Chapter 8
Simulating Protein Folding in Different Environmental Conditions

Dirar Homouz

Abstract Molecular dynamics simulations have become an invaluable tool in investigating the dynamics of protein folding. However, most computational studies of protein folding assume dilute aqueous simulation conditions in order to reduce the complexity of the system under study and enhance the efficiency. Nowadays, it is evident that environmental conditions encountered *in vivo* (or even *in vitro*) play a major role in regulating the dynamics of protein folding especially when one considers the highly condensed environment in the cellular cytoplasm. In order to factor in these conditions, we can utilize the high efficiency of well-designed low resolution (coarse-grained) simulation models to reduce the complexity of these added protein-milieu interactions involving different time and length scales. The goal of this chapter is to describe some recently developed coarse-grained simulation techniques that are specifically designed to go beyond traditional aqueous solvent conditions. The chapter also gives the reader a flavor of the things that we can study using such "smart" low resolution models.

Keywords Molecular dynamics • All-atom models • Coarse-grained models • Multi-scale methods • Proteins • Folding • Crowding • Urea • HP model • Gō model • Statistical potential • C_α models • Side-chain-C_α Model (SCM) • Boltzmann inversion • SCAAL • MultiSCAAL

D. Homouz (✉)
AMS Department, Khalifa University, P.O. Box 127788, Abu Dhabi, UAE
e-mail: dirar.homouz@kustar.ac.ae

K. Han et al. (eds.), *Protein Conformational Dynamics*, Advances in Experimental Medicine and Biology 805, DOI 10.1007/978-3-319-02970-2_8,

8.1 Introduction

Molecular Dynamics (MD) simulation is a computer computational method that utilizes the laws of classical statistical physics in order to predict the behavior of many particle systems. The history of MD is tied to the history of the development of computer technology. The first real system to be studied using MD simulations was in 1964 by Rahman [1] who simulated liquid argon at 94.4 K. The system simulated by Rahman was limited to only 864 particles. Studying bigger systems with more particles became increasingly more feasible with the continual growth in computational power and speed. The pioneering work of McCammon et al. [2] marked the beginning of new era in using MD simulations in the very important biological problem of protein folding.

Most of the functions performed in a living cell are carried out by different proteins. In order for these proteins to function properly they have to be in their functional shape or fold. Proteins are large biomolecules that consist of one or more chains of amino acids. Thus, understanding the dynamics of how a protein can go from unfolded sequence of amino acids into its functional three dimensional fold is one of the fundamental problems in biology. MD simulations became an invaluable tool for studying protein folding and unfolding dynamics. It is used in conjunction with several experimental techniques in order to understand and interpret the experimental results at the atomic level. For more details on the MD history and techniques in protein folding studies we refer the reader to the following review articles [3–5].

In recent years, it became very obvious that the folding of proteins is highly dependent on their environmental conditions. Thus, the native protein folds are likely to be different from the ones usually determined by experimental techniques such as x-ray crystallography and NMR as these methods don't account for the densely crowded cellular environment. Several experimental studies have recently started factoring in these crowding effects in their experimental design by adding synthetic chowders to mimic the macromolecular crowding in the cell [6–14]. In addition to crowding, other cellular conditions can affect protein folding and stability such as the concentration of different ions. Well-designed computer simulation schemes are needed in order to better understand the role that all these environmental factors play in determining protein structure. In order to efficiently simulate protein interactions *in vivo*, one has to account for different sizes of interacting particles and different time scales.

In this chapter we present a multi-scale molecular dynamics scheme that can be used to simulate protein interactions in different crowding and solvent conditions. This scheme is based on a low resolution simulation model Side-chain C_α Model (SCM) [15] that was previously implemented in studying the protein folding dynamics in crowded environment. However, this model can't handle other environmental factors with small length scales besides the large crowders. Thus, SCM is integrated into a multi-scale algorithm (MultiSCAAL) [16] that deals with both large macromolecular crowders and small interfering chemicals. This scheme enables us to simulate proteins in many cellular as well as experimental conditions.

The material in this chapter is organized as follows: Sect. 8.2 gives a short overview of molecular dynamics simulations in the context of protein folding applications. In Sect. 8.3 we describe SCM and how it is integrated into MultiSCAAL scheme. In Sect. 8.4 we discuss some of the applications of these various techniques. Finally, we close this chapter with conclusions.

8.2 Molecular Dynamics and Protein Folding

8.2.1 All-Atom Versus Coarse-Grained

Different Molecular Dynamics simulation schemes are distinguished by the models they use to represent proteins and their interactions. These models differ in the level of detail, or resolution, that they reflect. Traditionally, these models are classified into two classes; All-Atom (AA) and Coarse-Grained (CG) models. AA models, with their explicit solvent representation, provide a great deal of detail at very short time scales (picoseconds). However, the inverse relationship between the resolution and computational cost usually limits the applicability of AA models when it comes to simulating protein folding trajectories with long timescales (microseconds). In addition, the computational cost grows exponentially when one considers environmental interactions with solvent, crowders, and other ions.

On the other hand, CG models with implicit solvents average out all amino acid atomic sites and replace them with a smaller number of beads, typically one or two. Thus, with these CG models, the accuracy of atomistic details and the reliability of energy functions are reduced. However, this is the price that one has to pay in order to capture the main features of protein folding over reasonable biological times. CG models are capable of increasing the timescale of molecular simulations due to the huge reduction in the number of degrees of freedom in the systems simulated mainly due to replacing all the degrees of freedom of the solvent with a mean field implicit solvent representation with zero degrees of freedom. Thus, with existing computer technology, CG simulations seem to be the only viable solution in order to study protein folding especially when the right environmental conditions are considered.

