I,k)},\theta,\phi\right)$ and $\tilde{h}^{(I,k)}(\theta,\phi) \equiv \left(a^{(I,k)}\right)^2 h^{(I,k)}\left(a^{(I,k)},\theta,\phi\right)$, are also zero on the buried surface. Separately, we can express $\tilde{f}^{(I,k)}$ and $\tilde{h}^{(I,k)}$ in terms of $F_{nm}^{(I,k)}$ and $H_{nm}^{(I,k)}$ using Eqs. (3.4c), (3.8d) and (3.10a),

$$\tilde{f}^{(I,k)}(\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \frac{2n+1}{4\pi} F_{nm}^{(I,k)} Y_{nm}^{(I,k)}(\theta,\phi) \tag{3.15a}$$

$$\tilde{h}^{(I,k)}(\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \frac{2n+1}{4\pi} \frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} Y_{nm}^{(I,k)}(\theta,\phi) \tag{3.15b}$$

so the ‘zero-charge’ requirement at the buried boundary can be imposed as

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n} \frac{2n+1}{4\pi} F_{nm}^{(I,k)} Y_{nm}^{(I,k)}(\theta_B,\phi_B) = 0 \tag{3.16a}$$

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n} \frac{2n+1}{4\pi} \frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} Y_{nm}^{(I,k)}(\theta_B,\phi_B) = 0 \tag{3.16b}$$

Equations (3.13a), (3.13b), (3.16a) and (3.16b) specified the complete boundary value problem, from which $F_{nm}^{(I,k)}$ and $H_{nm}^{(I,k)}$ can be solved.

Solution of the boundary value coefficients and interaction energy

To solve for $F_{nm}^{(i,k)}$ and $H_{nm}^{(i,k)}$, we need to cast the boundary value problem as a linear system of equations. The infinite expansion series must first be truncated at a maximum pole order *p*, chosen depending on the desired level of accuracy versus computational cost (see Results). The obvious approach is to set up the boundary equations as a linear least square problem (Figure 3.2a), by discretizing sphere *k* into M_B buried and M_E exposed grid points, and then finding solutions of vectors $\boldsymbol{F}^{(I,k)}$ and $\boldsymbol{H}^{(I,k)}$ that best satisfy the appropriate boundary equations on all grid points. Using the DGELSY routine (complete orthogonal factorization) in LAPACK for $(M_E + M_B)$= 10,000 and $p = 60$, each sphere is solved in approximately 10 minutes. This is computationally intractable if the LPBE needs to be solved repeatedly for tens to hundreds of spheres during dynamics simulations.

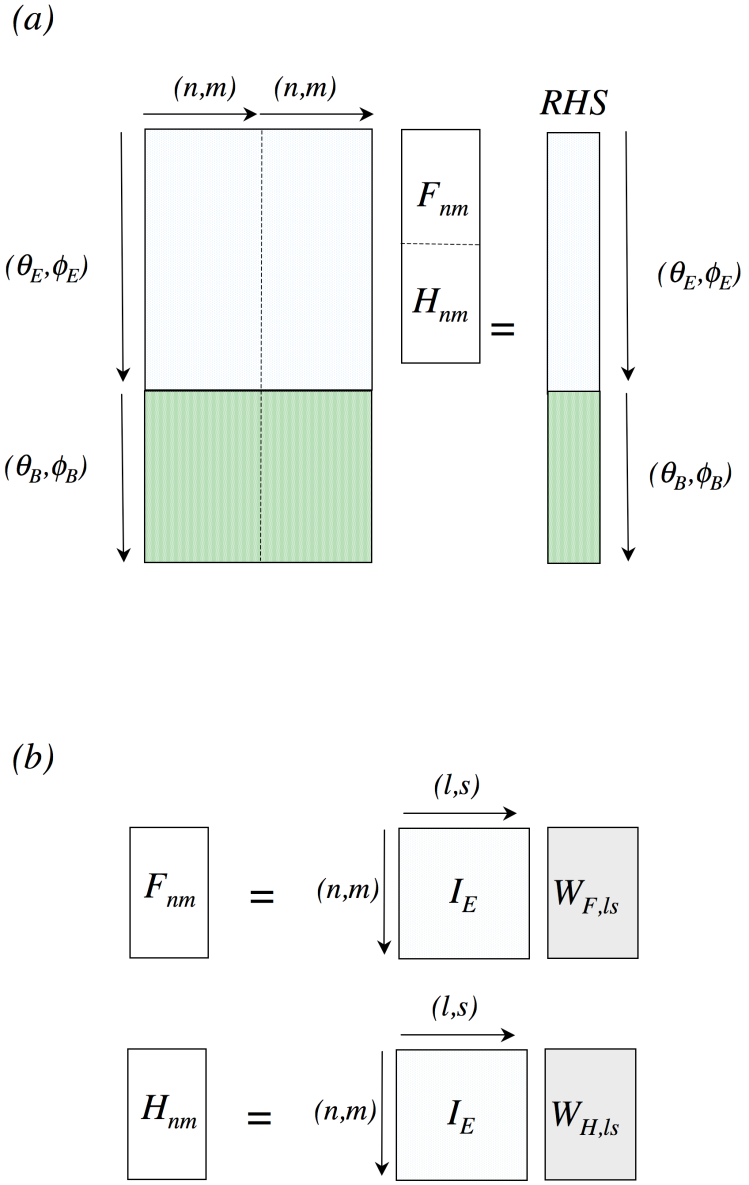

Figure 3.2. ***Setting up the boundary equation (Eqs. 3.13a-b, 3.16a-b).*** (a) As a Linear Least Square solve problem. (b) As a matrix-vector multiply operation.

Instead, we formulated a novel approach that makes use of spherical harmonics' orthogonal property (Eq. 3.4). It converts the problem to a direct matrix-vector multiply operation (Figure 3.2b), which can be evaluated two-orders of magnitude faster than the LLS approach. We first add $\sum_{n=0}^{\infty}\sum_{m=-n}^{n}(2n+1)\frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})}Y_{nm}^{(I,k)}(\theta_E,\phi_E)$ to both sides of Eq. (3.13a) and divide by 4π to arrive at:

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{2n+1}{4\pi}\frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a_{ki})}Y_{nm}^{(I,k)}(\theta_E,\phi_E)=\tilde{w}_{H,exposed}(\theta_E,\phi_E) \qquad (3.17a)$$

where

$$\tilde{w}_{H,exposed}(\theta,\phi)=\frac{1}{4\pi}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(H_{nm}^{(I,k)}\left(\frac{2n+1}{\hat{i}_n(\kappa a^{(I,k)})}-e^{-\kappa a^{(I,k)}}\hat{k}_n(\kappa a^{(I,k)})\right)+F_{nm}^{(I,k)}+XH_{nm}^{(I,k)}\right)Y_{nm}^{(I,k)}(\theta,\phi) \qquad (3.17b)$$

Similarly, we add $\sum_{n=0}^{\infty}\sum_{m=-n}^{n}(2n+1)F_{nm}^{(I,k)}Y_{nm}^{(I,k)}(\theta_E,\phi_E)$ to both sides of Eq. (3.13b) and then divide by 4π:

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{2n+1}{4\pi}F_{nm}^{(I,k)}Y_{nm}^{(I,k)}(\theta_E,\phi_E)=\tilde{w}_{F,exposed}(\theta_E,\phi_E) \qquad (3.18a)$$

$$\tilde{w}_{F,exposed}(\theta,\phi)=\frac{1}{4\pi}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\begin{pmatrix}e^{-ka^{(I,k)}}\left[n\hat{k}_n(ka^{(I,k)})-(2n+1)\hat{k}_{n+1}(ka^{(I,k)})\right]H_{nm}^{(I,k)}+\\(2n+1-n\varepsilon)F_{nm}^{(I,k)}+XF_{nm}^{(I,k)}\end{pmatrix}Y_{nm}^{(I,k)}(\theta,\phi) \qquad (3.18b)$$

Equations (3.17a) and (3.17b) (and similarly (3.18a) and (3.18b)) now completely describe functions $\tilde{w}_H(\theta,\phi)$ (and $\tilde{w}_F(\theta,\phi)$) over the entire surface of sphere k:

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{2n+1}{4\pi}\left[\frac{H_{nm}^{(\mathrm{I,k})}}{\hat{i}_n(\kappa a^{(\mathrm{I,k})})}\right]Y_{nm}^{(\mathrm{I,k})}(\theta,\phi)=\tilde{w}_H(\theta,\phi)=\begin{cases}\tilde{w}_{H,exposed}(\theta,\phi), & (\theta,\phi)\in\{\theta_E,\phi_E\}\\0, & (\theta,\phi)\in\{\theta_B,\phi_B\}\end{cases} \qquad (3.19a)$$

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{2n+1}{4\pi}\left[F_{nm}^{(I,k)}\right]Y_{nm}^{(I,k)}(\theta,\phi)=\tilde{w}_F(\theta,\phi)=\begin{cases}\tilde{w}_{F,exposed}(\theta,\phi), & (\theta,\phi)\in\{\theta_E,\phi_E\}\\0, & (\theta,\phi)\in\{\theta_B,\phi_B\}\end{cases} \qquad (3.19b)$$

The above equations now have the familiar form of spherical harmonic expansion of Eq. (3.4b), so we can directly evaluate the coefficients in square parentheses via the reciprocal transform Eq. (3.4c). We show below the derivation for $\boldsymbol{H}^{(I,k)}$:

$$
\begin{aligned}
\frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} &= \int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} \tilde{w}_H(\theta',\phi')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')\sin\theta' d\theta' d\phi' \\
&= \int_{\phi_E}\int_{\theta_E} \tilde{w}_{H,exposed}(\theta',\phi')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')\sin\theta' d\theta' d\phi' \\
&= \int_{\phi_E}\int_{\theta_E}\left\{\sum_{l=0}^{\infty}\sum_{s=-l}^{l}\begin{pmatrix} H_{ls}^{(I,k)}\left(\dfrac{2l+1}{\hat{i}_l(\kappa a^{(i,k)})} - e^{-\kappa a^{(I,k)}}\hat{k}_l(\kappa a^{(I,k)})\right) \\ +F_{ls}^{(I,k)} + XH_{ls}^{(I,k)} \end{pmatrix} Y_{ls}^{(I,k)}(\theta',\phi')\right\}\overline{Y_{nm}^{(I,k)}}(\theta',\phi')\sin\theta' d\theta' d\phi' \\
&= \sum_{l=0}^{\infty}\sum_{s=-l}^{l} I_{E,lsnm}^{(I,k)}\left(H_{ls}^{(I,k)}\left(\frac{2l+1}{\hat{i}_l(\kappa a^{(i,k)})} - e^{-\kappa a^{(I,k)}}\hat{k}_l(\kappa a^{(I,k)})\right) + F_{ls}^{(I,k)} + XH_{ls}^{(I,k)}\right)
\end{aligned}
\tag{3.20}
$$

where I_E, the exposed surface integral matrix, is computed using quadrature method with M_{grid} uniform surface grid points:

$$
\begin{aligned}
I_{E,lsnm}^{(I,k)} &\equiv \frac{1}{4\pi}\int_{\phi_E}\int_{\theta_E} Y_{ls}^{(I,k)}(\theta',\phi')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')\sin\theta' d\theta' d\phi' \\
&\approx \frac{1}{M_{grid}}\sum_{k=1}^{M_E} Y_{ls}^{(I,k)}(\theta_k,\phi_k)\overline{Y_{nm}^{(I,k)}}(\theta_k,\phi_k)
\end{aligned}
\tag{3.21}
$$

A similar transform to Eq. (3.20) can be written for $F_{nm}^{(I,k)}$. Finally, we truncate the series at pole order p to get the iterative equations

$$
\frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} = \sum_{l=0}^{p}\sum_{s=-l}^{l} I_{E,lsnm}^{(I,k)}\left(H_{ls}^{(I,k)}\left(\frac{2l+1}{\hat{i}_l(\kappa a^{(I,k)})} - e^{-\kappa a^{(I,k)}}\hat{k}_l(\kappa a^{(I,k)})\right) + F_{ls}^{(I,k)} + XH_{ls}^{(I,k)}\right) \tag{3.22a}
$$

$$
F_{nm}^{(I,k)} = \sum_{l=0}^{p}\sum_{s=-l}^{l} I_{E,lsnm}^{(I,k)}\begin{pmatrix} e^{-\kappa a^{(I,k)}}\left[l\hat{k}_l(\kappa a^{(I,k)}) - (2l+1)\hat{k}_{l+1}(\kappa a^{(I,k)})\right]H_{ls}^{(I,k)} \\ +(2l+1-l\varepsilon)F_{ls}^{(I,k)} + XF_{ls}^{(I,k)} \end{pmatrix} \tag{3.22b}
$$

Equations (3.22a-b), along with Eqs. (3.14a-b), represent a key result of this paper. The equations are iteratively evaluated, until the values of $\boldsymbol{F}^{(I,k)}$ and $\boldsymbol{H}^{(I,k)}$ converge to a stipulated tolerance. The operations are simply matrix-vector multiply, **y**=**Ax**, where the vector **x** is constantly updated using the latest values of $\boldsymbol{F}^{(I,k)}$ and $\boldsymbol{H}^{(I,k)}$. During computation, the surface integral coefficients $I_{E,lsnm}^{(I,k)}$, and fixed charge coefficients $E_{nm}^{(I,k)}$ and $LE_{nm}^{(I,k)}$ are pre-computed for each sphere *(I,k)* prior to simulation; while $LF_{nm}^{(I,k)}$, $LH_{nm}^{(I,k)}$, and $LHN_{nm}^{(I,k)}$ are updated via multipole-to-local operations (see implementation section below).

In summary, our approach to solve the LPBE is as follows: (1) For each sphere *k* in molecule *I*, we apply the addition theorems to express the potentials $\Phi_{in}(\mathbf{r})$ and $\Phi_{out}(\mathbf{r})$ as spherical harmonic expansions containing unknown coefficients ($F_{nm}^{(I,k)}$and $H_{nm}^{(I,k)}$) representing sphere *k's* polarization charges. (2) We impose boundary conditions on the sphere surface to derive boundary equations. (3) We account for charges from other spheres and molecules by re-expanding their polarization coefficients ($F_{nm}^{(J,l)}$and $H_{nm}^{(J,l)}$) about the center of sphere *k* using ‘multipole-to-local’ operations. (4) We then solve the boundary equations for $F_{nm}^{(I,k)}$and $H_{nm}^{(I,k)}$ iteratively using a novel fast iterative method (‘*inner*-iteration’). (5) We repeat steps (1)-(4) for all other spheres (‘*outer*-iteration’) until the convergence criteria is reached.