8.2.2 Coarse-Grained Models for Protein Folding

The famous experiments of Anfinsen et al. [17] in the early 1960s have instigated a large interest in the problem of protein folding. These experiments show that proteins can fold and refold reversibly to the same native state (functional state) which means that this state is thermodynamically stable and forms a global minimum. This conclusion raised the question of how can proteins reach this minimum starting from an unfolded state in a relatively short time ($\sim$ms) given the large number of possible conformations of any given protein. Levinthal [18] tried to resolve this paradox by suggesting that proteins follow a specified (encoded) kinetic "folding pathway" to reach its global minimum.

Several objections were raised against the idea of folding pathways and alternative views were proposed [19]. Among these alternative views, the Energy Landscape Theory was the most acceptable one. According to the Energy Landscape Theory, proteins don't follow a single pathway to reach the native state. Rather, they can follow multiple routes down a biased energy landscape towards the global minimum [20–22]. In other words, the energy landscape of protein folding process has a funnel-like shape and the folding is viewed as a flow process of an ensemble of routes down this funnel. The energy funnel is controlled by both its bias towards the native state and its roughness. In order for the protein to have fast folding, the roughness has to be small compared to the bias. This concept gave rise to the Principle of Minimal Frustration [23, 24] which can be justified by the fact that folding processes have evolved to make the native state more stable, favor stabilizing interactions, and make folding processes fast [25].

Coarse-grained computer models of proteins tried to conform to these competing views of protein folding processes. Early models used simplified geometries as well as energy functions. Lattice models achieved an early success due to the great simplification in the simulation geometry [24, 26–28]. In these models, proteins were modeled as self-avoiding polymer chains of one-bead amino acids where the beads on the chain are confined to move on a fixed three dimensional cubic lattice. These simplified models used fictitious energy functions such as HP [28] and Gō [29] energy functions. The HP model distinguishes between two types of monomers, H (Hydrophobic) and P (Polar), and assumes an attractive interaction between HH pairs and none between all other pairs. Gō model on the other hand tries to bias the energy function towards the native state by assuming attractive interactions for native contacts and repulsive interactions for none-native contacts. The Gō model gained more recognition later since it conforms to the Energy Landscape Theory and the principle of minimal frustration. Several Gō-like energy functions were developed later to be used with more advanced CG models [30].

The lattice models gave way to off-lattice models as computer power improved. This development allowed for more realistic representation of protein's geometry. Most of the early off-lattice models relied on simplified energy functions and one-bead amino acid representation [31–33]. These models are typically called C_α models since each amino acid is represented by one site located at the C_α carbon position. These C_α models started to take shape and give more faithful representation of protein by adopting more sophisticated energy functions (force fields) that included different type of structural as well as non-bonded interactions.

The difficulty in designing these dimensionally reduced C_α models lies in choosing the proper force field. There were different strategies for choosing the interaction energies between the 20 different types of beads (20 different amino acids). The structural energy terms (bond, angle, dihedral) were typically chosen such that they produce a thermodynamically stable structure. The non-bonded interactions could be still borrowed from earlier fictitious energy functions such as Gō model. However, more improved models tried to base these interaction energies on measured experimental values of amino acid pair potentials. Examples of such interaction maps are the Betancourt-Thirumalai (BT) statistical potential [34] and the Miyazawa-Jernigan (MJ) potential [35].

The C_α models gave way to more advanced models that incorporate more structural details of proteins. Cheung et al. [15] introduced one such a model in which each amino acid is represented by two beads; one at the C_α position and the second one at the center of mass of the side chain. This model called Side-chain C_α Model (SCM) falls between C_α and AA models and is capable of accounting for side-chain packing while keeping the computational cost low. This model was very successful in addressing protein folding interactions in crowded medium and confined geometries [6, 7, 36, 37]. With such improvements, the CG models start to look more like AA models and include more interactions which enable them to simulate different biological and experimental conditions. More information about CG models of protein folding can be found in these reviews [38, 39].

8.3 Flexible Low Resolution Simulation Techniques

The success of CG molecular dynamics stems from their ability to simulate protein folding and refolding events over large time scales. They do so by capturing the main features of the protein, stripping away complex details, and using implicit solvent models. In fact the greatest reduction in computation cost and time comes from replacing the atomic details of water with implicit solvent model. Thus, this approach works well for studying folding dynamics of isolated proteins or protein-protein interactions. In addition, the same CG models can be easily extended to studying protein folding in crowded medium where the dominant crowding agents are large macromolecules that can be themselves coarse-grained. However, this approach will be useless if one has to deal with environmental conditions that are controlled by small particles ($\sim$water molecule size) like urea. The reason being that the simplification and reduction in computational time achieved by removing water molecules will be undone by including a large number of these additional small molecules.

Taking these points into consideration, CG models have to be modified and a multi-scale approach is needed in order to capture both protein and environment details without sacrificing the computational efficiency. Here we present the details of the modifications that can be done to a simple two-bead model in order to develop it into a multi-scale algorithm. This is done by using SCM at the core to model proteins and large crowders, Langevin Dynamics to represent water solvent conditions, and adjusting force field parameters for different solvent conditions in order to account for chemical interference effects. The main elements of the final

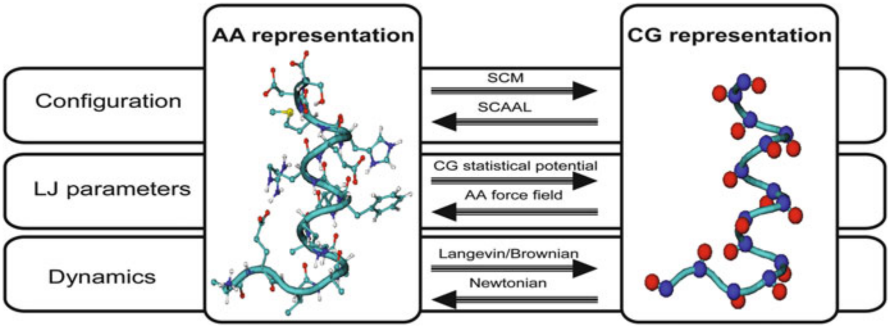


Fig. 8.1 A schematic diagram in a multi-scale algorithm where a protein configuration switches from all-atomistic (AA) to coarse-grained (CG) representation and vice versa. A side-chain-C_α model (SCM) is used as a coarse-grained model. The reconstruction of a protein in an AA representation from CG representation is achieved by SCAAL. The Lennard-Jones (LJ) parameters for an AA representation follow atomistic force field, while for a CG representation they follow a statistical potential based on bioinformatics and the potential of mean force from the AA molecular dynamic simulations via Boltzmann inversion method. The dynamics of an AA protein is governed by the Newtonian equations of motion. The dynamics of a CG protein is governed by the Langevin/Brownian equations of motion

multi-scale scheme, MultiSCAAL, are shown in Fig. 8.1 where we can see that SCM model is used to build the coarse-grained model starting from the corresponding all-atom representation. The scheme also includes the algorithm, Side-chain C Alpha to All-atom (SCAAL), which enables us to construct the all-atom representation of a protein starting from its course grained model. The Lennard-Jones (LJ) parameters for nonbonded interactions are based on a CG statistical potential. The dynamics that we use to sample the phase space of the protein is the Langevin Dynamics in order to account for the water solvent conditions implicitly. The details of these different elements and the implementation of the MultiSCAAL algorithm are given in the subsections below.

8.3.1 SCM Model (Representation & Hamiltonian)

A Sidechain-C_α (SCM) [15] coarse-grained model is used to represent proteins where each amino acid (except glycine) is modeled by two beads: a C_α bead and a side-chain bead located at the center of mass of the side-chain. The potential energy of a protein, E_p is the sum of three terms; the structural energy (E_{Struc}), the nonbonded energy (E_{NB}), and the Hydrogen bond energy (E_{HB})

$$E_p = E_{Struc} + E_{NB} + E_{HB} \tag{8.1}$$

8.3.1.1 Structural Energy

The structural energy, E$_{\text{Struc}}$, consists of the terms that account for all of the topological constraints of our structure. It is the sum of bond-length potential (E$_{\text{bond}}$), bond-angle potential (E$_{\text{angle}}$), dihedral potential (E$_{\text{dih}}$), and chiral interactions (E$_{\text{chi}}$).

$$E_{Struc} = E_{bond} + E_{angle} + E_{dih} + E_{chi} \tag{8.2}$$

The bond-length potential (E$_{\text{bond}}$) and the bond-angle potential (E$_{\text{angle}}$) are represented by harmonic springs as follows:

$$E_{bond} = \sum_{bonds} k_b (r - r_0)^2 \tag{8.3}$$

$$E_{angle} = \sum_{angles} k_\theta (\theta - \theta_0)^2 \tag{8.4}$$

Dihedral potential (E$_{\text{dih}}$) for every four consecutive C$_\alpha$ beads is represented by:

$$E_{dih} = \sum_{dihedrals}^{C_\alpha - C_\alpha - C_\alpha - C_\alpha} k_\phi^{(n)} \left[1 - \cos\left(n\left(\phi - \phi_0\right)\right)\right] \tag{8.5}$$

where ϕ is the dihedral angle, r is the distance between two adjacent beads and θ is the angle of three consecutive beads. The equilibrium values of ϕ_O, θ_O, and r_O are calculated based on the native all-atom structure of a protein. The force constants are given these values $k_b = 100\varepsilon, k_\theta = 20\varepsilon$, $k_\phi^{(1)} = \varepsilon$, and $k_\phi^{(3)} = 0.5\varepsilon$, where $\varepsilon = 0.6$ kcal/mol.

The chiral energy (E$_{\text{chi}}$) accounts for an L-isoform preference of side chains. This energy is given by:

$$E_{chi} = \sum_{chiral} k_c (c - c_0)^2 \tag{8.6}$$

where c is the triple scalar product defined as $c = \overrightarrow{r}_{C_\alpha^i C_{SC}^i} \cdot \left(\overrightarrow{r}_{C_\alpha^i C_\alpha^{i-1}} \times \overrightarrow{r}_{C_\alpha^i C_\alpha^{i+1}}\right)$, c_O is determined based on the native structure of the protein and $k_c = 20\varepsilon$. C_α^i and C_{SC}^i are the C$_\alpha$ bead and side-chain bead of the ith residue of the protein, respectively.

8.3.1.2 Nonbonded Energy

Nonbonded interaction energy E_{NB}^{ij} between a pair of i and j side-chain beads at a distance r has an LJ potential of the form,

$$E_{NB}^{ij} = \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{8.7}$$

where $\sigma_{ij} = f(\sigma_i + \sigma_j)$, σ_i and σ_j are the Van der Waals (VdW) radii of side-chain beads, $|i–j| > 2$, and f is a control scaling factor that is used to prevent clashes that might destabilize the native state. The values of ε_{ij} are based on the solvent-mediated interaction between pairs of residues. For water solvent conditions we use the Betancourt-Thirumalai statistical potential map [34]. For other solvents this map can be modified according to the recipe give in Sect. 8.3.3.