Convergence is monitored using relative change in $\boldsymbol{H}^{(I,k)}$between the t^{th} and $(t\text{-}1)^{th}$ outer iterations

$$\mu_{H,t}^{(I,k)} \equiv \frac{\sum_{n=0}^{p}\sum_{m=-n}^{n}\left|H_{nm,t}^{(I,k)} - H_{nm,t-1}^{(I,k)}\right|}{\frac{1}{2}\sum_{n=0}^{p}\sum_{m=-n}^{n}\left|H_{nm,t}^{(I,k)}\right| + \left|H_{nm,t-1}^{(I,k)}\right|} \tag{3.23}$$

We now can calculate the interaction energies from converged values of $\boldsymbol{H}$. The interaction energy of sphere *k* is the inner product of its effective charge distribution with the potential due to external sources. The interaction energy $W^{(I)}$of each molecule *I* is the sum of interaction energies of its constituent spheres

$$W^{(I)} = \sum_{k=1}^{N_S^{(I)}}\left\langle \mathbf{LHN}^{(I,k)}, \mathbf{H}^{(I,k)}\right\rangle = \sum_{k=1}^{N_S^{(I)}}\sum_{n=0}^{p}\sum_{m=-n}^{n} LHN_{nm}^{(I,k)}\overline{H}_{nm}^{(I,k)} \tag{3.24}$$

Implementation of re-expansion operations

To solve for $\boldsymbol{F}^{(I,k)}$ and $\boldsymbol{H}^{(I,k)}$, we need to first account for the polarization charges from all other spheres via $\boldsymbol{LF}^{(I,k)}$, $\boldsymbol{LH}^{(I,k)}$, and $\boldsymbol{LHN}^{(I,k)}$. To do this, we convert source multipoles $\boldsymbol{F}$ and $\boldsymbol{H}$ from other spheres to target local expansions centered at $\boldsymbol{c}^{(I,k)}$. If the source and target spheres are well-separated (see criterion below), the re-expansion can be accomplished analytically through multipole-to-local operators $\boldsymbol{T}_0$ and $\boldsymbol{T}_\kappa$. The procedure for computing coefficients of $\boldsymbol{T}_0$ and $\boldsymbol{T}_\kappa$ has been previously detailed in reference [3]. For *intramolecular* re-expansions (i.e. from spheres *j* to center of sphere *k* in the same molecule *I*)

$$\mathbf{LF}^{(I,k)} = \sum_{j\neq k}^{N_S^{(I)}}\mathbf{T}_0^{(I,k)(I,j)}\mathbf{F}^{(I,j)} \quad ; \quad \mathbf{LH}^{(I,k)} = \sum_{j\neq k}^{N_S^{(I)}}\mathbf{T}_\kappa^{(I,k)(I,j)}\mathbf{H}^{(I,j)} \tag{3.25}$$

or *intermolecular* re-expansions (i.e. from spheres *l* on molecule *J* to center of sphere *k* in the same molecule *I*)

$$\mathbf{LHN}^{(I,k)} = \sum_{J\neq I}^{Nmol}\sum_{l=1}^{N_S^{(J)}}\mathbf{T}_\kappa^{(I,k)(J,l)}\mathbf{H}^{(J,l)} \tag{3.26}$$

The analytical re-expansion operators are only valid when the target center $\boldsymbol{c}^{(I,k)}$ lies outside the bounding sphere of the source charge distribution, so they cannot be used in cases where source and target spheres overlap. Nonetheless, the local expansions $\boldsymbol{LF}^{(I,k)}$ and $\boldsymbol{LH}^{(I,k)}$ are still well-defined and could be directly computed using discrete versions of Eqs. (3.8c) and

(3.10b) – a procedure we termed ‘numerical re-expansion’, as described below. To our knowledge this method of circumventing the restriction by analytical re-expansion has not been previously documented.

We first discretize the surface of source sphere *j* uniformly into M_p patches, with each patch *b* centered at $\boldsymbol{r}_b^{(I,j)} = \left[a^{(I,j)}, \theta_b^{(I,j)}, \phi_b^{(I,j)}\right]$. We then compute the surface charge on the b^{th} patch $\tilde{q}_b^{(I,j)} = 4\pi\tilde{q}^{(I,j)}\left(\theta_b^{(I,j)},\phi_b^{(I,j)}\right)/M_p$, where $\tilde{q}^{(I,j)} = \tilde{f}^{(I,j)}$ or $\tilde{h}^{(I,j)}$ from Eqs. (3.15a) and (3.15b). The local expansions of sphere *j*’s multipoles re-centered on *k* are then approximated from Eqs. (3.8c) and (3.10b) as

$$LF_{nm}^{(I,k)} \approx \sum_{b=1}^{M_p} \frac{f_b^{(I,j)}}{r_b^{(I,k)}} \left(\frac{a^{(I,k)}}{r_b^{(I,k)}}\right)^n \overline{Y_{nm}}(\theta_b^{(I,k)}, \phi_b^{(I,k)}) \tag{3.27a}$$

$$LH_{nm}^{(I,k)} \approx \sum_{b=1}^{M_p} \frac{h_b^{(I,j)}}{r_b^{(I,k)}} \left(\frac{a^{(I,k)}}{r_b^{(I,k)}}\right)^n e^{-\kappa r_b^{(I,k)}} \hat{k}_n(\kappa r_b^{(I,k)}) \overline{Y_{nm}}(\theta_b^{(I,k)}, \phi_b^{(I,k)}) \tag{3.28b}$$

where $\boldsymbol{r}_b^{(I,k)} = \boldsymbol{r}_b^{(I,j)} - \left(\boldsymbol{c}^{(I,k)} - \boldsymbol{c}^{(I,j)}\right)$. The re-expansion becomes exact as M_p approaches infinity, although in practice we find that a value of $M_p \approx 2.5p^2$ adequately captures features of the surface charge distributions. Numerical re-expansion is also used in cases where the source and target spheres are non-overlapping but not well-separated, which we defined as when the distance between sphere surfaces is less than 5Å. At such short distance, analytical re-expansion requires a high number of poles for a stipulated level of error. Since both computational time and memory for $\boldsymbol{T}$ scales with p^3 it is more efficient to perform the re-expansion using direct numerical method.

We have also derived a formula using Greengard’s error bound[121] to adaptively determine the minimum pole order adequate for a re-expansion operation. To re-expand sphere *j*’s multipole to a local expansion at target center *k* within an error of ε_X, the pole order required is given by

$$p = \log\left(\frac{\sum_{\text{charges on j}} \left|\tilde{q}\right|}{\varepsilon_X a^{(I,j)}(c-1)}\right) / \log(c) - 1 \tag{3.29}$$

where $c = \dfrac{\left|\boldsymbol{c}^{(I,k)} - \boldsymbol{c}^{(I,j)}\right|}{a^{(I,j)}} - 1$, and $\tilde{q} = \tilde{f}$ or $\tilde{h}$ are the surface polarization charges. The optimal pole order is calculated on the fly every outer iteration.

Further implementation details

The surface integral coefficients $I_{E,lsnm}^{(I,k)}$ involve numerical quadratures that are pre-computed for each sphere *(I,k)*; we have found that the number of quadrature points should scale with pole number as $M_{grid} \sim 20p^2$, which we found to be adequate for capturing the spatial features of the integrand in Eq. (3.21).

To prepare a target molecule for computation, we must discretize it into a collection of overlapping spheres. To do so, we first convert its PDB file to PQR format using the PDB2PQR webserver[12-13]. We then obtain its solvent excluded surface (SES) using MSMS[122] and a chosen probe radius r_p in Å. We proceed with a Monte Carlo search algorithm to find the minimum number of spheres and corresponding radii that satisfying the following criteria:

1. The sphere surface must be at least d (in Å) away from the outermost atom center. The distance d can be held constant, or set to the van de Waals radius of each atom.
2. The surface of the spheres cannot protrude more than t (in Å) from the SES surface.

The search is terminated when each atom is encompassed by at least one sphere.

Finally, the code is implemented in C++, and is parallelized in a shared memory framework using openMP 2.0. Timings for PB-SAM and APBS in Results are based on single processor runs on an Intel(R) Xeon(R) CPU 2.27GHz processor with 24GB of physical memory; we did this to compare PB-SAM in a serial version against the APBS serial code. We performed APBS calculations section with the following parameters:

```
mg-auto, cgcent 0 0 0, fgcent 0 0 0, cglen 200 200 200, fglen 100 100 100,
mol 1,  lpbe, bcfl mdh, pdie 4.0,  sdie 78.0,  srfm mol, chgm spl2, sdens
10.00, srad 0.0, swin 0.30, temp 298.15, ion charge 1 conc 0.05 radius 0.0,
ion charge -1 conc 0.05 radius 0.0.
```

Results

Non-overlapping spherical test cases

We first assess the accuracy of PB-SAM and APBS finite difference solutions against analytical values for three test systems involving 2, 27, and 343 non-overlapping spherical dielectric cavities (of diameter 20Å, 15Å, and 5Å, respectively) with internal charges placed near the dielectric boundaries (Table 3.1). For large spheres this corresponds to a highly asymmetric charge arrangement, while as sphere size decreases the charge distribution approaches a monopole. The exact analytical solution of the PBE for multiple non-overlapping spheres has only become available recently[3]. In all cases, the salt concentration is set to 0.05M, corresponding to $\kappa = 0.07374$. Convergence is reached when the relative change $\mu_{H,t}^{(I,k)}$ falls below 10^{-2} for all spheres.

Table 3.1: *Spherical test systems for comparison of APBS and PB-SAM to analytical model solution in Tables 3.2 and 3.3. Cavities have surface-to-surface separation of 1Å from one another*

Test System	Description	Charge Configuration [position from center], charge [e]	
1	2 dielectric cavities of radius 20Å	Cavity 1 Cavity 2	[18, 0, 0], +3 [-18, 0, 0], -3
2	27 dielectric cavities of radius 15Å	All Cavities	[13, 0, 0], +1; [-13, 0, 0], -1 [0, 13, 0], +2; [0, -13, 0], -2 [0, 0, 13], +1; [0, 0, -13], -1
3	343 dielectric cavities of radius 5Å	All Cavities	[3, 0, 0], +1; [-3, 0, 0], -1 [0, 3, 0], +2; [0, -3, 0], -2 [0, 0, 3], +1; [0, 0, -3], -1

For test system 1 (two non-overlapping spheres), we computed the APBS solutions at four different grid resolutions that are typically used in biomolecular applications, and compared the potential value over the entire surface against the analytical model, as well as reporting the corresponding memory requirements and timings (Table 3.2). At the most coarse resolution we find that the APBS error can be as high as ~20% of the theoretical result; as the APBS grid spacing decreases the APBS accuracy increases, reaching ~5% of the true value. The APBS timing scales cubically with the number of grid points, as does memory cost that largely reached the limit of 27GB on our computing node at the highest resolution we tested. Using this highest resolution grid but increasing the number of spheres in test systems 2 and 3, the APBS solution gets corresponding better as the charge distribution simplifies, with average errors of ~2% and ~1%, respectively. Table 3.3 shows that the corresponding result for our PB-SAM model, in which we can quickly exceed the accuracy of the APBS solution at a fraction of the cost and memory requirements for all three systems. In all three test cases, very few poles ($p \leq 40$) are needed to define a high accuracy solution, primarily because there are no problematic deep cusp dielectric geometries in the non-overlapping sphere case.

Table 3.2. *Comparison of APBS against analytical model for test systems described in Table 3.1.*

Test System	Grid Points	Resolution (Å)	Run time [s]	Memory [GB]	Overall Relative Error	Maximum Relative Error
1	65x65x65	1.5625	3	0.08	19.7%	34.8%
1	129x129x129	0.7813	29	0.47	14.4%	24.7%
1	257x257x257	0.3906	142	3.50	11.2%	31.7%
1	513x513x513	0.1953	1315	27.8	4.9%	11.4%
2	513x513x513	0.1953	1216	27.8	1.9%	5.3%
3	513x513x513	0.1953	1421	27.8	1.1%	4.9%

Table 3.3. *Comparison of PB-SAM against analytical model for test systems described in Table 3.1.*

Test System	Number of multipoles	Run time [s]	Memory [GB]	Overall Relative Error	Maximum Relative Error
1	30	4.3	23	13.5%	17.6%
1	35	12.1	31	4.3%	4.6%
1	40	20.7	51	2.4%	1.9%
2	10	1.4	15	13.6%	26.7%
2	15	2.3	21	6.4%	11.8%
2	20	7.6	33	2.2%	4.1%
2	30	46.5	82	0.4%	4.4%
3	5	22.2	108	4.4%	9.6%
3	10	28.4	167	0.1%	0.3%

Overlapping spherical test cases

Our second comparison involves two overlapping spheres of various sizes. In this case no analytical solution is known, but we can define a benchmark calculation based on a high quality PB-SAM solution computed at p=140 and M_p=200,000 (PB-SAM140). In Table 3.4, we compare the relative difference in surface potential against PB-SAM140 as sphere size increases. We considered the worst-case scenario by placing the positive charge close to the surface, at a fixed distance of 1.73 Å below the cusp region, so that as sphere size grows it results in higher asymmetry of the charge distribution. For each sphere radius, we first compared the surface potential computed by APBS against that of PB-SAM140, and then seek a corresponding set of PB-SAM parameters that provide solutions with comparable relative errors. PB-SAM at 40 and 60 poles is able to achieve relative difference comparable to APBS with comparable total solve time, and with less memory requirements. We want to point out that the total solve time for PB-SAM reported in Table 3.4 is principally dominated by the one-time cost of surface integral computation (1140s), while the actual time for solving the iterative equations, Eqs. (3.22a and 3.22b), are between 9 seconds to 2 minutes.

Table 3.4: *Two overlapping spheres with varying sphere sizes.* Comparison of the surface potential computed with APBS and PB-SAM (M_{grid}=100k, M_p=2.5p^2) against PB-SAM140.

	APBS					**PB-SAM**				
Sphere Size	Grid Size [Å]	Solve Time [s]	Memory [GB]	Relative Error	Maximum Relative Error	Pole Order	Solve Time [s]	Memory [GB]	Relative Error	Maximum Relative Error
2	0.0195	960	27.8	0.6%	1.6%	20	1141	0.018	12.1%	16.1%
						30	1143	0.030	7.0%	9.2%
						40	1149	0.057	3.8%	5.6%
						60	1209	0.230	1.5%	2.1%
5	0.0391	1018	27.8	1.6%	3.5%	20	1141	0.018	14.4%	20.6%
						30	1143	0.030	7.7%	10.5%
						40	1148	0.057	5.5%	7.8%
						60	1315	0.229	2.3%	3.9%
15	0.1172	1,158	27.8	4.6%	9.7%	20	1141	0.018	21.7%	30.4%
						30	1143	0.030	12.5%	18.4%
						40	1148	0.057	8.6%	14.4%
						60	1223	0.229	4.3%	6.9%
50	0.3906	1,276	27.8	16.8%	32.8%	20	1141	0.018	41.8%	36.2%
						30	1142	0.030	27.7%	25.3%
						40	1148	0.057	19.6%	18.2%
						60	1180	0.229	11.3%	13.3%

It is interesting to note that, for a fixed number of poles, PB-SAM's relative error increases with increasing sphere sizes. Since the boundary equations are formulated and solved in scaled representations, they should be independent of sphere sizes. The two potential sources of error are if M_p is insufficient in discriminating the positions of the source surface charges for numerical re-expansion, or the scaled fixed charge multipole $E_{nm}^{(I,k)}$ decays more slowly with poles with increasing charge asymmetry. While we found that increasing M_p by a factor of 40 resulted in no change in potential, the term $(r_\alpha/a^{(I,k)})^n$ in $E_{nm}^{(I,k)}$ converges slower at large sphere radii, hence more poles are needed to describe the corresponding increase in charge asymmetry. In practice, charges in realistic biomolecules are more evenly distributed, hence their fixed charge multipoles will converge much faster. The convergence improves further when smaller spheres are used to define higher resolution dielectric boundaries. Hence Table 3.4 shows that PB-SAM's relative error decreases with smaller spheres and higher pole, and our simplified case with maximum charge asymmetry provides a worse case upper bound on the relative error for $20 < p < 60$. This will inform estimates of error in our calculation of the bromine mosaic virus in the following section.

The Bromine Mosaic Virus

We have also applied our PB-SAM method to solve for the potential around a biological molecule, the T=1 particle of the brome mosaic virus (BMV) capsid (PBD code: 1YC6). The virus has been shown to convert from T=3 (comprising of 180 monomers) to T=1 (comprising of 60 monomers) capsid under proteolytic conditions[4]. Each capsid protein monomer is comprised of 154 amino acids. To prepare the PDB file for calculation, we converted chain A of the PDB file into PQR format using the PDB2PQR server[123, 124], which also assigned partial atomic charges using the AMBER 99 force field[125]. We then discretized the protein into a collection of overlapping spheres using an in-house algorithm (see implementation details in Methods). Using discretization criteria that varies in spatial resolution, we generated three representations of the protein monomer with 107, 354 and 712 spheres, and Figure 3.3 compares the dielectric boundary representation against the solvent excluded surface computed using MSMS[122] with probe radius $r_p = 1.4$ Å.