Repulsive hard-core potential is used to model excluded volume interactions between C_α–Side-chain nonbonded pairs. This potential is given by this form:

$$E_{NBrep}^{ij} = \varepsilon\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \tag{8.8}$$

8.3.1.3 Hydrogen Bond Energy

For backbone hydrogen bonding interactions, an angular-dependent function is used to capture directional properties of backbone hydrogen bonds. For a pair of i and j C_α beads, the hydrogen bond interaction is given by:

$$E_{HB}^{ij} = A(\rho)\, E_{NB}^{ij} \tag{8.9}$$

$$A(\rho) = \frac{1}{\left[1 + (1 - \cos^2\rho)\left(1 - \frac{\cos\rho}{\cos\rho_a}\right)\right]^2} \tag{8.10}$$

where E_{NB}^{ij} has the same form as in Eq. (8.8), except that $\varepsilon_{\mathrm{ij}}$ for backbone hydrogen bonding is 0.6 kcal/mol and σ_{ij} is the hydrogen bond length, 4.6 Å.

The Lorentzian function $A(\rho)$ in Eq. (8.10) restricts the structural alignment of two interacting strands such that local backbone orientational configurations of parallel β sheets, antiparallel β sheets, or left and right-handed α helices are favored. The parameter ρ is the pseudo-dihedral angle between two interacting strands of the backbone. The function $A(\rho)$ will have its maximum value of 1 when $\rho = 0$ (the alignment that points to β-strands or α-helices) or when $\rho = \rho_a$ (the pseudo-dihedral angle of a canonical helical turn, 0.466 rad). For all other pseudo-dihedral angles (ρ) the value of $A(\rho)$ will be diminished (much smaller than 1).

8.3.1.4 SCM with Gō-Like Hamiltonian

The energy terms presented above are used to model proteins with non-specific nonbonded interactions. However, these terms can be manipulated easily to produce a topologically based Gō-like model that provides a minimally frustrated energy landscape. In such a model, the nonbonded interactions found in the native structure of the protein retain their sticky interaction represented by LJ potential of the form give in Eq. (8.7). All other non-native nonbonded pairs will be assigned repulsive interaction of the form in Eq. (8.8). The same rule can be applied to hydrogen bonding where native interactions will be represented by Eq. (8.9) while the non-native ones are represented by repulsive potential.

This kind of flexibility enables to tailor the SCM model to our computational needs. While the SCM model with non-specific Hamiltonian can explore bigger regions of the energy landscape than a one with Gō-like Hamiltonian, it is more expensive computationally. Thus, when we are interested in protein folding problems where the focus is on transitions out or into the native state we can utilize the Gō-like based SCM model.

8.3.2 Langevin Dynamics (Implicit Solvent)

To account for the effect of the solvent on the protein dynamics the Langevin equation of motion [40] is used to describe the dynamics in SCM coarse-grained molecular simulations. The solvent is treated implicitly in the Langevin equation through a stochastic term. The Langevin equation of motion for a general coordinate x is:

$$m\ddot{x} = -\frac{\partial U}{\partial x} - \zeta \dot{x} + \varGamma, \tag{8.11}$$

where m is the mass and U is the potential energy of the molecule. The drag term, $-\zeta \dot{x}$, or the dissipation term, is caused by friction which is compensated by a random force $\varGamma$ representing random collisions with solvent molecules. $\varGamma$ is sampled from a distribution of a white noise (Gaussian noise).

Fast motions of large biomolecules are quickly damped in a viscous solvent such as water. As a result, they follow random trajectories referred to as the Brownian motion. The inertia term is dropped in Eq. (8.11) and we get the first order ordinary differential equation for the Brownian motion given by:

$$\zeta \dot{x} = -\frac{\partial U}{\partial x} + \varGamma. \tag{8.12}$$

8.3.3 Different Solvent Conditions (Modifying LJ Parameters)

The techniques implemented in SCM were designed to simulated protein folding in aqueous medium. However, we are presented with many situations where it is important to study protein folding/refolding in different solvent conditions. One such a situation arises when one wants to simulate the experimental unfolding of proteins in different urea concentrations or the experimental folding of a protein in the presence of small molecules such as salt and alcohol, or small crowders such as glycerol. Extending SCM to cover these situations presents us with great challenge since these small molecules have the same length scale as water. These solvent conditions can be readily handled in AA simulations. Thus, one has to devise a multi-scale approach that can benefit from AA models of these solvents and feeds back into GC simulations. This is the approach used here in order to implicitly account for chemical interference in solvents by adjusting the solvent-mediated amino acid pair interaction energies. The details of the technique used to adjust these parameters are given below.

8.3.3.1 The Choice of Parameters ε_{ij}

In order to design a coarse-grained model that can accommodate the chemical properties of different amino acids we chose our nonbonded LJ interaction parameters in Eq. (8.7) based on knowledge-based potentials. These knowledge-based (or statistical) potentials are matrices (of 210 elements) that give the solvent-mediated interaction energies between all pairs of amino acids. There are several schemes for calculating these potentials such as those of Miyazawa and Jernigan [35], Kolinski and Skolnick [41], or Betancourt-Thirumalai [34]. Our model is based on the Betancourt-Thirumalai statistical potential [34]. This statistical potential addresses sequence variations where the reference interaction, $\varepsilon = 0.6$ kcal/mol, is based on the Thr-Thr pairwise interaction.

8.3.3.2 The Statistical Potential Map in a Different Solvent

All of the available statistical potential maps give the interactions energies between amino acids in water. Using SCM model to simulate proteins in other solvents such as urea requires expanding the idea of statistical potential maps to other solvents. In principle, the statistical potential between two residues should be the same as the potential of mean force (PMF) between these residues. The effect of the solvent is implicitly accounted for in the statistical potential. Calculating the potential of mean force is inherently complex and inefficient. The direct calculation of the residue-residue interaction from the PMF is therefore not attainable. However, creating the statistical potential parameter map (SPPM) is a much simpler problem.

In order to get the statistical potential parameter map (SPPM) for a certain solvent, we compute the PMFs of pairs of amino acids using all-atom simulations of free residues in that solvent. We circumvent the inherent difficulty of calculating this PMF by simulating a large number of copies of each pair at once, instead of one pair. For instance, in order to calculate the parameter ε_{TT} between two Threonine (Thr) residues we run a simulation of a large number of solvated free Thr residues. This method helps enhance the sampling and converge the PMF for this pair of residues. We make two approximations in order to further simplify the calculation of the statistical potential map. First, we approximate the PMF between a pair of amino acids by a two point correlation function of the distance between the two centers of mass of the side chains. Second, we fit the calculated PMF to a Lennard Jones (LJ) potential and set the statistical potential parameter to be equal to the depth of the resulting LJ potential. This calculation is done by using the Boltzmann inversion method discussed below.

8.3.3.3 Boltzmann Inversion

The CG energy function that accommodates chemical interference can be created using Boltzmann inversion [42–44] based on data obtained from all-atomistic molecular dynamics simulations. The pair correlation function between any two amino acid types i and j at a distance r in type α solvent is $g_{ij}^{\alpha}(r)$. This function is related to the potential of mean force, $U_{ij}^{\alpha}(r)$, between the same pair of amino acids through Boltzmann inversion at temperature T by the following formula [45]:

$$U_{ij}^{\alpha}(r) = -k_B T \ln\left[\frac{g_{ij}^{\alpha}(r)}{\rho_o}\right], \tag{8.13}$$

where ρ_o is the average density of the system (amino acid pairs and the solvent) and k_B is the Boltzmann constant. The average density ρ_o is used to normalize the pair correlation function at distances greater than the excluded volume radius. The solvent mediated interactions $\varepsilon'^{\alpha}_{ij}$ for every pair of amino acids i and j is equal to $U^{\alpha}{}_{ij}(r^*)$

$$\varepsilon'^{\alpha}_{ij} = U_{ij}^{\alpha}\left(r^*\right), \tag{8.14}$$

where r^* denotes the first highest peak of $g_{ij}^{\alpha}(r)$. Next $\varepsilon'^{\alpha}_{ij}$ is shifted by a constant, V_o.,

$$\varepsilon_{ij}^{\alpha} = \varepsilon'^{\alpha}_{ij} + V_o. \tag{8.15}$$

where V_o is obtained from a Threonine pair by setting $\varepsilon'^{\alpha}_{TT}$ (in water) from the simulation equal to $\varepsilon^{\alpha}_{TT}$ from the statistical potential of the same amino acid pair [34].

A Lennard-Jones potential (LJ), $V_{ij}^{\alpha}(r)$, is used to approximate the overall profile of $U_{ij}^{\alpha}(r)$ [46] and it is the energy function for the same type of amino acids in coarse-grained molecular simulation:

$$V_{ij}^{\alpha}(r) = \varepsilon_{ij}^{\alpha}\left[\left(\frac{r_{ij}^{o}}{r}\right)^{12} - 2\left(\frac{r_{ij}^{o}}{r}\right)^{6}\right]. \tag{8.16}$$

$\varepsilon_{ij}^{\alpha}$ is the solvent-mediated interaction of an amino acid pair i and j in solvent type α. r_{ij}^{o} is the bonding distance. Figure 8.2 shows how Boltzmann inversion