It is our intention to study the dynamics of BMV capsid assembly via Brownian dynamics in future work. Therefore Table 3.5 describes the computational time and memory resources for PB-SAM to calculate the self-polarization of one 1YC6 monomer, and the mutual polarization of an array of 60 monomers that make up the unassembled BMV capsid. We therefore consider the breakdown of computational cost and memory as (1) a one time cost to prepare the surface integral of the chosen dielectric representation of the 1YC6 monomer, (2) the one-time cost to self-polarize each monomer, and (3) the cost to mutually polarize the 60 monomers. In the context of a Brownian dynamic simulation, Table 3.5 represents the cost of the initialization phase that will require "cold" guesses for $\boldsymbol{F}$ and $\boldsymbol{H}$ for steps (2) and (3) and the timings will be non-optimal relative to later solutions that will provide better initial guesses as the dynamics algorithm proceeds as the capsid assembles.

The PB-SAM computational cost depends on the number of poles and number of spheres, and timings are faster or slower depending on how much of the calculation can be done in memory. Using Table 3.4, we will focus our PB-SAM solutions at a ~5-7% error by choosing pole order $20 < p < 60$ and keeping average size of spheres of the dielectric boundary representation between 2-5Å. For step (1), Table 3.5 shows the one time surface integral cost of the 1YC6 monomer, which scales as $O(M_{grid}p^4)$ (see methods), varies between several minutes to several hours. However, a nice benefit is that as resolution increases the sphere size and hence M_{grid} decreases as do the number of needed poles, which together mitigates the time of calculating more spheres. The cost to self-polarize will depend on the available memory; in memory-saving mode the re-expansion operators $\boldsymbol{T_0}$ and $\boldsymbol{T_\kappa}$ are computed on the fly, instead of being stored in memory and hence increase the cost of the calculation. In Table 3.5, the self-polarization timings are based on a "cold" guess of $\boldsymbol{F}^{(I,k)} = 0$ and $\boldsymbol{H}^{(I,k)}$ approximated using the fixed charges, and iterated until the relative change in $\boldsymbol{H}^{(I,k)}$ falls below 10^{-2} for all spheres. For the 1YC6 monomer, the APBS result is necessarily evaluated at a low resolution of 0.22Å based on the maximum allowed grid points of 513^3 given our maximum memory of 24GB; Table 3.4 suggests that the APBS relative error would be ~10-12% for this system. Therefore it is evident that with 30-40 poles for the representations of 107 and 353 spheres, and 20-30 poles for 712 spheres, we have arguably a higher quality solution at a comparable cpu cost and memory of the APBS solution.

Table 3.5. *Computational timing and memory resources using PB-SAM for capsid assembly.* Self-polarization of 1YC6 monomer and mutual polarization of 60 monomers of BMV capsid for various dielectric representations (Figure 3.3)

Number and median sphere radius	Poles	Time to calculate surface integrals [s]	Self-polarization		Mutual-polarization	
			Time [s]	Memory [GB]	Time [s]	Memory [GB]
107 spheres	40	1,083	280	3.6	2,589	4.4 **
4.40 Å	50	4,131	552	7.2		
	60	12,336	1,180	13.3		
354 spheres	30	423	603	7.8		
3.06 Å	40	2,380	2,091	7.1**	9,365	13.5 **
	50	9,079	4,934	17.1**		
712 spheres	20	70	271	8.6		
1.91 Å	30	802	1,177	17.5	16,046	33.8 **
	40	4,508	3,707	14.2**		

** memory-saving mode

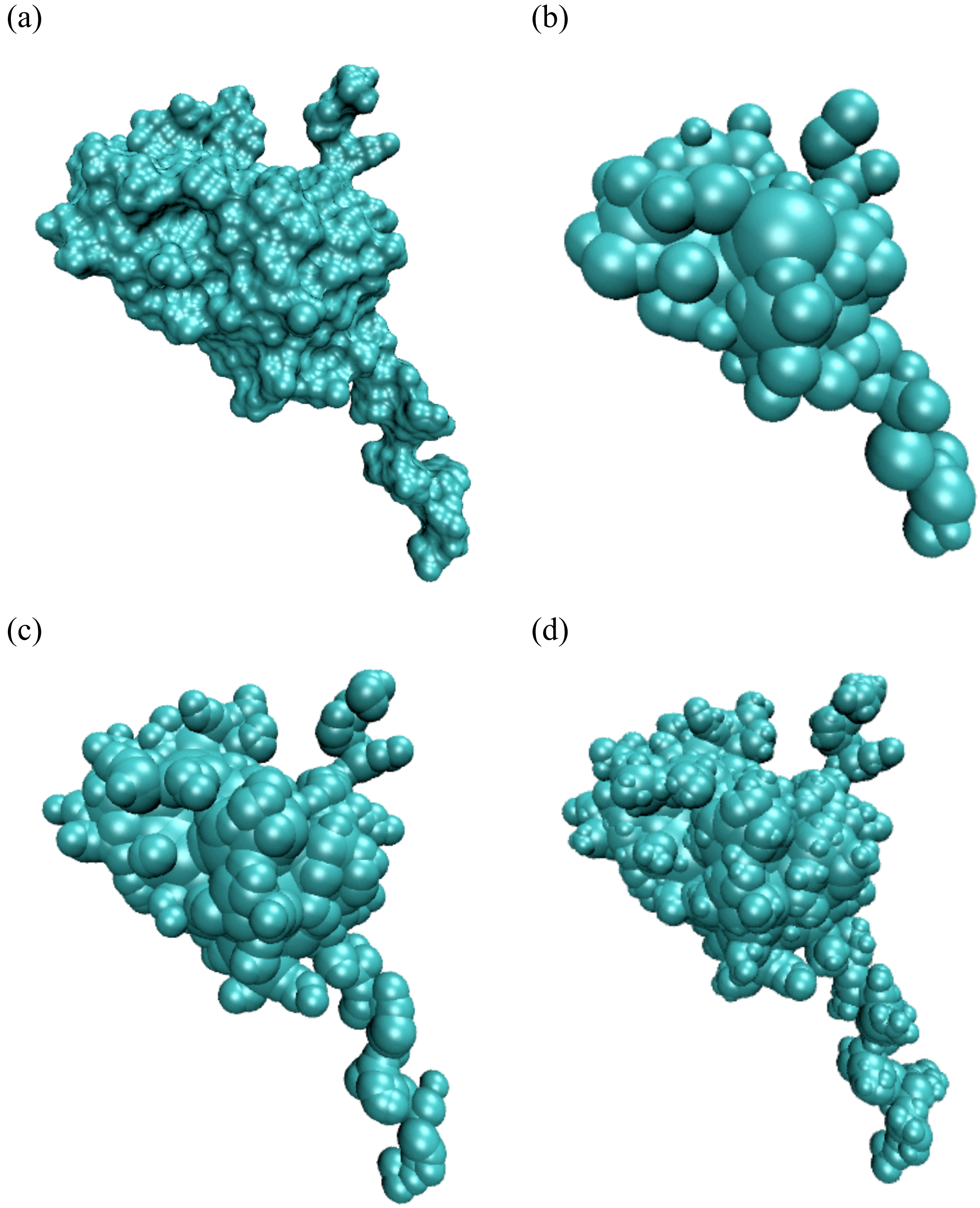

Figure 3.3. ***Representations of 1YC6 monomer based on different discretization criteria.*** (a) the solvent excluded surface computed using MSMS with p = 1.4Å. (b) 107 spheres with p = 1 Å, d = 1 Å, t = 2 Å, (c) 354 spheres with p = 1 Å, d = atomic vdW radii, t = 1 Å (d) 712 spheres with p = 1 Å, d = atomic vdW radii, t = 0.5 Å.

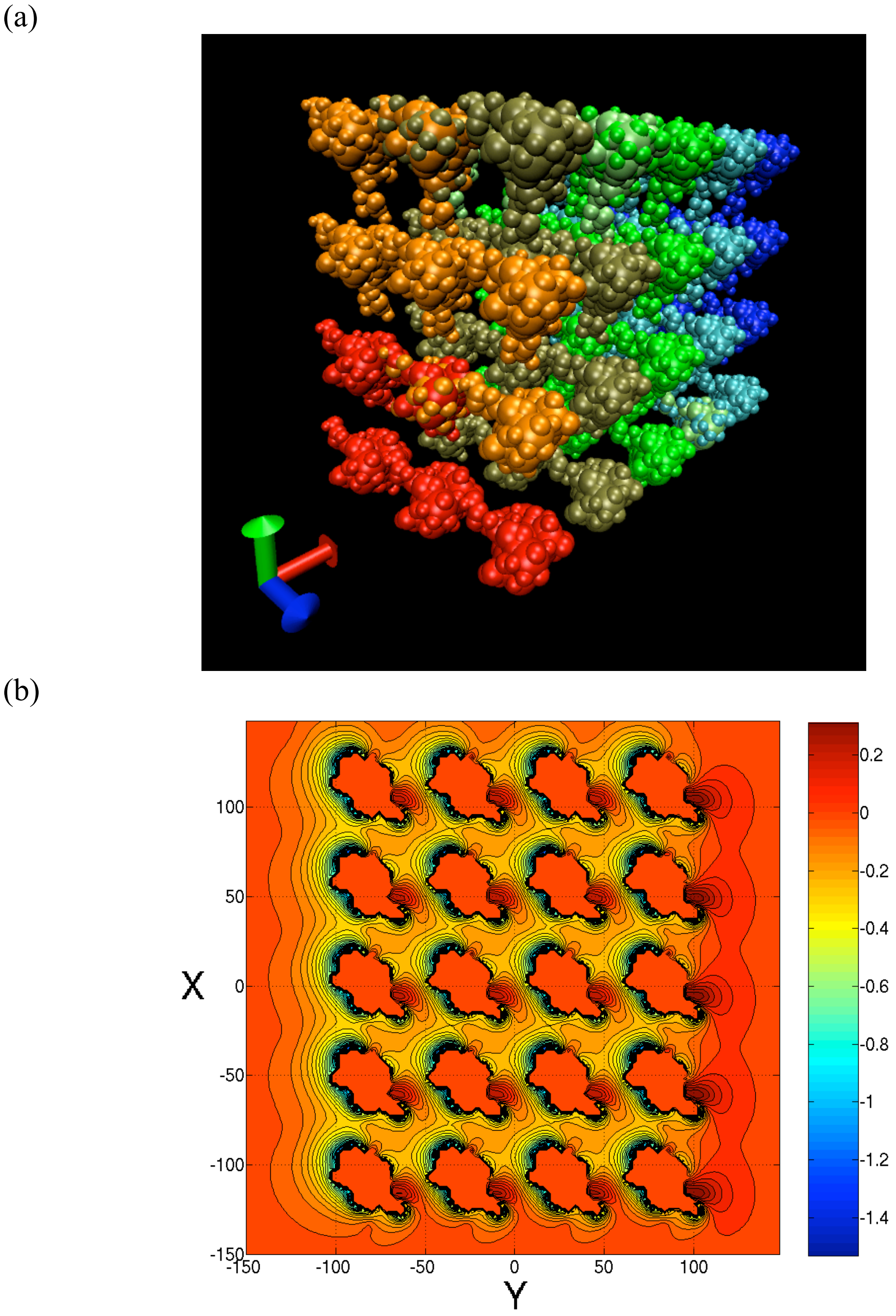

Figure 3.4. ***Array of 60 virus monomers.*** (a) Array configuration (b) Potential profile of a cross-section through the z=0 plane with twenty monomers. Contour lines at 0.05 kT.

Finally we have evaluated the potential of an assembly of 60 copies of 1YC6 monomers in a 5 x 4 x 3 array, corresponding to a system size of 165 Å x 220 Å x 275 Å (Figure 3.4). The array configuration is intended to mimic late stage assembly, at which the entire capsid system is compact and mutual polarization becomes significant and more difficult to converge (as opposed to the 60 monomers being well separated). All monomers were given the same initial guess of $\boldsymbol{F}^{\text{self}}$ and $\boldsymbol{H}^{\text{self}}$ from the converged self-polarization step, and the computational time and memory to calculate the total (self and mutual) polarization is given in Table 3.5. The memory for the 712-sphere representation required 33GB of virtual memory, which is not as efficient if it were able to fit in the available 24GB of physical memory. The fact that the calculation of a high quality solution is doable on a single standard commodity node is a strength of the PB-SAM approach, although further optimization will be explored in the future.

Conclusion

We have developed a novel method for solving the linearized Poisson Boltzmann equation by discretizing the protein surface as a collection of spheres, in which the surface charges can be iteratively solved by our recent analytical solution of the PBE equations for spherical geometries in which mutual polarization is treated exactly[3]. We have compared PB-SAM and the finite difference PB solver APBS against two new benchmarks never before available to compare numerical methods. First we show that PB-SAM converges to the analytical solution of hundreds of spheres with better accuracy and at greatly reduced cost relative to APBS. Second the PB-SAM solution using 140 poles allows us to define a high quality benchmark to describe the electrostatic potential for two overlapping spheres that are models for cusp-like features of protein active sites, in which we show that our PB-SAM solution converges to the correct solution with the same computational cost or better than the finite difference solution. Finally we illustrate the strength of the PB-SAM approach by computing the potential profile of a close configuration of 60 T1-particle forming monomers of the bromine mosaic virus (PDB code 1YC6), with clear improvements in accuracy relative to other numerical PB solutions, given a fixed hardware configuration of physical memory.

Further development is necessary to enable PB-SAM's application in large-scale Brownian dynamic simulations. The current version of PB-SAM expends significant computational time solving Eqs. 3.22(a-b) iteratively. This step was implemented simply as repeated calls to the BLAS matrix-vector multiply routine *dgemv*, but can be accelerated by preconditioning Eqs. 3.22(a-b) and using a more sophisticated linear system solving method, such as generalized minimal residual method. We also noted during our benchmarking studies that when our current convergence criterion is relaxed, the resulting surface potential is unchanged, so there is room explore a less stringent but adequate convergence criterion. Finally, forces and torques are required for Brownian dynamic simulation. We have derived in reference [3] how forces and torques can be computed analytically for spherical dielectrics. The same formulation can be extended to the overlapping sphere representation in PB-SAM via superposition, which is on-going work in our lab.

Acknowledgements We gratefully acknowledge support from an NIH Multiscale grant and NERSC for computational resources, and Dr. Itay Lotan for providing source code for reference [3] and clarifications.

Chapter 4

Brownian Dynamics of Particles in a Poisson Boltzmann Continuum

Introduction

In Chapter 3 we presented a new mathematical formalism, the Poisson-Boltzmann Semi-Analytical Method (PB-SAM), for solving the linearized Poisson Boltzmann equation. In this chapter, we derive the corresponding force and torque equations, using a variational approach to account for mutual polarization.

When a dielectric cavity is placed in a medium of a different dielectric constant, the cavity develops surface charges in response to the dielectric discontinuity across the boundary, in a process we termed *self-polarization*. This is in accordance to Gauss's Law, which states that the net normal electric displacement flux ($\oint_S \varepsilon \mathbf{E} \cdot \partial \mathbf{A}$) emanating from a closed surface must be equal to the net charge in that enclosed volume. Now, when two dielectric cavities are placed in close proximity to each other, the electric field from the first cavity's fixed and self-polarized charges induces additional polarization on the second cavity, which in turn induces additional charges on the first. This iterative charging process is termed *mutual polarization*.

Mutual polarization is negligible for well separated cavities, but becomes dominant at small separation distance *d*. Forces and torques computed with and without accounting for mutual polarization essentially agree for $d > 40$Å (for what size cavity and charge distribution), but differ by more than 80% at $d = 2$ Å[3]. It is hence clear that inclusion of (at least some) mutual polarization effect is essential for realistic modeling of intermolecular electrostatic interactions at close range.

To account for mutual polarization, we need to solve for two quantities: the mutual polarized charge distribution and its gradient. Solving the mutual polarized charge distribution allows us to compute a system's total energy. If we are interested in calculating the force on a particular site *i*, we must also compute the gradient, i.e. how the mutual polarized charge distribution will change with respect to the position of *i*.

Accounting for mutual polarization effects requires significant computation effort. To compute the force at each atom *i*, one could employ the 'virtual work method'[126], i.e. a finite difference solution, in which each atom is displaced slightly and the system re-solved. Three separate solutions are required (one in each direction) to compute the gradient at atom *i*. The process is then repeated for other atoms. Gilson[127] derived force expression using a Maxwell stress tensor that has been widely used[48, 49], although its use is limited to a single cavity and suffers from hyper-singularity issues when deployed on multiple cavities[128].

Itay and Head-Gordon[3] and Lu *et. al.*[128] independently developed solutions for solving the lineared PBE for multiple cavities. Reference [3] was based on an analytical approach, while [128] was based on numerical boundary element methods, but both used essentially the same fundamental theory and variational approach to force[127] calculation.

In this chapter we derived expressions for forces and torques using the same variational approach described by references [3] and [128], and then implement the derived equations with the PB-SAM model. The force and torque computation is then coupled to a Brownian dynamics algorithm. Force and torque calculations were performed on two monomers of the T1-particle forming brome mosaic virus for two successive timesteps, and the timings are reported.

Methods

The method for solving mutually polarized charge distribution has been presented in Chapter 3. Here we first summarize the principal result, namely the solution of the charge distribution due to mutual polarization, followed by derivation of expressions for force and torque.

Solution of mutually polarized charge distribution

The system of interest comprises of N_{mol} macromolecules immersed in an implicit aqueous salty solvent. Each macromolecule I is embedded with $N_C^{(I)}$ fixed partial charge and represented as a collection of $N_S^{(I)}$ overlapping spheres with dielectric constant ε_{in}. The solvent is treated as a continuum with dielectric constant ε_{out}, with screening effects due to mobile ions captured via the inverse Debye length κ. The linearized PB equation (LPBE) gives the potential Φ at any point **r** in space $\Re^3$ as

$$-\nabla\left[\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})\right]+\kappa^2\Phi(\mathbf{r})=4\pi\rho_{fixed}(\mathbf{r}) \tag{4.1}$$

where ε is the relative dielectric function, ρ_{fixed} is the charge density due to the fixed protein partial charges, and $\kappa=\sqrt{8\pi\bar{n}e^2/\varepsilon_{out}k_BT}$, where $\bar{n}$ is the bulk concentration of monovalent salt in the solution, e is the fundamental electronic charge, k_B the Boltzmann constant, and T the absolute temperature.

The potentials inside and outside a molecule I are given respectively by:

$$\Phi_{out}(\mathbf{r})=\sum_{I=1}^{N_{mol}}\left(\frac{1}{4\pi}\int_{d\Omega^{(I)}}\frac{e^{-\kappa|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}h^{(I)}(\mathbf{r}')d\mathbf{r}'\right) \tag{4.2a}$$

$$\Phi_{in}^{(I)}(\mathbf{r}) = \sum_{\alpha=1}^{N_C^{(I)}} \frac{1}{\left|\mathbf{r}-\mathbf{r}_\alpha^{(i)}\right|}\frac{q_\alpha^{(I)}}{\varepsilon_{in}} + \frac{1}{4\pi}\int_{d\Omega^{(I)}} \frac{1}{\left|\mathbf{r}-\mathbf{r}'\right|} f^{(I)}(\mathbf{r}')d\mathbf{r}' \qquad (4.2b)$$

where *f(**r**)* and *h(**r**)* are the mutually polarized, *reactive* and *effective* surface charge distributions respectively. Charges on the surface of sphere *k,* denoted as $f^{(I,k)}(\mathbf{r})$ and $h^{(I,k)}(\mathbf{r})$, can be transformed into *reactive* and *effective* multipoles:

$$F_{nm}^{(I,k)} \equiv \frac{1}{4\pi}\int_{d\Omega^{(I,k)}} f^{(I,k)}(\mathbf{r}')\left(\frac{a^{(I,k)}}{r'}\right)^{n+1}\overline{Y_{nm}^{(I,k)}}(\theta',\phi')d\mathbf{r}' \qquad (4.3a)$$

$$H_{nm}^{(I,k)} \equiv \frac{1}{4\pi}\int_{d\Omega^{(i,k)}} h^{(I,k)}(\mathbf{r}')\left(\frac{r'}{a^{(I,k)}}\right)^{n}\hat{i}_n(\kappa r')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')d\mathbf{r}' \qquad (4.3b)$$

Coefficients of $\boldsymbol{F}^{(k,I)}$ and $\boldsymbol{H}^{(k,I)}$ can then be solved iteratively using

$$F_{nm}^{(I,k)} = \left\langle \mathbf{I}_{E,nm}^{(I,k)}, \mathbf{WF}^{(I,k)} \right\rangle \qquad (4.4a)$$

$$\frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} = \left\langle \mathbf{I}_{E,nm}^{(I,k)}, \mathbf{WH}_H^{(I,k)} \right\rangle \qquad (4.4b)$$

where $\boldsymbol{WF}^{(k,I)}$ and $\boldsymbol{WH}^{(k,I)}$ are scaled multipoles computed from fixed charges and polarization charges from other spheres (see Chapter 3 for definitions) , and $\boldsymbol{I_E}^{(k,I)}$ is a surface integral over the exposed surface, defined by

$$I_{E,lsnm}^{(I,k)} \equiv \frac{1}{4\pi}\int_{\phi_E}\int_{\theta_E} Y_{ls}^{(I,k)}(\theta',\phi')\overline{Y_{nm}^{(I,k)}}(\theta',\phi')\sin\theta' d\theta' d\phi' \qquad (4.5)$$

We can evaluate the interaction energies upon solution of $\boldsymbol{F}^{(k,I)}$ and $\boldsymbol{H}^{(k,I)}$. The interaction energy of sphere *k* with the external field is the inner product of its effective multipole $\boldsymbol{H}^{(k,I)}$ with its local expansion of external (i.e. intermolecular) effective charges, $\boldsymbol{LHN}^{(k,I)}$. The total interaction energy of molecule *I* is in turn the sum of interaction energies of all constituent spheres.

$$W^{(I)} = \sum_{k=1}^{N_S^{(I)}} \left\langle \mathbf{LHN}^{(I,k)}, \mathbf{H}^{(I,k)} \right\rangle = \sum_{k=1}^{N_S^{(I)}}\sum_{n=0}^{p}\sum_{m=-n}^{n} LHN_{nm}^{(I,k)}\overline{H}_{nm}^{(I,k)} \qquad (4.6)$$

Force on an effective charge

The surface of each sphere can be discretized into M_P grid points, of which M_E (M_B) are exposed (buried). Since buried grid points have no surface charge and experience no force, we only need to consider the force experienced at each exposed grid point *P*.

We begin by deriving an expression for the force at *P*, and then summing up contributions from all exposed charges to derive the total force and torque on molecule *I*. We only consider forces due to external field, because forces due to intramolecular effective charges cancel out and do not contribute to the overall force and torque on molecule *I*.

We shall use the shorthand h_P to denote $h(\mathbf{r}_P)$, the effective charge at space position *P*. The multipole coefficient $\mathbf{H}_P^{(I,k)}$ is the product of h_P and the spherical harmonic $\mathbf{Y}_P^{(I,k)} \equiv \mathbf{Y}^{(I,k)}(\theta_P, \varphi_P)$. The force $\boldsymbol{f}_P$ acting on the effective charge at point *P* is the negative gradient of the interaction energy of charge h_P with the external field:

$$\begin{aligned}
\mathbf{f}_P &= -\nabla_P W_P = -\nabla_P \left\langle \mathbf{LHN}^{(I,k)}, \mathbf{H}_P^{(I,k)} \right\rangle \\
&= -\left\langle \nabla_P \mathbf{LHN}^{(I,k)}, \mathbf{H}_P^{(I,k)} \right\rangle - \left\langle \mathbf{LHN}^{(I,k)}, \nabla_P \mathbf{H}_P^{(I,k)} \right\rangle \\
&= -\left\langle \nabla_P \mathbf{LHN}^{(I,k)}, \mathbf{H}_P^{(I,k)} \right\rangle - \left\langle \mathbf{LHN}^{(I,k)}, \nabla_P h_P \cdot \mathbf{Y}_P^{(I,k)} \right\rangle - \left\langle \mathbf{LHN}^{(I,k)}, h_P \cdot \nabla_P \mathbf{Y}_P^{(I,k)} \right\rangle
\end{aligned} \tag{4.7}$$

In rigid body dynamics, the translational force on a molecule acting through its center of mass is the sum of all forces acting on all its constituent parts. Summing up $\boldsymbol{f}_P$ from Eq. (4.7) from all exposed points, we get the translational force $\boldsymbol{f}_I$ as

$$\begin{aligned}
\mathbf{f}_I &= \sum_{all\,P} \mathbf{f}_P = \sum_{k}^{Nk^{(I)}} \sum_{P \in k} \mathbf{f}_P \\
&= -\sum_{k}^{Nk^{(I)}} \sum_{P \in k} \left\langle \nabla_P \mathbf{LHN}^{(I,k)}, \mathbf{H}_P^{(I,k)} \right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, \nabla_P h_P \cdot \mathbf{Y}_P^{(I,k)} \right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, h_P \cdot \nabla_P \mathbf{Y}_P^{(I,k)} \right\rangle
\end{aligned} \tag{4.8}$$

The last inner product represents the traditional 'direct' force between charges and is equivalent to $\sum_P \sum_{\text{ext}\,Q} h_P h_Q \nabla_P \left(\exp(-\kappa R_{PQ}) / R_{PQ} \right)$. The first two inner products are best understood from a variational perspective: the operator ∇_P measures how a scalar field changes with the position of *P*, so the first two inner products account for how the magnitude of mutually polarized charges change as *P* moves. These nonlinear changes are dependent of the instantaneous configuration of the molecules, and must be solve numerically. In cases where mutual polarization is neglected, polarization charges are fixed to their self-polarized value, and moving *P* does not induce any changes in polarization charges. The first two inner products are hence dropped and we are left with the familiar direct force expression used in fast multipole methods for discrete fixed charges.

If we wish to include mutual polarization in our force computation, all three components need to be included. The large number of surface charges makes it impractical to consider a variational (or 'virtual work') treatment of each charge. However, since charges on one molecule are constrained to move concertedly, we can use the variational approach about the center of mass $\boldsymbol{c}^{(I)}$ of each molecule, and consider how the quantity in question changes with the position of $\boldsymbol{c}^{(I)}$. This is the method used in references [3] and [128], although they did not discuss the following caveat: the approach does not consider variation due to rotational movement of molecule *I* when determining the mutual polarization forces. That is, we compute the gradients by exploring only a sub-portion of the neighboring configuration space where molecules maintain their same orientations but are translated with respect to each other. Fortunately we are justified in this approximation because the translational diffusion constants are much larger than rotational diffusion constants. We denote this ‘gradient at I under fixed orientation’ as $\tilde{\nabla}_I$ to highlight its difference from convention gradient operator ∇. Since all points P move concertedly with I under translation, we can replace ∇_P in Eq. (4.8) with $\tilde{\nabla}_I$:

$$\begin{aligned}
\mathbf{f}_I &= -\sum_{k}^{Nk^{(I)}} \sum_{P \in k} \left\langle \tilde{\nabla}_I \mathbf{LHN}^{(I,k)}, \mathbf{H}_P^{(I,k)} \right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, \tilde{\nabla}_I q_P \cdot \mathbf{Y}_P^{(I,k)} \right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, q_P \cdot \tilde{\nabla}_I \mathbf{Y}_P^{(I,k)} \right\rangle \\
&= -\sum_{k}^{Nk^{(I)}} \sum_{P \in k} \left\langle \tilde{\nabla}_I \mathbf{LHN}^{(I,k)}, \mathbf{H}_P^{(I,k)} \right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, \tilde{\nabla}_I q_P \cdot \mathbf{Y}_P^{(I,k)} \right\rangle \\
&= -\sum_{k}^{Nk^{(I)}} \left\langle \tilde{\nabla}_I \mathbf{LHN}^{(I,k)}, \sum_{P \in k} \mathbf{H}_P^{(I,k)} \right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, \sum_{P \in k} \tilde{\nabla}_I \mathbf{H}_P^{(I,k)} \right\rangle \\
&= -\sum_{k}^{Nk^{(I)}} \left\langle \tilde{\nabla}_I \mathbf{LHN}^{(I,k)}, \mathbf{H}^{(I,k)} \right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, \tilde{\nabla}_I \mathbf{H}^{(I,k)} \right\rangle
\end{aligned} \tag{4.9}$$

The last inner product on the first line was dropped because the spherical harmonic $\boldsymbol{Y}_P^{(I,k)}$ centered at $\boldsymbol{c}^{(I,k)}$ is unchanged if molecule I does not change its orientation.

Solution mutually polarized gradients

We now need to compute gradients $\tilde{\nabla}_I \mathbf{H}^{(I,k)}$ and $\tilde{\nabla}_I \mathbf{LHN}^{(I,k)}$ to account for how the position of molecule I changes the polarization charges. The gradient $\tilde{\nabla}_I \mathbf{LHN}^{(I,k)}$ is given by

$$\tilde{\nabla}_I \mathbf{LHN}^{(I,k)} = \sum_{J \neq I} \sum_{j}^{N_S^{(J)}} \tilde{\nabla}_I \mathbf{T}_\kappa^{(I,k)(J,j)} \mathbf{H}^{(J,j)} + \sum_{J \neq I} \sum_{j}^{N_S^{(J)}} \mathbf{T}_\kappa^{(I,k)(J,j)} \tilde{\nabla}_I \mathbf{H}^{(J,j)} \tag{4.10}$$

where $\boldsymbol{T}_\kappa^{(I,k)(J,j)}$ denotes the multipole-to-local re-expansion operator (see Chapter 3). The first sum can be computed from the converged solutions of effective multipoles $\boldsymbol{H}$. For the second sum, we would need $\tilde{\nabla}_I \mathbf{H}^{(J,j)}$. That is, for each sphere, we need to compute the gradient of its effective multipole $\mathbf{H}$, with respect to *every* molecule I.

The gradient polarization step thus comprises of three nested iteration loops. The outermost loop goes over $1 \leq J \leq N_{mol}$ to compute gradients with respect to each J. The middle and innermost loops then solve for $\tilde{\nabla}_J \mathbf{H}^{(I,k)}$ of all spheres in a procedure analogous to the multipole polarization loops described in Chapter 3. Below we detailed the formulism for solving $\tilde{\nabla}_J \mathbf{H}^{(I,k)}$ for a sphere *(I,k)*.