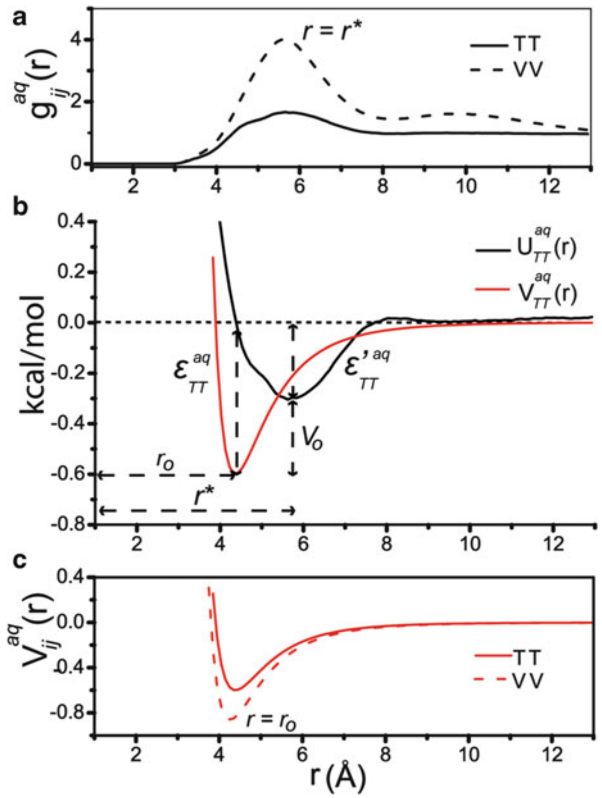


Fig. 8.2 (**a**) Pair correlation function $g^{aq}_{ij}(r)$ for Thr-Thr (*solid line*) and Val-Val pairs (*dotted line*) derived from all–atomistic molecular dynamics simulations in aqueous condition. $r = r*$ is at the maximum $g^{aq}_{ij}(r^*)$. (**b**) The potential of mean force $U^{aq}_{TT}(r)$ of Thr-Thr interaction (in *black*) is obtained from all–atomistic molecular simulations under aqueous condition through Boltzmann inversion (Eq. 8.13) as a function of *r*, distance between the chosen atoms (i.e. C_β atom for Thr.) that are in closest proximity to the center of mass of the side chain in threonine. r^* denotes the position of the major peak of the pair correlation function $g^{aq}TT(r)$ in (**a**) and $\varepsilon'^{aq}_{TT} = U^{aq}_{TT}(r^*)$. The Betancourt-Thirumalai statistical potential follows a Lennard-Jones interaction $V^{aq}_{TT}(r)$ (Eq. 8.16) for the same pair of amino acid in coarse-grained molecular simulations (in *red*). *r* is the interacting distance between the coarse-grained side-chain beads of the amino acids (i.e. center of mass of side chains). r_o is the bonding distance σ_{TT} in Eq. (8.7). $\varepsilon^{aq}_{TT} = V^{aq}_{TT}(r_o)$ is taken from the Betancourt–Thirumalai statistical potential. The reference potential from Eq. (8.15) is V_o. (**c**) $V^{aq}_{ij}(r)$ for Thr-Thr (*solid line*) and Val-Val pairs (*dotted line*) in aqueous solvent. r_o is the same bonding distance in (**b**)

is applied in practice to generate LJ parameters for amino acid pairs in water. In addition, Fig. 8.3 shows the accuracy of this process by comparing the SPPM for all amino acid pairs in water with the BT map. The end process result of this process is to generate a new SPPM of the parameters $\varepsilon^{\alpha}_{ij}$. Once this map is generated it can be then used for any CG simulation with the corresponding solvent. Important examples of these maps would be the maps of solvent mediated interaction for all 210 amino acid pairs in different concentrations of urea published in [16].

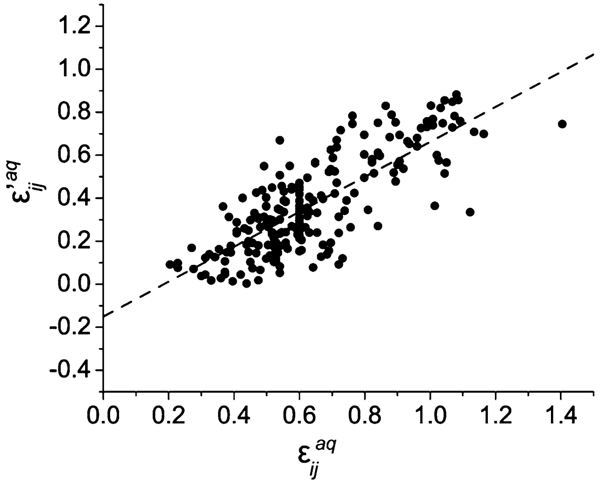


Fig. 8.3 The correlation between the aqueous solvent-mediated interactions between amino acids i and j, ε'^{aq}_{ij}, which are derived from the molecular dynamics simulations and the ones from the Betancourt-Thirumalai statistical potential ε^{aq}_{ij}. The linear correlation coefficient is 0.79

8.3.4 Crowders and Ions

More modifications were devised in order to account for other environmental factors such as large molecular crowders, electrostatic interaction, and ions. A short description of these modifications follows.

8.3.4.1 Macromolecular Crowders

Intracellular crowding can be mimicked experimentally by adding high concentrations of inert synthetic or natural macromolecules, termed crowding agents, to the systems *in vitro*. Inert large synthetic macromolecules such as Ficoll 70 and dextran can be readily included in CG simulations because of their large sizes. The atomic details of these particles will be irrelevant when we investigate their excluded volume effect on protein folding. Thus, they can be represented as hard particles with shapes that capture the geometry of each molecule. For instance Ficoll 70 can be modeled as a hard sphere and dextran as a hard dumbbell (two bonded spheres) of relevant size. In terms of the Hamiltonian, all the interactions that involve crowders (crowder-crowder, crowder-protein) will be repulsive with the same form given in Eq. (8.8). These repulsive interactions model the nonspecific steric space-filling repulsions due to the excluded volume effect of crowding. For other types of crowders such as the macromolecules in the cellular environment, a polydisperse CG model of these particles can be employed in order to mimic their different sizes and shapes.

8.3.4.2 Electrostatics and Ionic Concentration

In order to improve the accuracy and the performance of the coarse-grained (SCM) model, we included electrostatic interactions by adding a Debye-Hückel energy term [47]. This added term is supposed to represent screened Columbic interactions between charged sites. The charges are obtained using quantum chemistry calculations of the electronic structures of the all-atomistic representation of all residues in the protein. However, adding this term means that our Lennard Jones (LJ) potential parameters have to be adjusted. The original LJ parameters in the coarse-grained model were obtained from knowledge-based statistical potential which measures the solvent mediated interaction energies between different amino acid pairs including electrostatic interactions.

In order to adjust the LJ parameters in the coarse-grained simulation we first adjust the LJ parameters for every amino acid pair *(i,j)* as follows:

$$\varepsilon'_{ij} = \varepsilon_{ij} + \frac{e^2}{4\pi\varepsilon}\left(\frac{q_i q_j}{\sigma_{ij}}\right) = \varepsilon_{ij} + \alpha\left(\frac{q_i q_j}{\sigma_{ij}}\right). \tag{8.17}$$

where q_i and q_j are the charges of the two amino acids and σ_{ij} is the position of the minimum in the original LJ potential. Then, we can adjust the nonbonded interactions to have this form:

$$E^{ij}_{NB+Elect} = \varepsilon'_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \alpha\left(\frac{q_i q_j}{r_{ij}}\right). \tag{8.18}$$

The effects of ionic concentration in the solvent will be captured through a screening factor that changes Eq. (8.18) to this form

$$E^{ij}_{NB+Elect+I} = \varepsilon'_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \alpha\left(\frac{q_i q_j}{r_{ij}}\right)\exp\left(-r_{ij}\sqrt{(4\pi\alpha I/k_B T)}\right), \tag{8.19}$$

where *I* is the ionic concentration. The LJ parameters in Eq. (8.19) retain the same modified values according to Eq. (8.17).