To compute $\tilde{\nabla}_J \mathbf{H}^{(I,k)}$ with respect to molecule J, we begin by applying the gradient operator $\tilde{\nabla}_J$ to equations (3.22a) and (3.22b):

$$\tilde{\nabla}_J F_{nm}^{(I,k)} = \left\langle \mathbf{I}_{E,nm}^{(I,k)}, \tilde{\nabla}_J \mathbf{WF}^{(I,k)} \right\rangle \tag{4.11a}$$

$$\tilde{\nabla}_J \frac{H_{nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})} = \left\langle \mathbf{I}_{E,nm}^{(I,k)}, \tilde{\nabla}_J \mathbf{WH}_H^{(I,k)} \right\rangle \tag{4.11b}$$

If molecule *I* is fixed in orientation, $\boldsymbol{I}_E$, $\boldsymbol{E}^{(I,k)}$, and $\boldsymbol{LE}^{(I,k)}$ do not depend on position of *I* since their positions move concertedly with *I*, so we are only concerned with $\tilde{\nabla}_J \mathbf{WF}^{(I,k)}$and $\tilde{\nabla}_J \mathbf{WH}^{(I,k)}$:

$$\begin{aligned}\tilde{\nabla}_J WF_{nm}^{(I,k)} = {} & e^{-\kappa a^{(I,k)}}\left[l\hat{k}_l(\kappa a^{(I,k)}) - (2l+1)\hat{k}_{l+1}(\kappa a^{(I,k)})\right]\tilde{\nabla}_J H_{ls}^{(I,k)} \\ & + (2l+1-l\varepsilon)\tilde{\nabla}_J F_{ls}^{(I,k)} - n\varepsilon a^{(I,k)}\tilde{\nabla}_J LF_{nm}^{(I,k)} \\ & + a^{(I,k)}\left[n\hat{i}_n(ka^{(I,k)}) + \frac{\left(\kappa a^{(I,k)}\right)^2 \hat{i}_{n+1}(ka^{(I,k)})}{2n+3}\right]\left(\tilde{\nabla}_J LH_{nm}^{(I,k)} + \tilde{\nabla}_J LHN_{nm}^{(I,k)}\right)\end{aligned} \tag{4.12a}$$

$$\begin{aligned}\tilde{\nabla}_J WH_{nm}^{(I,k)} = {} & \left(\frac{2l+1}{\hat{i}_l(\kappa a^{(I,k)})} - e^{-\kappa a^{(I,k)}}\hat{k}_l(\kappa a^{(I,k)})\right)\tilde{\nabla}_J H_{ls}^{(I,k)} + \tilde{\nabla}_J F_{ls}^{(I,k)} \\ & + a^{(I,k)}\tilde{\nabla}_J LF_{nm}^{(I,k)} - a^{(I,k)}\hat{i}_n(ka^{(I,k)})\left(\tilde{\nabla}_J LH_{nm}^{(I,k)} + \tilde{\nabla}_J LHN_{nm}^{(I,k)}\right)\end{aligned} \tag{4.12b}$$

During each middle-loop iteration, we consider one sphere *(I,k)*, and compute the local expansions $\tilde{\nabla}_J \mathbf{LF}^{(I,k)}$, $\tilde{\nabla}_J \mathbf{LH}^{(I,k)}$, and $\tilde{\nabla}_J \mathbf{LHN}^{(I,k)}$ from outer spheres' polarized gradients. The local expansions are defined below:

$$\tilde{\nabla}_J \mathbf{LF}^{(I,k)} = \sum_{j\neq k}^{N_S^{(I)}} \mathbf{T}_0^{(I,k)(I,j)}\tilde{\nabla}_J \mathbf{F}^{(I,j)} \quad ; \quad \tilde{\nabla}_J \mathbf{LH}^{(I,k)} = \sum_{j\neq k}^{N_S^{(I)}} \mathbf{T}_\kappa^{(I,k)(I,j)}\tilde{\nabla}_J \mathbf{H}^{(I,j)} \tag{4.13a}$$

$$\tilde{\nabla}_J \mathbf{LHN}^{(I,k)} = \sum_{J\neq M}\sum_{m}^{N_S^{(M)}} \tilde{\nabla}_J \mathbf{T}_\kappa^{(I,k)(M,m)}\mathbf{H}^{(M,m)} + \sum_{J\neq I}\sum_{j}^{N_S^{(M)}} \mathbf{T}_\kappa^{(I,k)(M,m)}\tilde{\nabla}_J \mathbf{H}^{(M,m)} \tag{4.13b}$$

Note that intramolecular re-expansions (within same molecule *I*) does not have a $\tilde{\nabla}_J \mathbf{T}\cdot\mathbf{H}$ component, since for intramolecular re-expansion the operation $\boldsymbol{T}$ does not change with position of *I*.

As discussed in Chapter 3, analytical re-expansions can be performed between well-separated spheres. Details of computing $\tilde{\nabla}_J \mathbf{T}$ coefficients were described in reference [3]. For spheres in close spatial proximity the analytical re-expansions break down, and we resort to numerical re-expansions. To do this, we first compute the vector representing the gradient at each exposed points *P*:

$$\left(\tilde{\nabla}_J h\right)_\alpha\left(\theta_P,\phi_P\right) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{2n+1}{4\pi}\frac{\tilde{\nabla}_J H_{\alpha,nm}^{(I,k)}}{\hat{i}_n(\kappa a^{(I,k)})}Y_{nm}^{(I,k)}(\theta_P,\phi_P) \qquad \alpha = x,y,z \tag{4.14a}$$

$$\left(\tilde{\nabla}_J f\right)_\alpha\left(\theta_P,\phi_P\right) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{2n+1}{4\pi}\tilde{\nabla}_J F_{\alpha,nm}^{(I,k)} Y_{nm}^{(I,k)}(\theta_P,\phi_P) \tag{4.14b}$$

Finally, with all local expansions computed, we can enter the innermost loop to solve for $\tilde{\nabla}_J\mathbf{F}^{(I,k)}$ $\tilde{\nabla}_J\mathbf{H}^{(I,k)}$ using Eqs. (4.11a-b) and (4.12a-b).

Since mutual polarization is a short-range effect, a cutoff r_{cut} can be used during charge and gradient polarization to simplify the computations. *Intermolecular* spheres whose surface-to-surface distances are greater than r_{cut} will not be included in each other's external field. The validity of cutoffs for intramolecular spheres needs to be further investigated.

Expressions for force and torque

The translational force on molecule I is given

$$\mathbf{f}_I = -\sum_{k}^{Nk^{(I)}}\mathbf{f}_{I,k} \tag{4.15}$$

where

$$\mathbf{f}_{I,k} = -\left\langle \tilde{\nabla}_I \mathbf{LHN}^{(I,k)}, \mathbf{H}^{(I,k)}\right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, \tilde{\nabla}_I \mathbf{H}^{(I,k)}\right\rangle \tag{4.16}$$

The torque on a charge at position P about the molecule I's center of mass $\boldsymbol{c}^{(I)}$ is given by the cross product of its position $\boldsymbol{r}_P^{(I)}$ with respect to $\boldsymbol{c}^{(I)}$ and the force it experienced, $\boldsymbol{f}_P$. The total torque on molecule I is then the sum of all torques:

$$\boldsymbol{\tau}_I = \sum_{k}^{Nk^{(I)}}\sum_{P\in k}\boldsymbol{\tau}_P = \sum_{k}^{Nk^{(I)}}\sum_{P\in k}\mathbf{r}_P^{(I)}\times\mathbf{f}_P \tag{4.17}$$

We can re-express $\boldsymbol{r}_P^{(I)}$ as the sum of vectors from center of molecule I to center of sphere k ($\boldsymbol{c}^{(I,k)}$), and from center of sphere k to point P ($\boldsymbol{r}_P^{(I,k)}$). The total torque about the center of molecule I is then:

$$\boldsymbol{\tau}_I = \sum_{k}^{Nk^{(I)}}\mathbf{c}^{(I,k)}\times\ \mathbf{f}_k\ + \sum_{k}^{Nk^{(I)}}\sum_{P\in k}\mathbf{r}_P^{(I,k)}\times\mathbf{f}_P \tag{4.18}$$

where

$$\mathbf{f}_P = -\sum_{k}^{Nk^{(I)}}\left\langle \tilde{\nabla}_I \mathbf{LHN}^{(I,k)}, \mathbf{H}_P^{(I,k)}\right\rangle + \left\langle \mathbf{LHN}^{(I,k)}, \tilde{\nabla}_I \mathbf{H}_P^{(I,k)}\right\rangle \tag{4.19}$$

and

$$\mathbf{H}_P^{(I,k)} = h(\theta_P,\phi_P)\mathbf{Y}_{nm}^{(I,k)}(\theta_P,\phi_P), \qquad \alpha = x,y,z \tag{4.20a}$$

$$\tilde{\nabla}_j\mathbf{H}_{P,\alpha}^{(I,k)} = \left(\tilde{\nabla}_j h(\theta_P,\phi_P)\right)_\alpha \mathbf{Y}_{nm}^{(I,k)}(\theta_P,\phi_P) \tag{4.20b}$$

Equations (4.15-4.16) and equations (4.18-4.20b) will be used to compute the force and torque respectively.

Brownian Dynamics

We have adopted the Brownian dynamics simulation protocol developed by Ermak and McCammon[29]. Each macromolecule *I* is treated as a Brownian particle experiencing a conservative force $\boldsymbol{f_I}$ and torque $\boldsymbol{\tau_I}$, in addition to hydrodynamic interactions with the solvent. The Langevin equation describing a system of *N* Brownian particles is given in [28] as

$$m_I \dot{v}_i = -\sum_j \xi_{ij} v_j + F_i + \sum_j \alpha_{ij} f_j \qquad (4.21)$$

where the index *I* runs over particles $1 \le I \le N$, and indices *i* and *j* ($1 \le i,j \le 3N$) run over *x*, *y*, and *z* particle coordinates, m_I is mass of particle *I*, F_i is the sum of systematic interparticle and external forces acting in direction *i*, v_i is the velocity in the direction *i*, ξ_{ij} is the configuration-dependent friction tensor between directions *i* and *j*. The sum $\sum_j \alpha_{ij} f_j$ represents the random fluctuating force exerted on particle *I*, the coefficients α are related to the friction tensor by $\xi_{ij} = \sum_l \alpha_{il}\alpha_{jl} / k_B T$. The above equation can be integrated twice to yield an equation for displacement *r*. If we further stipulate that our timestep *Δt* is much longer than $\tau_{ii}^0 = m_I / \xi_{ij} = m_I D_{ii}^0 / k_B T$, the relaxation time for velocity correlation for particle *I*, we obtain a displacement equation to propagate the Brownian dynamics [29]

$$r_i - r_i^0 = \sum_j \frac{\partial D_{ij}^0}{\partial r_j} \Delta t + \sum_j \frac{D_{ij}^0 F_j^0}{k_B T} \Delta t + S_i(\Delta t) \qquad (4.22)$$

where the random displacement *S* has the property

$$\left\langle S_i(\Delta t) \right\rangle = 0 \qquad (4.23a)$$

$$\left\langle S_i(\Delta t) S_j(\Delta t) \right\rangle = 2 D_{ij}^0 \Delta t \qquad (4.23b)$$

The hydrodynamic interactions between particles can be approximated using the Oseen and Rogne-Prager diffusion tensors.These two tensors have the property that $\sum_j \partial D_{ij}^0 / \partial r_j = 0$, so the first term on the right of Eq. (4.22) can be dropped. The Oseen tensor is given by[129]

$$\begin{aligned} D_{ij} &= \frac{k_B T}{c\pi\eta R_{hyd,I}} \delta_{ij}, \quad i, j \text{ on the same particle} \\ D_{ij} &= \frac{k_B T}{8\pi\eta R_{hyd,I}} \left(\overset{=}{I} + \frac{\mathbf{r}_{ij}\mathbf{r}_{ij}^T}{r_{ij}^2} \right), \quad i, j \text{ on the different particle} \end{aligned} \qquad (4.24)$$

while the Rotne-Prager tensor is given by[130]

$$D_{ij} = \frac{k_B T}{c\pi\eta R_{hyd,I}}\delta_{ij}, \quad i, j \text{ on the same particle}$$

$$D_{ij} = \frac{k_B T}{8\pi\eta R_{hyd,I}}\left(\left(\bar{\bar{I}} + \frac{\mathbf{r}_{ij}\mathbf{r}_{ij}^T}{r_{ij}^2}\right) + \frac{2R_{hyd,I}^2}{r_{ij}^2}\left(\frac{1}{3}\bar{\bar{I}} - \frac{\mathbf{r}_{ij}\mathbf{r}_{ij}^T}{r_{ij}^2}\right)\right), \quad i, j \text{ on the different particle} \tag{4.25}$$

where η is the viscosity of the solvent, c = 6 or 4 for stick or slip boundary condition respectively, and $R_{hyd,I}$ is the hydrodynamic radius of molecule I, and $\boldsymbol{r}_{ij}$ is the vector connecting the center of particles associated with indices i and j.

Assuming no hydrodynamic interaction between the macromolecules, the displacement $\Delta\boldsymbol{r}_I$ and angular rotation $\Delta\boldsymbol{\vartheta}_I$ per timestep Δt are given by

$$\Delta\mathbf{r}_I = \frac{D_{I,trans}\Delta t}{k_B T}\mathbf{f}_I + \mathbf{S}_I(\Delta t) \tag{4.26a}$$

$$\Delta\vartheta_I = \frac{D_{I,rot}\Delta t}{k_B T}\boldsymbol{\tau}_I + \Theta_I(\Delta t) \tag{4.26b}$$

where the stochastic displacement (S) and rotation (Θ) have the properties

$$\langle S_\alpha \rangle = 0, \qquad \langle {S_\alpha}^2 \rangle = 2D_{I,trans}\Delta t \tag{4.27a}$$

$$\langle \Theta_\alpha \rangle = 0, \qquad \langle {\Theta_\alpha}^2 \rangle = 2D_{I,rot}\Delta t \qquad \alpha = \mathrm{x,y,z} \tag{4.27b}$$

The translational and rotational diffusion constants of each molecule I are given by

$$D_{I,trans} = \frac{k_B T}{c\pi\eta R_{hyd,I}} \tag{4.28a}$$

$$D_{I,rot} = \frac{k_B T}{8\pi\eta {R_{hyd,I}}^3} \tag{4.28b}$$

Model and Parameters

We consider a system of two brome mosaic virus (PDB: 1YC6[4]) monomers whose dynamics are simulated using Eqs. (4.26a-b) under the PB-SAM model. The first monomer ("M1") is fixed at the origin, while the second monomer (M2) is positioned 100Å away, and allowed to move with respect to M1 (Figure 4.1). We represent each monomer's dielectric boundary using different number of spheres: 107 spheres ('107S'), 354 spheres ('354S'), and 712 spheres ('712S') (see Figure 3.3). We note that this same system would have been represented on a 300Å x 300Å x 300Å grid, which translates to a 0.6Å resolution, in a finite difference PBE solver. Thus the accuracy of our PB-SAM solution is far superior.

During the initialization step, the system is solved with full polarization ($r_{cut} = \infty$). The mobile monomer M2 is then propagated for one time step, and the system solved again, with r_{cut} = 50Å. Convergence is reached when the relative changes in $\boldsymbol{H}^{(I,k)}$ and $\tilde{\nabla}_I H_{nm}^{(I,k)}$ fall below 10^{-2} for all spheres. For each representation, we solve for the multipole and gradient polarization at maximum pole order p = 10 and p = 20. The system temperature was set to 298.15 K. The dielectric constants were ε_s = 78 and ε_p = 4. The inverse Debye length κ = 0.07374, corresponding to a salt concentration of 0.05M. The viscosity of water is 0.001002 kg m^{-1} s^{-1} at 20°C, and we use c = 6 corresponding to stick boundary conditions. The hydrodynamic radius of 1YC6 is set to 33Å, based on an equivalent sphere with identical volume. A timestep of 10 ps is used. All computations were performed on a single processor on an Intel(R) Xeon(R) CPU 2.27GHz processor with 24GB of physical memory.

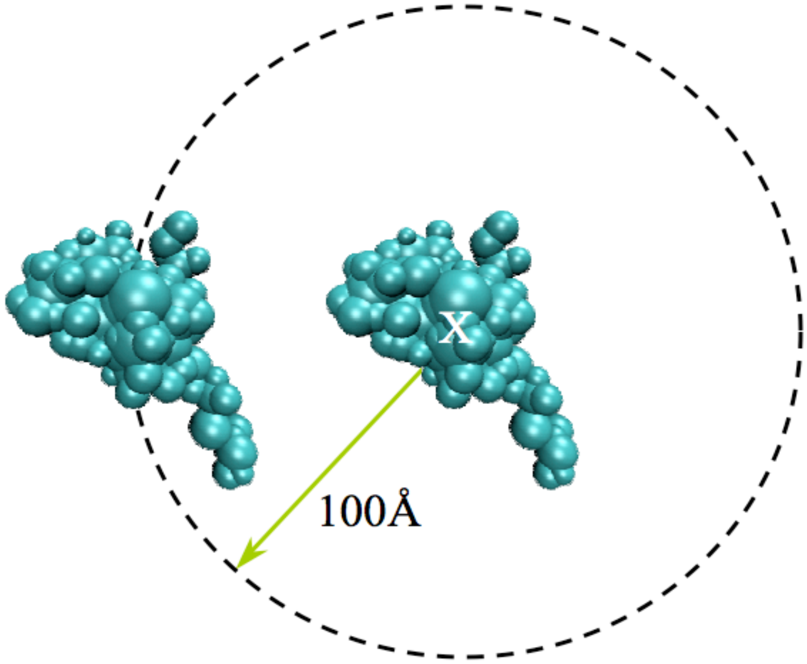

Figure 4.1 ***Starting configuration for force and torque calculations.*** Figure shows two 1YC6 monomers using 107S representations. Forces and torques were also computed using representations with 354 and 712 spheres. Figure is not to scale.