8.3.5 Reconstructing the AA Coordinates (SCAAL)

Several methods of reconstructing reduced representation into all-atomistic structures have been developed over the last few years [48–51]. These include methods that can either recover the atomistic details of a protein's backbone with the knowledge of C_α beads [48], or reconstructing a full protein with the knowledge of its four heavy backbone atoms [49]. Methods that reconstruct all–atomistic structures from the information provided by a C_α bead and the center of mass of the side chain are also available [50]. However, the main purpose of the methods above is to reconstruct protein conformations that are very close to the crystal structures obtained by X-ray or NMR experiments. The use of rotamer libraries, obtained from PDB structures, in all these algorithms has made possible the development of very fast and accurate reconstruction methods. However, when reconstructing far – from the native state protein structures, which is most often the case in the course of a multi-scale simulation, it is questionable whether the accuracy of such methods can still be achieved. For this reason we have used a very simple approach based on the physics principle of harmonic constraints to reconstruct all-tom structures from coarse-grained ones in multi-scale simulation scheme.

In order to reconstruct the desired all-atom structure from coarse-grained models we use the positions from coarse-grained SCM as a part of harmonic constraints and apply them to an all-atom protein template through a process of energy minimization. For each residue, C_α positions from the SCM will be used as position constraints for C_α in the backbones from the all-atom template. As for the constraint of a side-chain position, it will be imposed on a heavy atom with the closest proximity to the actual center of mass of the side chain, in which the distance between the two is typically less than 1 Å. By doing this, the calculation of the center of mass of the side chain during a reconstruction algorithm is avoided by paying a small price on accuracy as long as we keep the harmonic spring constants at a reasonable range. During the reconstruction procedure that takes in both a SCM protein structure and an all-atom template as an input, the harmonic constraints imposed by a few chosen beads will carry the all-atom template to the desired structure, through driving forces of energy minimization, without the need for building a protein from individual atoms. The use of this "template concept" for protein reconstruction is depicted schematically in Fig. 8.4a and the flowchart of the SCAAL reconstruction algorithm is shown in Fig. 8.4b. The details of this method can be found in previous studies [7, 16].

8.3.6 MultiSCAAL: SCM + SCAAL

The improvements described above have transformed the SCM into a multifaceted algorithm that can be used to simulate protein folding in many different conditions. It can simulate the folding behavior in crowded environment that resembles the cellular conditions or reproduce the effect of synthetic crowding agents used in experimental studies to mimic cellular crowding. The modified SCM is capable of simulating experimental refolding events in the presence of denaturing factors such as urea or in the presence of other ions. Any combination of these different conditions (crowding, urea, ionic concentration) becomes accessible for simulation using low resolution protein representation.

In addition, combining reconstruction algorithm SCAAL with SCM results in a more sophisticated multi-scale scheme that combines AA simulations with CG ones.

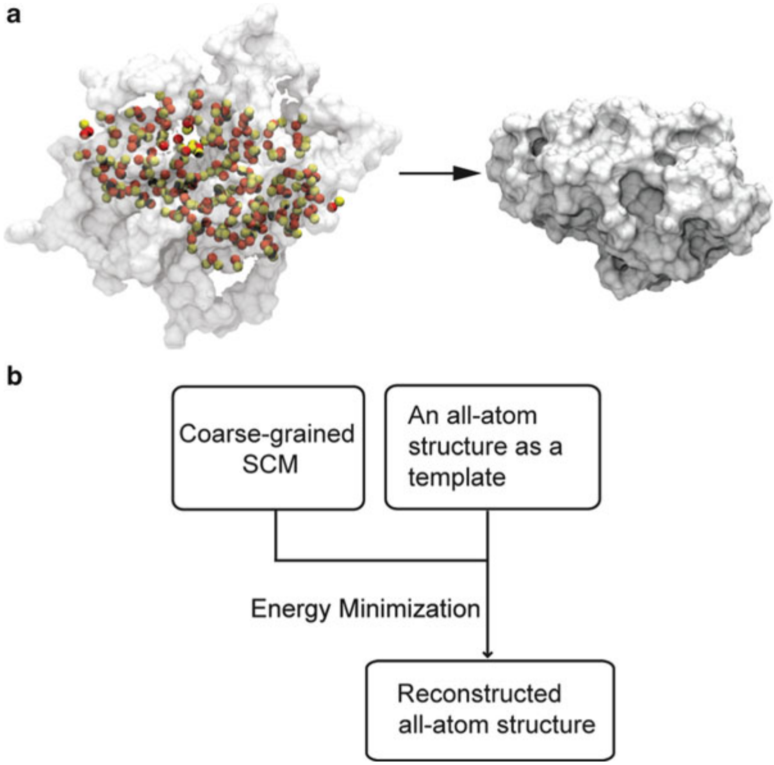

a
b
Coarse-grained SCM
An all-atom structure as a template
Energy Minimization
Reconstructed all-atom structure

Fig. 8.4 A schematic representation of the SCAAL reconstruction method with the use of an all-atomistic protein structure as a template and the positions of coarse-grained side-chain-C$_\alpha$ model (SCM) as harmonic constraints. (**a**) (*Left*) C$_\alpha$ beads are in *red* and the heavy side-chain beads are in *yellow*. The two beads hold the positions through harmonic constraints for a projected reconstructed all-atomistic protein model. A randomly chosen all-atomistic protein structure that can be far from the crystal structure is introduced as a structural template and shown in a solvent accessible surface area mode. (*Right*) After the structural reconstruction by SCAAL, an all-atomistic representation of a projected protein structure is created (myoglobin, PDBID 1A6M, is used for illustration). (**b**) Flow chart of the SCAAL algorithm

This combined multi-scale scheme “MultiSCAAL” builds on the capabilities of the modified SCM which can handle different solvent and environment condition and on the accurate reconstruction of all-atom protein structures from SCM provided by SCAAL. Both these steps are necessary to incorporate crowding and chemical interferences in a multi-scale molecular simulation.