Results and Discussion

The total time taken for force and torque computation at each timestep is presented in Table 4.1. This time includes the solution of (i) mutually polarized multipoles $\boldsymbol{F}^{(I,k)}$ and $\boldsymbol{H}^{(I,k)}$ for each sphere, and (ii) corresponding gradient for each sphere with respect to M1 and M2; (iii) computing forces and torques according to Eqs. (4.15-4.16) and Eqs. (4.18-4.20b). The timing for the first step (which includes one time cost of initialization) is presented in italic while the timing for the subsequent step is in parentheses.

Table 4.1. Timings for force and torque computation. Timings presented for three boundary representations (107S, 354S, 712S), and two pole orders. Timings for the initializing step are in italic; timings for the subsequent step are in parentheses.

	Timing [hours]		
Pole Order	107S	354S	712S
10	*0.02* (0.01)	*0.17(0.06)*	*0.34 ()*
20	*0.17* (0.10)	*1.18* (0.87)	*2.02* (2.14)

The initializing step involves solving the mutual polarization quantities from a 'cold-start', using initial guesses from self-polarized values. Consequently they are generally longer than the subsequent step, although the difference diminishes as the number of spheres *Ns* increases. A possible reason for this trend is that, as *Ns* increases, it becomes harder to obtain a converged solution within the stipulated number of iterations during the initialization step. Therefore, subsequent steps have to continue solving for the quantities using imperfect initial guesses, using up comparable computation time.

The timing scales with $O(p^{3 \sim 4})$. This behavior is dominated by two routines that scale with p^4: (i) matrix-vector multiply operations in the innermost iterative loops of charge and gradient polarizations, and (ii) numerical re-expansions of surface charges and gradients. Future efforts to improve scaling with pole order must address these two areas. The timing scales approximately linear with *Ns*, the number of spheres. This is encouraging because this gives us some freedom to select the appropriate resolution without incurring excessive computational cost.

While PB-SAM enable us to model mutual polarization effects in systems of hitherto inaccessible spatial dimensions, the current computation time per step could be further improved in order to perform multiple trajectories for longer durations to collect association statistics: firstly, the computation is currently performed on a single processor, and can be trivially modified to run on 8-processor shared memory platform to given a speed-up factor of 8. Secondly, we can improve the convergence of the innermost iterative loop through pre-conditioning, and replace the repeated matrix-vector multiply subroutine calls with a fit-for-purpose linearized equation solution routine such as Generalized minimal residual method (GMRES). Thirdly, more studies need to be done to determine how much we can relax the convergence criteria while maintaining a stipulated level of accuracy. In addition, the choice of cutoff distance, r_{cut}, for inter- and intramolecular polarizations can be further optimized. Lastly, we observe that the gradient calculation step constitutes ¾ of the total polarization time. A promising approach to reducing the computation costs while including some aspect of gradient polarization could involve approximating the polarized gradients from analytical calculation using spherical dielectric boundaries.

Conclusions

We have derived the formalism for force and torque calculation in the context of our new Poisson Boltzmann solution algorithm, PB-SAM, and incorporated within the framework of a Brownian dynamics simulation algorithm. The formulism accounts for mutual polarization in both the zero and first order derivative of the surface charges. We demonstrated for the first time dynamic propagation of Brownian particles with accurate accounting of mutual polarization effects for successive timesteps, using a model system of two monomers of brome mosaic virus (1YC6), with resolutions ranging from a spatial coarsening at the residue-level to atomistic resolution. The time taken to perform the initialization step and subsequent step are collected, and future strategies for algorithm accelerations were proposed.

Future simulations of PB-SAM could be extended to address periodic boundary condition. For a simulation box of length *l*, the force attenuates rapidly with $exp(-\kappa R)/R^2$ at physiological salt concentrations (*n* ~ 0.05M), so a simple truncation of the Yukawa potential with *l*/2, along with minimum image convention, adequately addresses the periodic boundary condition. At lower salt concentrations where the forces decay more slowly, long range contributions would have to be included using an Ewald sum for Yukawa potential[131, 132].

Chapter 5

Strategies for Multiscale Simulations

Introduction

Many processes in nature are inherently multiscale, spanning the microscopic to mesoscopic and macroscopic lengthscales. Coupling simulation models of different length- and timescales will allow us to study interesting phenomena with sufficient *breadth* (long timescales, macroscopic lengthscales, reliable statistics), without sacrificing *depth* (atomistic details).

The term 'multiscale' typically refers to strategies for connecting different spatial resolutions such as all atomic (AA) to coarse-grained (CG) representations of materials. A broader interpretation of multiscale strategies could also include connecting different temporal resolutions such as strategies to couple continuous, dynamic simulations with discrete event simulations or replacement of basic quantum mechanically motivated interactions in terms of effective interactions. Multiscale methods can be roughly categorized into parallel and serial strategies[133]. Below we provide examples of methods in each category, and discuss approaches relevant to studies of protein-protein association.

Parallel Multiscale Strategies

In parallel multi-scaling approaches, simulations using models of different resolutions are carried out concurrently. Information is constantly exchanged between the different models in real time. In the 'mixed resolution' approach, a selected region (or molecular species) is simulated in atomistic resolution, while the other regions (or species) are simulated using coarse-grained models. This approach is analogous to hybrid quantum mechanics / molecular mechanics (QM/MM) methods. Mixed resolution simulations have been used to study membrane-bound ion channels by coarse graining the lipid and water molecules while using an all-atom representation for the polypeptide ion channel[134].

In another approach, known as 'resolution exchange' or 'model swapping'[135], replicas of a system are evolved concurrently using different resolution models, and exchanges are attempted at regular time intervals. The approach is analogous to replica exchange or parallel tempering, and allows movement between different levels of structural detail in order to cross energy barriers, in. One challenge in using this method is the ability to regenerate realistic atomistic details from coarse-grained models. Recent effort in this area[136] has introduced an efficient and reliable algorithm to generate all-atom details from alpha-carbon only protein models.

Serial Multiscale Strategies

In serial multiscale approaches, also known as field theoretic approaches, different resolution models are employed in sequence, so there is no real-time coupling between the simulations. Instead, information is transferred in a bottom-up approach, such that emergent parameters (e.g. diffusion constants, transmission coefficients) are extracted from simulations at a finer resolution, and used as input parameters for the coarser model.

Force-matching [137] is a technique to obtain classical force fields from trajectory and force databases produced by *ab initio* MD simulations. The force-matching procedure includes a fit of short-ranged nonbonded forces, bonded forces, and atomic partial charges from ab initio MD and MD simulations. The technique has been applied to parameterize coarse-grained water and protein forcefields [138, 139].

In another example, hydrodynamic parameters were extracted from atomistic MD to model coarse-grained dynamics. Coarse-graining introduces spurious, accelerated diffusional behavior because the fluctuating forces associated with missing molecular degrees of freedom are eliminated. To correct this, one can approximate the frictional constant from the instantaneous difference between the mean CG forces and the exact all-atom MD forces[140], and use the frictional constant to model Langevin dynamics in the CG model.

Multiscale Strategies for Protein-Protein Kinetics

Multiscale algorithms have been used on static protein docking problems[141]. Below we discuss ways in which multiscale strategies can be applied to investigate kinetics of large-scale protein assembly kinetics.

a) Built-in Multiscale Forcefields

Conventional multiscale strategies approach the challenge from the angle of spatial coarse-graining. Alternatively, one can also think about how multi-scaling can be implicitly built into force-fields. An illustrative example is the fast multipole method, in which well-separated objects are automatically collectivized into coarser resolutions and less information (pole) propagated back to the local field. In our PB-SAM methodology, we employ a similar philosophy, adaptively using the minimum pole order necessary to perform interactions with entities at different separation lengthscales.

b) Nested Northrup-Allison-McCammon (NAM) method for bi-molecular kinetics

Bimolecular association rates can be obtained from Brownian dynamic simulations using the Northrup-Allison-McCammon ('NAM') formulism[5]. Here we propose a procedure to convert the basic NAM methodology into a multiscale framework using the two coarse-grained models that we developed.

In the basic NAM method, the first molecule ('A') is positioned at the origin and the second molecule ('B') randomly on a sphere with $r = b$, chosen such that the potential of mean force and reactive flux of B are centrosymmetric for $r \geq b$. Molecule B is then evolved in time until it either satisfy the collision criteria for encounter complex or escape to distance q ($q > b$). The above is repeated for 1000 or more trajectories to obtain the collision frequency δ. The intrinsic association rate k can then evaluated with

$$k = \frac{k(b)\delta}{1-(1-\delta)k(b)/k(q)} \tag{5.1}$$

where *k(b)* and *k(q)* are rates at which a molecule B starting infinity reaches $r = b$ and $r = q$ respectively, evaluated analytically from the Smouluchowski rate equation

$$k(R) = \left[\frac{1}{4\pi(D_A + D_B)}\int_R^\infty \frac{\exp(q_1 q_2 / k_B T r)}{r^2} dr\right]^{-1} \tag{5.2}$$

To extend the NAM formulism to a multi-scale framework, the simulation will be divided into overlapping regimes R1 and R2 that uses the *protein level* and *residue level* models respectively (see Figure 5.1). In regime R1, defined as $s < r < q$, the protein level model is employed, treating proteins as rigid bodies that interact through electrostatics only. Simulations begin with B at $r = b$, and stopping at either the $r = s$ boundary ('collided') or $r = q$ boundary ('escaped'). The collision frequency thus collected, δ_1, can then be used to evaluate

$$k(s) = \frac{k(b)\delta_1}{1-(1-\delta_1)k(b)/k(q)} \tag{5.3}$$

A separate set of simulations is performed for regime R2 ($r < b$) using the residue level model, starting with B at $r = s$ and stopping when the encounter complex is formed or at $r = b$, to collect collision frequency δ_2. The overall intrinsic association rate is then

$$k = \frac{k(s)\delta_2}{1-(1-\delta_2)k(s)/k(b)} \tag{5.4}$$

The method can in theory accommodate multiple simulation shells, each using a different resolution model.

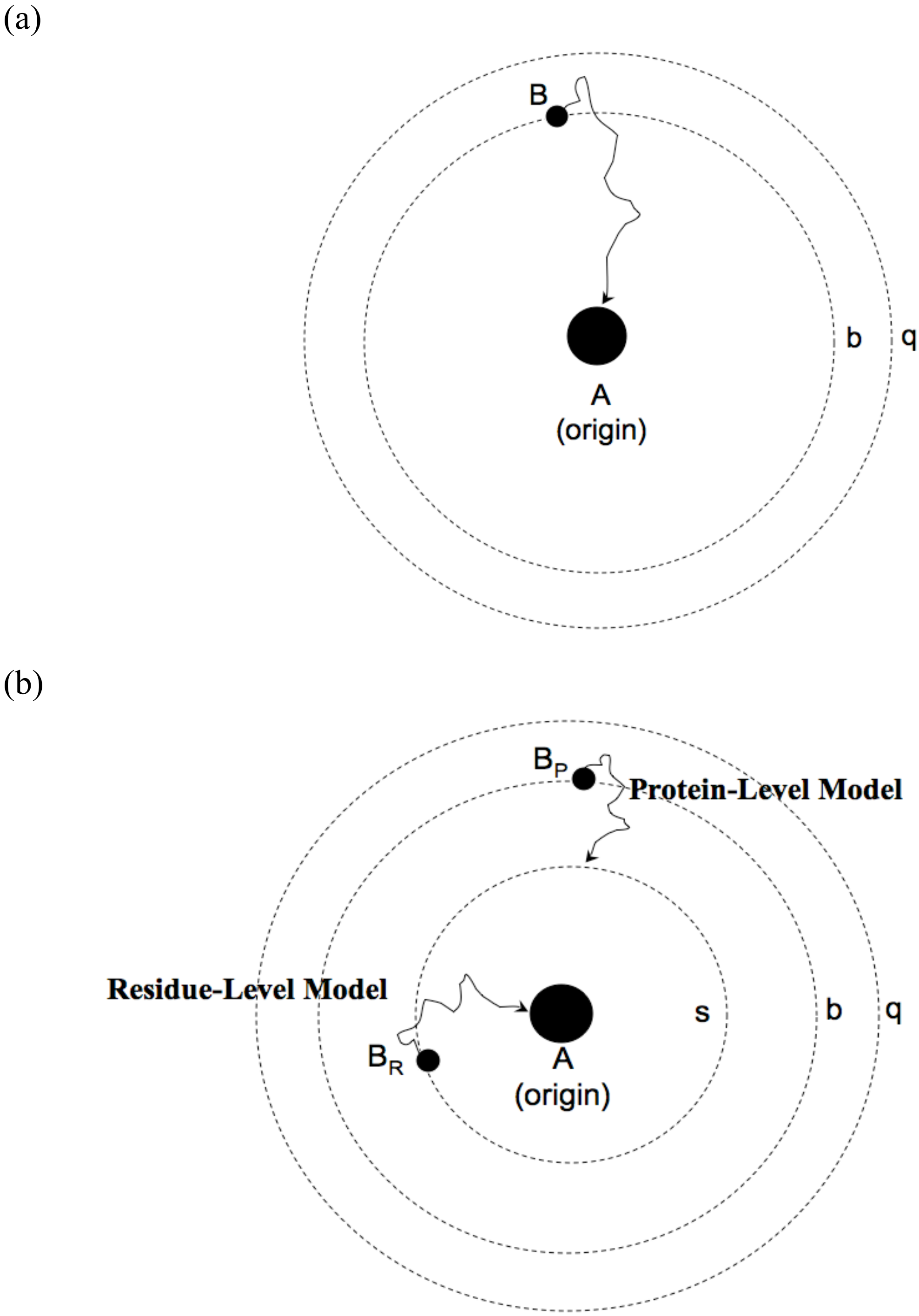

Figure 5.1. ***Simulation procedure for association rate calculations.*** (a) Original NAM method (b) Multi-scale extension of NAM method. The protein-level model will be used for to simulate trajectories starting at *b* and ending at either *s* or *q*. The residue-level model will be used to simulate trajectories starting at s and ending at either encounter complex formation or *b*.

c) Continuous Dynamic Simulations as inputs to Chemical Master Equation

Coarse-grained simulations, such as the binary association presented in (b), can provide us with rate constants that can be in turn plugged into chemical master equations (CME). For a system with *N* species $\{S_1, \ldots, S_N\}$, interacting through *M* number of reaction channels, the CME describes the evolution of the state vector $\boldsymbol{x} = \boldsymbol{X(t)} = \{X_1(t), \ldots, X_N(t)\}$ [142]

$$\frac{\partial P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)}{\partial t} = \sum_{j=1}^{M}\left[a_j(\mathbf{x}-\mathbf{v}_j)P(\mathbf{x}-\mathbf{v}_j,t\,|\,\mathbf{x}_0,t_0) - a_j(\mathbf{x})P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)\right] \tag{5.5}$$

where $P(\boldsymbol{x},t|\boldsymbol{x}_0,t_0)$ is the conditional probability that the system is in state $\boldsymbol{x}$ at time *t*, given that it is at state $\boldsymbol{x}_0$ at time t_0, $\boldsymbol{v}_j$ is the state change vector describing the changes in species populations associated with reaction *j*; and $a_j(\boldsymbol{x})$ is the propensity for reaction *j* to occur given the system is in state $\boldsymbol{x}$.

Solutions of CME will provide us with detailed information about how the concentration of each species varies with time, from which a dominant assembly pathway can be determined. Sept and McCammon[143] investigated the nucleation pathway of actin by first characterizing the association and dissociation rate constants for all possible pairwise associations using Brownian dynamics. The kinetic parameters were inputted into a chemical master equation for nucleation–elongation. The CME was solved to obtain the time course of polymerization and identify the dominant nucleation pathway.

d) Continuous Dynamic Simulations as inputs to Discrete-Event Simulations

While the chemical master equation provides full detail of the time course of each species, its direct solution is only tractable for low dimensions (number of species < ~10). For systems of higher dimensions, the kinetics can be studied by generating trajectories based on the underlying Markov process. This is the basis of the Gillespie's stochastic simulation algorithm[144]. In such cases, the continuous dynamic simulations using residue level and/or the protein level model(s) can provide kinetic rates, or transition probability in Markovian terminology, to direct the stochastic simulations. Chodera et. al. described the long-time statistical dynamics of solvated terminally blocked alanine peptide using a discrete-state Markov chain model constructed from short MD trajectories[145, 146]. In another example, Hemberg et. al. performed stochastic kinetics to simulate the viral capsid assembly using an updated Gillespie algorithm that is modified for heterogeneous solvent conditions[147].

Conclusions

This thesis work focuses on the theory and algorithm development of coarse-grained implicit solvent models that could be deployed within a multiscale framework to enable computational studies of large-scale protein associations. A multiscale coarse-graining approach is ideal for such studies, because different stages of the association process fall naturally into different time and length-scales regimes. At very short separation distance (e.g. during docking), all atom molecular dynamics is necessary to model side chain packing and capture short-range interactions such as van de Waal forces, hydrogen bonding. As we move to intermediate separation distances characteristic of encounter complexes (separation by one to two water layers), a residue level model propagated by Langevin dynamics is sufficient to account for backbone conformational fluctuations and hydrophobic interactions. At even greater separation distances, conformational fluctuations become insignificant, so proteins can be represented as a rigid bodies moving according to Brownian dynamics, and only long range electrostatic interaction persists.

The residue level $\alpha-$carbon model presented in chapter 2 incorporates a novel forcefield term to model directional backbone hydrogen bond, leading to more stable and realistic $\alpha-$helices and $\beta-$sheets. In addition, the addition of a fourth bead flavor provides a more graded spectrum of attractive interaction energies that better reflect the hydrophobicity range of the 20 naturally occurring amino acids, reducing energetic frustrations and competition from misfolded states. The model retains a strong connection between sequence and folding mechanism for proteins L and G, shows increased folding cooperativity, and a greater structural faithfulness to experimentally solved structures. The computational efficiency of the model has also permitted us to develop molecular models of the Alzheimer's $A\beta_{1-40}$ fibril in order to determine the critical nucleus, stability with chain size, and fibril elongation[1, 2], providing a good proof-of-concept and setting the foundation for applications to other protein-protein assembly processes.

Chapters 3 and 4 describe the development of a protein level model to simulate proteins during diffusional search. Chapter 3 focuses on the theory and implementation of a new approach, Poisson-Boltzmann Semi-Analytical Method (PB-SAM) to model electrostatic interactions by efficiently solving the linearized PBE. This method represents the macromolecular surface as a collection of overlapping spheres, for which polarization charges can then be iteratively solved using analytical multipole method[3]. Unlike finite difference solvers, PB-SAM is not constrained spatially by the box size, making it suitable for dynamics. This method realizes better accuracy at reduced cost relative to either finite difference or boundary element PBE solvers.

We then incorporated the PB-SAM solver into a protein level Brownian dynamics simulation algorithm (chapter 4). We derived the formalism for force and torque calculation that account for mutual polarization in both the zero and first order derivative of the surface charges, and demonstrated for the first time dynamic propagation of multiple Brownian particles with accurate accounting of mutual polarization effects for successive timesteps, using a model system of two monomers of brome mosaic virus (1YC6). While PB-SAM enable us to model mutual polarization effects in systems of hitherto inaccessible spatial dimensions, the current computation time per step can be further improved through parallelization, pre-conditioning and more efficient solution of the iterative equations, careful choices of convergence criteria and polarization cutoff distances, and approximating mutual polarization effects from cheap analytical calculation based on spherical dielectric boundaries.

Lastly, we discussed in chapter 5 multiscale strategies to connect the two models described above for large-scale protein assembly studies. The two models can be employed successively in a novel nested variant of the Northrup-Allison-McCammon[5] formalism to compute bi-molecular kinetics rates. These rates will in turn be inputs to chemical master equations or, for more complex systems, stochastic simulations such as Markov chains and the Gillespie algorithm[144]. The framework can be applied to study the role of protein-protein interactions in recruitment of adapter and binding proteins in signal transduction to form organized scaffolds known as “signalosomes”[148] and virus capsid assembly for drug delivery research[149-153]. Multiscale simulations using the two coarse-grained models can be performed to determine kinetics rates and the order of association, and help investigate how modifying the physical interactions (e.g. through mutation or changing solution conditions) can alter the association rates, and consequently modify the overall sequences of association.

Bibliography

1. Fawzi, N. L.; Okabe, Y.; Yap, E. H.; Head-Gordon, T., Determining the critical nucleus and mechanism of fibril elongation of the Alzheimer's A-beta(1-40) peptide. *Journal of Molecular Biology* **2007,** 365, (2), 535-550.
2. Fawzi, N. L.; Yap, E. H.; Okabe, Y.; Kohlstedt, K. L.; Brown, S. P.; Head-Gordon, T., Contrasting disease and nondisease protein aggregation by molecular simulation. *Accounts of Chemical Research* **2008,** 41, (8), 1037-1047.
3. Lotan, I.; Head-Gordon, T., An analytical electrostatic model for salt screened interactions between multiple proteins. *Journal of Chemical Theory and Computation* **2006,** 2, (3), 541-555.
4. Lucas, R. W.; Kuznetsov, Y. G.; Larson, S. B.; McPherson, A., Crystallization of brome mosaic virus and T=1 brome mosaic virus particles following a structural transition. *Virology* **2001,** 286, (2), 290-303.
5. Northrup, S. H.; Allison, S. A.; McCammon, J. A., Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions. *Journal of Chemical Physics* **1984,** 80, (4), 1517-1526.
6. Socci, N. D.; Onuchic, J. N., Folding Kinetics of Proteinlike Heteropolymers. *Journal of Chemical Physics* **1994,** 101, (2), 1519-1528.
7. Werlen, G.; Palmer, E., The TCR signalosome: a dynamic structure with expanding complexity. *Current Opinion in Immunology* **2002,** 14, (3), 299-305.
8. Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** 98, (8), 4569-4574.
9. Cleland, W. W., Enzyme Kinetics. *Annual Review of Biochemistry* **1967,** 36, 77-&.
10. Schreiber, G., Kinetic studies of protein-protein interactions. *Current Opinion in Structural Biology* **2002,** 12, (1), 41-47.
11. Schreiber, G.; Fersht, A. R., Energetics of Protein-Protein Interactions - Analysis of the Barnase-Barstar Interface by Single Mutations and Double Mutant Cycles. *Journal of Molecular Biology* **1995,** 248, (2), 478-486.
12. Abagyan, R.; Totrov, M., High-throughput docking for lead generation. *Current Opinion in Chemical Biology* **2001,** 5, (4), 375-382.
13. Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D., Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology* **2003,** 331, (1), 281-299.
14. Smith, G. R.; Sternberg, M. J. E., Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology* **2002,** 12, (1), 28-35.
15. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005,** 26, (16), 1668-1688.

16. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry* **2009,** 30, (10), 1545-1614.

17. Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Van Gunsteren, W. F., The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry* **2005,** 26, (16), 1719-1751.

18. Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K., Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **2006,** 14, (3), 437-449.

19. Klein, M. L.; Shinoda, W., Large-scale molecular dynamics simulations of self-assembling systems. *Science* **2008,** 321, (5890), 798-800.

20. Sanbonmatsu, K. Y.; Tung, C. S., High performance computing in biology: Multimillion atom simulations of nanoscale systems. *Journal of Structural Biology* **2007,** 157, (3), 470-480.

21. Earl, D. J.; Deem, M. W., Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* **2005,** 7, (23), 3910-3916.

22. Bussi, G.; Laio, A.; Parrinello, M., Equilibrium free energies from nonequilibrium metadynamics. *Physical Review Letters* **2006,** 96, (9), 1558-1568.

23. Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D., Transition path sampling and the calculation of rate constants. *Journal of Chemical Physics* **1998,** 108, (5), 1964-1977.

24. E, W. N.; Ren, W. Q.; Vanden-Eijnden, E., String method for the study of rare events. *Physical Review B* **2002,** 66, (5), 052301.

25. Henkelman, G.; Uberuaga, B. P.; Jonsson, H., A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *Journal of Chemical Physics* **2000,** 113, (22), 9901-9904.

26. Langevin, P., The theory of brownian movement. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences* **1908,** 146, 530-533.

27. Lemons, D. S.; Gythiel, a., Paul Langevin's 1908 paper "On the theory of Brownian motion". *American Journal of Physics* **1997,** 65, (11), 1079-1081.

28. Deutch, J. M.; Oppenhei.I, Molecular Theory of Brownian Motion for Several Particles. *Journal of Chemical Physics* **1971,** 54, (8), 3547-&.

29. Ermak, D. L.; Mccammon, J. A., Brownian Dynamics with Hydrodynamic Interactions. *Journal of Chemical Physics* **1978,** 69, (4), 1352-1360.

30. Privalov, P. L.; Gill, S. J., Stability of Protein-Structure and Hydrophobic Interaction. *Advances in Protein Chemistry* **1988,** 39, 191-234.

31. Geiger, A.; Stillinger, F. H.; Rahman, A., Aspects of the Percolation Process for Hydrogen-Bond Networks in Water. *Journal of Chemical Physics* **1979,** 70, (9), 4185-4193.

32. Stillinger, F. H.; David, C. W., Study of the Water Octamer Using the Polarization Model of Molecular-Interactions. *Journal of Chemical Physics* **1980,** 73, (7), 3384-3389.

33. Dill, K. A., Dominant Forces in Protein Folding. *Biochemistry* **1990,** 29, (31), 7133-7155.

34. Spolar, R. S.; Ha, J. H.; Record, M. T., Hydrophobic Effect in Protein Folding and Other Noncovalent Processes Involving Proteins. *Proceedings of the National Academy of Sciences of the United States of America* **1989,** 86, (21), 8382-8385.

35. Eisenberg, D.; Mclachlan, A. D., Solvation Energy in Protein Folding and Binding. *Nature* **1986,** 319, (6050), 199-203.

36. Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B., Reconciling the Magnitude of the Microscopic and Macroscopic Hydrophobic Effects. *Science* **1991,** 252, (5002), 106-109.

37. Chothia, C., Hydrophobic Bonding and Accessible Surface-Area in Proteins. *Nature* **1974,** 248, (5446), 338-339.

38. Chandler, D., Interfaces and the driving force of hydrophobic assembly. *Nature* **2005,** 437, (7059), 640-647.

39. Lin, M. S.; Fawzi, N. L.; Head-Gordon, T., Hydrophobic potential of mean force as a solvation function for protein structure prediction. *Structure* **2007,** 15, (6), 727-740.

40. Sharp, K. A., Electrostatic Interactions in Macromolecules. *Current Opinion in Structural Biology* **1994,** 4, (2), 234-239.

41. Murphy, W. F., The Rayleigh depolarization ratio and rotational Raman spectrum of water vapor and the polarizability components for the water molecule. *Journal of Chemical Physics* **1977,** 67, 434794-434800.

42. Takashim.S; Schwan, H. P., Dielectric Dispersion of Crystalline Powders of Amino Acids Peptides and Proteins. *Journal of Physical Chemistry* **1965,** 69, (12), 4176-&.

43. Nicholls, A.; Honig, B., A Rapid Finite-Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson-Boltzmann Equation. *Journal of Computational Chemistry* **1991,** 12, (4), 435-445.

44. Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A., Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** 98, (18), 10037-10041.

45. Bordner, A. J.; Huber, G. A., Boundary element solution of the linear Poisson-Boltzmann equation and a multipole method for the rapid calculation of forces on macromolecules in solution. *Journal of Computational Chemistry* **2003,** 24, (3), 353-367.

46. Boschitsch, A. H.; Fenley, M. O.; Zhou, H. X., Fast boundary element method for the linear Poisson-Boltzmann equation. *Journal of Physical Chemistry B* **2002,** 106, (10), 2741-2754.

47. Juffer, A. H.; Botta, E. F. F.; Vankeulen, B. A. M.; Vanderploeg, A.; Berendsen, H. J. C., The Electric-Potential of a Macromolecule in a Solvent - a Fundamental Approach. *Journal of Computational Physics* **1991,** 97, (1), 144-171.

48. Lu, B. Z.; Cheng, X. L.; Huang, J. F.; McCammon, J. A., An Adaptive Fast Multipole Boundary Element Method for Poisson-Boltzmann Electrostatics. *Journal of Chemical Theory and Computation* **2009,** 5, (6), 1692-1699.

49. Lu, B. Z.; McCammon, J. A., Improved boundary element methods for Poisson-Boltzmann electrostatic potential and force calculations. *Journal of Chemical Theory and Computation* **2007,** 3, (3), 1134-1142.

50. Zauhar, R. J.; Morgan, R. S., A New Method for Computing the Macromolecular Electric-Potential. *Journal of Molecular Biology* **1985,** 186, (4), 815-820.

51. Zhou, H. X., Boundary-Element Solution of Macromolecular Electrostatics - Interaction Energy between 2 Proteins. *Biophysical Journal* **1993,** 65, (2), 955-963.

52. Camacho, C. J.; Weng, Z. P.; Vajda, S.; DeLisi, C., Free energy landscapes of encounter complexes in protein-protein association. *Biophysical Journal* **1999,** 76, (3), 1166-1178.

53. Sheinerman, F. B.; Norel, R.; Honig, B., Electrostatic aspects of protein-protein interactions. *Current Opinion in Structural Biology* **2000,** 10, (2), 153-159.

54. GO, N., Theoretical-Studies of Protein Folding. *Annual Review of Biophysics and Bioengineering* **1983,** 12, 183-210.

55. Plaxco, K. W.; Simons, K. T.; Baker, D., Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of Molecular Biology* **1998,** 277, (4), 985-994.

56. Brown, S.; Head-Gordon, T., Intermediates and the folding of proteins L and G. *Protein Science* **2004,** 13, (4), 958-970.

57. Brown, S.; Fawzi, N. J.; Head-Gordon, T., Coarse-grained sequences for protein folding and design. *Proceedings of the National Academy of Sciences of the United States of America* **2003,** 100, (19), 10712-10717.

58. Yap, E. H.; Fawzi, N. L.; Head-Gordon, T., A coarse-grained alpha-carbon protein model with anisotropic hydrogen-bonding. *Proteins-Structure Function and Bioinformatics* **2008,** 70, (3), 626-638.

59. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G., Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis. *Proteins-Structure Function and Genetics* **1995,** 21, (3), 167-195.

60. Cheung, M. S.; Finke, J. M.; Callahan, B.; Onuchic, J. N., Exploring the interplay between topology and secondary structural formation in the protein folding problem. *Journal of Physical Chemistry B* **2003,** 107, (40), 11193-11200.

61. Onuchic, J. N.; LutheySchulten, Z.; Wolynes, P. G., Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry* **1997,** 48, 545-600.

62. Guo, Z.; Thirumalai, D., Kinetics and thermodynamics of folding of a de Novo designed four-helix bundle protein. *Journal of Molecular Biology* **1996,** 263, (2), 323-343.

63. Guo, Z. Y.; Thirumalai, D., Kinetics of Protein-Folding - Nucleation Mechanism, Time Scales, and Pathways. *Biopolymers* **1995,** 36, (1), 83-102.

64. Guo, Z. Y.; Thirumalai, D.; Honeycutt, J. D., Folding Kinetics of Proteins - a Model Study. *Journal of Chemical Physics* **1992,** 97, (1), 525-535.

65. Honeycutt, J. D.; Thirumalai, D., Metastability of the Folded States of Globular-Proteins. *Proceedings of the National Academy of Sciences of the United States of America* **1990,** 87, (9), 3526-3529.

66. Sorensen, J. M.; Head-Gordon, T., Toward minimalist models of larger proteins: A ubiquitin-like protein. *Proteins-Structure Function and Genetics* **2002,** 46, (4), 368-379.

67. Sorenson, J. M.; Head-Gordon, T., Redesigning the hydrophobic core of a model beta-sheet protein: Destabilizing traps through a threading approach. *Proteins-Structure Function and Genetics* **1999,** 37, (4), 582-591.

68. Sorenson, J. M.; Head-Gordon, T., Matching simulation and experiment: A new simplified model for simulating protein folding. *Journal of Computational Biology* **2000,** 7, (3-4), 469-481.

69. Sorenson, J. M.; Head-Gordon, T., Protein engineering study of protein L by simulation. *Journal of Computational Biology* **2002,** 9, (1), 35-54.

70. Fawzi, N. L.; Chubukov, V.; Clark, L. A.; Brown, S.; Head-Gordon, T., Influence of denatured and intermediate states of folding on protein aggregation. *Protein Science* **2005,** 14, (4), 993-1003.

71. Dobson, C. M., Principles of protein folding, misfolding and aggregation. *Seminars in Cell & Developmental Biology* **2004,** 15, (1), 3-16.

72. Buchete, N. V.; Straub, J. E.; Thirumalai, D., Orientational potentials extracted from protein structures improve native fold recognition. *Protein Science* **2004,** 13, (4), 862-874.

73. Das, P.; Matysiak, S.; Clementi, C., Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** 102, (29), 10141-10146.

74. Klimov, D. K.; Betancourt, M. R.; Thirumalai, D., Virtual atom representation of hydrogen bonds in minimal off-lattice models of alpha helices: effect on stability, cooperativity and kinetics. *Folding & Design* **1998,** 3, (6), 481-496.

75. Klimov, D. K.; Thirumalai, D., Mechanisms and kinetics of beta-hairpin formation. *Proceedings of the National Academy of Sciences of the United States of America* **2000,** 97, (6), 2544-2549.

76. Smith, A. V.; Hall, C. K., Alpha-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins-Structure Function and Genetics* **2001,** 44, (3), 344-360.

77. Onuchic, J. N.; Wolynes, P. G.; Lutheyschulten, Z.; Socci, N. D., Toward an Outline of the Topography of a Realistic Protein-Folding Funnel. *Proceedings of the National Academy of Sciences of the United States of America* **1995,** 92, (8), 3626-3630.

78. Marcus, Y.; Ben-Naim, A., A Study of the Structure of Water and Its Dependence on Solutes, Based on the Isotope Effects on Solvation Thermodynamics in Water. *Journal of Chemical Physics* **1985,** 83, (9), 4744-4759.

79. Silverstein, K. A. T.; Haymet, A. D. J.; Dill, K. A., A simple model of water and the hydrophobic effect. *Journal of the American Chemical Society* **1998,** 120, (13), 3166-3175.

80. Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M., A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein-G. *Science* **1991,** 253, (5020), 657-661.

81. Wikstrom, M.; Drakenberg, T.; Forsen, S.; Sjobring, U.; Bjorck, L., 3-Dimensional Solution Structure of an Immunoglobulin Light Chain-Binding Domain of Protein-L - Comparison with the Igg-Binding Domains of Protein-G. *Biochemistry* **1994,** 33, (47), 14011-14017.

82. Alexander, P.; Fahnestock, S.; Lee, T.; Orban, J.; Bryan, P., Thermodynamic Analysis of the Folding of the Streptococcal Protein-G Igg-Binding Domains B1 and B2 - Why Small

Proteins Tend to Have High Denaturation Temperatures. *Biochemistry* **1992,** 31, (14), 3597-3603.

83. Alexander, P.; Orban, J.; Bryan, P., Kinetic-Analysis of Folding and Unfolding the 56-Amino Acid Igg-Binding Domain of Streptococcal Protein-G. *Biochemistry* **1992,** 31, (32), 7243-7248.

84. Kim, D. E.; Fisher, C.; Baker, D., A breakdown of symmetry in the folding transition state of protein L. *Journal of Molecular Biology* **2000,** 298, (5), 971-984.

85. Krantz, B. A.; Mayne, L.; Rumbley, J.; Englander, S. W.; Sosnick, T. R., Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *Journal of Molecular Biology* **2002,** 324, (2), 359-371.

86. McCallister, E. L.; Alm, E.; Baker, D., Critical role of beta-hairpin formation in protein G folding. *Nature Structural Biology* **2000,** 7, (8), 669-673.

87. Park, S. H.; ONeil, K. T.; Roder, H., An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core. *Biochemistry* **1997,** 36, (47), 14277-14283.

88. Park, S. H.; Shastry, M. C. R.; Roder, H., Folding dynamics of the B1 domain of protein G explored by ultrarapid mixing. *Nature Structural Biology* **1999,** 6, (10), 943-947.

89. Roder, H.; Maki, K.; Cheng, H., Early events in protein folding explored by rapid mixing methods. *Chemical Reviews* **2006,** 106, (5), 1836-1861.

90. Roder, H.; Maki, K.; Cheng, H.; Shastry, M. C. R., Rapid mixing methods for exploring the kinetics of protein folding. *Methods* **2004,** 34, (1), 15-27.

91. Scalley, M. L.; Yi, Q.; Gu, H. D.; McCormack, A.; Yates, J. R.; Baker, D., Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L. *Biochemistry* **1997,** 36, (11), 3373-3382.

92. Karanicolas, J.; Brooks, C. L., The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Science* **2002,** 11, (10), 2351-2361.

93. Kuszewski, J.; Clore, G. M.; Gronenborn, A. M., Fast Folding of a Prototypic Polypeptide - the Immunoglobulin Binding Domain of Streptococcal Protein-G. *Protein Science* **1994,** 3, (11), 1945-1952.

94. Andersen, H. C., Rattle - a Velocity Version of the Shake Algorithm for Molecular-Dynamics Calculations. *Journal of Computational Physics* **1983,** 52, (1), 24-34.

95. Ferguson, D. M.; Garrett, D. G., Simulated annealing - Optimal histogram methods. *Monte Carlo Methods in Chemical Physics* **1999,** 105, 311-336.

96. Ferrenberg, A. M.; Swendsen, R. H., Optimized Monte-Carlo Data-Analysis. *Physical Review Letters* **1989,** 63, (12), 1195-1198.

97. Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S., On the transition coordinate for protein folding. *Journal of Chemical Physics* **1998,** 108, (1), 334-350.

98. Shindyalov, I. N.; Bourne, P. E., Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* **1998,** 11, (9), 739-747.

99. Feig, M.; Karanicolas, J.; Brooks, C. L., MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics & Modelling* **2004,** 22, (5), 377-395.

100. Marianayagam, N. J.; Fawzi, N. L.; Head-Gordon, T., Protein folding by distributed computing and the denatured state ensemble. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** 102, (46), 16684-16689.

101. DeLano, W. L. The PyMOL Molecular Graphics System.

102. Davis, M. E.; Mccammon, J. A., Electrostatics in Biomolecular Structure and Dynamics. *Chemical Reviews* **1990,** 90, (3), 509-521.

103. Chapman, D. L., A contribution to the theory of electrocapillarity. *Philosophical Magazine Series 6* **1913,** 25, (148), 475 - 481.

104. Gouy, G., The mutual action of two cathodes in a magnetic field. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences* **1910,** 150, 1652-1655.

105. Debye, P.; Huckel, E., The theory of electrolytes I. The lowering of the freezing point and related occurrences. *Physikalische Zeitschrift* **1923,** 24, 185-206.

106. Derjaguin, B.; Landau, L., Theory of the Stability of Strongly Charged Lyophobic Sols and of the Adhesion of Strongly Charged-Particles in Solutions of Electrolytes. *Progress in Surface Science* **1993,** 43, (1-4), 30-59.

107. Verwey, E. J. W., Theory of the Stability of Lyophobic Colloids. *Philips Research Reports* **1945,** 1, (1), 33-49.

108. Lu, B. Z.; Zhou, Y. C.; Holst, M. J.; McCammon, J. A., Recent progress in numerical methods for the Poisson-Boltzmann equation in biophysical applications. *Communications in Computational Physics* **2008,** 3, (5), 973-1009.

109. Kirkwood, J. G., Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions. *Journal of Chemical Physics* **1934,** 2, (7), 351-362.

110. Fenley, A. T.; Gordon, J. C.; Onufriev, A., An analytical approach to computing biomolecular electrostatic potential. I. Derivation and analysis. *Journal of Chemical Physics* **2008,** 129, (7), 075101-075111.

111. Mcclurg, R. B.; Zukoski, C. F., The electrostatic interaction of rigid, globular proteins with arbitrary charge distributions. *Journal of Colloid and Interface Science* **1998,** 208, (2), 529-542.

112. Phillies, G. D., Effects of Intermacromolecular Interactions on Diffusion.2. 3-Component Solutions. *Journal of Chemical Physics* **1974,** 60, (3), 983-989.

113. Sader, J. E.; Lenhoff, A. M., Electrical double-layer interaction between heterogeneously charged colloidal particles: A superposition formulation. *Journal of Colloid and Interface Science* **1998,** 201, (2), 233-243.

114. Rocchia, W.; Alexov, E.; Honig, B., Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *Journal of Physical Chemistry B* **2001,** 105, (28), 6507-6514.

115. Chen, L.; Holst, M. J.; Xu, J. C., The finite element approximation of the nonlinear Poisson-Boltzmann equation. *Siam Journal on Numerical Analysis* **2007,** 45, (6), 2298-2320.

116. Holst, M.; Baker, N.; Wang, F., Adaptive multilevel finite element solution of the Poisson-Boltzmann equation I. Algorithms and examples (vol 21, pg 1319, 2000). *Journal of Computational Chemistry* **2001,** 22, (4), 475-475.

117. Zhou, Z. X.; Payne, P.; Vasquez, M.; Kuhn, N.; Levitt, M., Finite-difference solution of the Poisson-Boltzmann equation: Complete elimination of self-energy. *Journal of Computational Chemistry* **1996,** 17, (11), 1344-1351.

118. Gumerov, N. A.; Duraiswami, R., Recursions for the computation of multipole translation and rotation coefficients for the 3-D helmholtz equation. *Siam Journal on Scientific Computing* **2003,** 25, (4), 1344-1381.

119. Arfken, G., *Mathematical methods for physicists*. 3rd ed.; Academic Press: 1985.

120. Chen, G.; Zhou, J., *Boundary Element Methods*. 1st ed.; Academic Press: 1992.

121. Cheng, H.; Greengard, L.; Rokhlin, V., A fast adaptive multipole algorithm in three dimensions. *Journal of Computational Physics* **1999,** 155, (2), 468-498.

122. Sanner, M. F.; Olson, A. J.; Spehner, J. C., Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **1996,** 38, (3), 305-320.

123. Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research* **2007,** 35, W522-W525.

124. Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A., PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research* **2004,** 32, W665-W667.

125. Wang, J. M.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **2000,** 21, (12), 1049-1074.

126. Davis, M. E.; Mccammon, J. A., Calculating Electrostatic Forces from Grid-Calculated Potentials. *Journal of Computational Chemistry* **1990,** 11, (3), 401-409.

127. Gilson, M. K.; Davis, M. E.; Luty, B. A.; Mccammon, J. A., Computation of Electrostatic Forces on Solvated Molecules Using the Poisson-Boltzmann Equation. *Journal of Physical Chemistry* **1993,** 97, (14), 3591-3600.

128. Lu, B. Z.; Zhang, D. Q.; McCammon, J. A., Computation of electrostatic forces between solvated molecules determined by the Poisson-Boltzmann equation using a boundary element method. *Journal of Chemical Physics* **2005,** 122, (21), 214102-214109.

129. Yamakawa, H., *Modern Theory of Polymer Solutions*. Harper and Row: New York, 1971.

130. Rotne, J.; Prager, S., Variational Treatment of Hydrodynamic Interaction in Polymers. *Journal of Chemical Physics* **1969,** 50, (11), 4831-&.

131. Johnson, R. E.; Ranganathan, S., Generalized approach to Ewald sums. *Physical Review E* **2007,** 75, (5), 056706-056714.

132. Salin, G.; Caillol, J. M., Ewald sums for Yukawa potentials. *Journal of Chemical Physics* **2000,** 113, (23), 10459-10463.

133. Ayton, G. S.; Noid, W. G.; Voth, G. A., Multiscale modeling of biomolecular systems: in serial and in parallel. *Current Opinion in Structural Biology* **2007,** 17, (2), 192-198.

134. Shi, Q.; Izvekov, S.; Voth, G. A., Mixed atomistic and coarse-grained molecular dynamics: Simulation of a membrane-bound ion channel. *Journal of Physical Chemistry B* **2006,** 110, (31), 15045-15048.

135. Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M., Resolution exchange simulation. *Physical Review Letters* **2006,** 96, (2), 028105-028109.

136. Heath, A. P.; Kavraki, L. E.; Clementi, C., From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins-Structure Function and Bioinformatics* **2007,** 68, (3), 646-661.

137. Ercolessi, F.; Adams, J. B., Interatomic Potentials from 1St-Principles Calculations - the Force-Matching Method. *Europhysics Letters* **1994,** 26, (8), 583-588.

138. Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A., Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching. *Journal of Chemical Physics* **2004,** 120, (23), 10896-10913.

139. Izvekov, S.; Voth, G. A., A multiscale coarse-graining method for biomolecular systems. *Journal of Physical Chemistry B* **2005,** 109, (7), 2469-2473.

140. Izvekov, S.; Voth, G. A., Modeling real dynamics in the coarse-grained representation of condensed phase systems. *Journal of Chemical Physics* **2006,** 125, (15), 151101-151105.

141. Glick, M.; Grant, G. H.; Richards, W. G., Docking of flexible molecules using multiscale ligand representations. *Journal of Medicinal Chemistry* **2002,** 45, (21), 4639-4646.

142. Gillespie, D. T., Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry* **2007,** 58, 35-55.

143. Sept, D.; McCammon, J. A., Thermodynamics and kinetics of actin filament nucleation. *Biophysical Journal* **2001,** 81, (2), 667-674.

144. Gillespie, D. T., Exact Stochastic Simulation of Coupled Chemical-Reactions. *Journal of Physical Chemistry* **1977,** 81, (25), 2340-2361.

145. Chodera, J. D.; Dill, K. A.; Swope, W. C.; Pitera, J. W., Constructing master equation models of protein folding and dynamics from atomistic simulation. *Protein Science* **2004,** 13, 101-102.

146. Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A., Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Modeling & Simulation* **2006,** 5, (4), 1214-1226.

147. Hemberg, M.; Yaliraki, S. N.; Barahona, M., Stochastic kinetics of viral capsid assembly based on detailed protein structures. *Biophysical Journal* **2006,** 90, (9), 3029-3042.

148. Burack, W. R.; Shaw, A. S., Signal transduction: hanging on a scaffold. *Current Opinion in Cell Biology* **2000,** 12, (2), 211-216.

149. Cuillel, M.; Berthetcolominas, C.; Krop, B.; Tardieu, A.; Vachette, P.; Jacrot, B., Self-Assembly of Brome Mosaic-Virus Capsids - Kinetic-Study Using Neutron and X-Ray Solution Scattering. *Journal of Molecular Biology* **1983,** 164, (4), 645-650.

150. Cuillel, M.; Zulauf, M.; Jacrot, B., Self-Assembly of Brome Mosaic-Virus Protein into Capsids - Initial and Final-States of Aggregation. *Journal of Molecular Biology* **1983,** 164, (4), 589-603.

151. Berthetcolominas, C.; Cuillel, M.; Koch, M. H. J.; Vachette, P.; Jacrot, B., Kinetic-Study of the Self-Assembly of Brome Mosaic-Virus Capsid. *European Biophysics Journal with Biophysics Letters* **1987,** 15, (3), 159-168.

152. Chen, C.; Kao, C. C.; Dragnea, B., Self-assembly of brome mosaic virus capsids: Insights from shorter time-scale experiments. *Journal of Physical Chemistry A* **2008,** 112, (39), 9405-9412.

153. Hagan, M. F.; Chandler, D., Dynamic pathways for viral capsid assembly. *Biophysical Journal* **2006,** 91, (1), 42-54.