The MultiSCAAL scheme works on enhancing the sampling of all-atomistic simulations by utilizing a large set of initial conditions sampled from the SCM distributions. These selected initial CG structures are reconstructed into AA ones using SCAAL. Then we let the all-atom simulation visit and refine all the conformations that are predicted by the more efficient SCM model. Our scheme is not based on the concept of Resolution Exchange. Thus, we don’t perform any conformation exchanges between the CG and AA simulations. Instead, we concentrate on the proper selection of initial AA conformations based on a knowledge-based CG model that can be adjusted to different environmental conditions.

In summary, the MultiSCAAL scheme follows these steps:

(1) The energy function for SCM molecular dynamics simulations is derived from the potential of mean force (PMF) from the all-atomistic simulations that contain certain chemical interference using Boltzmann inversion method.
(2) SCM protein representations in a thermodynamic ensemble of interest are selected according to a Metropolis criterion [52] and all-atomistic protein conformation are promptly reconstructed using SCAAL.
(3) Folding free energy landscape of a protein is effectively simulated by all-atomistic molecular dynamics that uses reconstructed all-atomistic protein models built from step (2) as initial conformations.

8.4 Protein Folding in Different Conditions: Examples

8.4.1 Crowding and Protein Folding

The living cell is a highly crowded environment due to the presence of large amounts of soluble and insoluble macromolecules, including proteins, nucleic acids, ribosomes, and carbohydrates. This cellular crowding limits the available space for biochemical interactions including protein folding. It is estimated that the concentration of macromolecules in the cytoplasm is in the range of 80–400 mg/ml which amounts to a volume fraction between 10 and 40 % [53–56]. Crowding can be mimicked experimentally by adding high concentrations of inert synthetic crowders. In addition, crowding can be modeled using CG molecular dynamics simulations. There are established effects of crowding on protein folding such that crowding stabilizes the folded protein, compacts denatured states. These effects have been investigated using different theoretical and experimental techniques [6, 8, 36, 53, 57, 58]. Here, we present examples of other interesting effects of macromolecular crowding on protein folding. These studies utilized the power and efficiency of CG simulations based on the SCM model.

8.4.1.1 Crowding Changes Protein Shape

SCM based molecular dynamics [7] simulations were used to investigate the secondary structure changes in protein *Borrelia burgdorferi* VlsE in experimental crowded conditions [59]. VlsE is an aspherical protein with marginal stability: It is best described as having an elongated football shape with a helical core surrounded by floppy loops at each end [60]. Experiments using Ficoll 70 as an inert synthetic crowding agent have shown that VlsE folded state is stabilized in the presence of increasing concentration of crowders. However, when the same crowding experiments were repeated in the presence of urea, crowding seemed to destabilize the folded state.

In order to understand these varying effects of crowding on the folding of VlsE, CG molecular simulations were used to calculate energy landscape of VlsE in different volume fractions of Ficoll 70 and at different temperatures. VlsE was modeled using SCM with nonspecific interactions. Ficoll 70 molecules are modeled as hard spheres that provide nonspecific repulsive interactions in the simulations. The thermodynamic properties of VlsE in aqueous solvent and in crowded environments (volume fractions, ϕ_c, of 0, 15, and 25 %) were studied by molecular simulations with Langevin dynamics. The replica exchange method (REM) [61, 62] was used in order to enhance the efficiency of sampling. The resulting trajectories were analyzed using the weighted histogram analysis method (WHAM) [63, 64].

The resulting energy landscape is shown in Fig. 8.5. This energy landscape shows that the combination of crowding and denaturing agents (temperature in simulations versus urea in experiments) can produce conformational changes in VlsE between three dominant states. These three states are the native structure (football shaped), a bean-like structure, and a collapsed globular structure. The all-atomic structures for these three states were reconstructed using SCAAL as shown in Fig. 8.6. The simulations have also shown that these conformational (shape) changes were accompanied by secondary structure transformations that lead to the exposure of a hidden antigenic region in agreement with experiments.

8.4.1.2 Crowding and Protein Folding Routes

The folding energy landscape of an α/β protein, apoflavodoxin, in the presence of inert macromolecular crowding agents was studied using *in silico* and *in vitro* approaches [65]. The crowding conditions were created using two crowding agents with different shapes, the spherical Ficoll 70 and the rod-like dextran. Parallel kinetic folding experiments were performed on purified apoflavodoxin in the presence of Ficoll 70 and dextran. These experiments have shown that time-resolved folding pathway of apoflavodoxin is modulated by crowding agent geometry.

In the CG molecular simulations, apoflavodoxin was constructed using the SCM model with a Gō-like Hamiltonian. Ficoll 70 was modeled as hard sphere. The rod-like dextran was modeled as dumbbell consisting of two bonded hard spheres (Ficoll 70). As with VlsE above, Langevin dynamics, REM, and WHAM were used. The results of the simulations showed that these different types of crowders stabilize the native state of apoflavodoxin (Fig. 8.7). In addition, the geometry of the crowder tends to play an important role in manipulating the folding route. The simulations show that the early formation of contacts around the β_1 sheet of apoflavodoxin creates a topological frustrated structure. In order for the protein to proceed in its folding, it has to unfold and undo these early formed contacts. This topological frustration is affected by the crowded environment. More specifically, the shape of the crowder can worsen or remedy the early topological frustration as can be seen in Fig. 8.8.

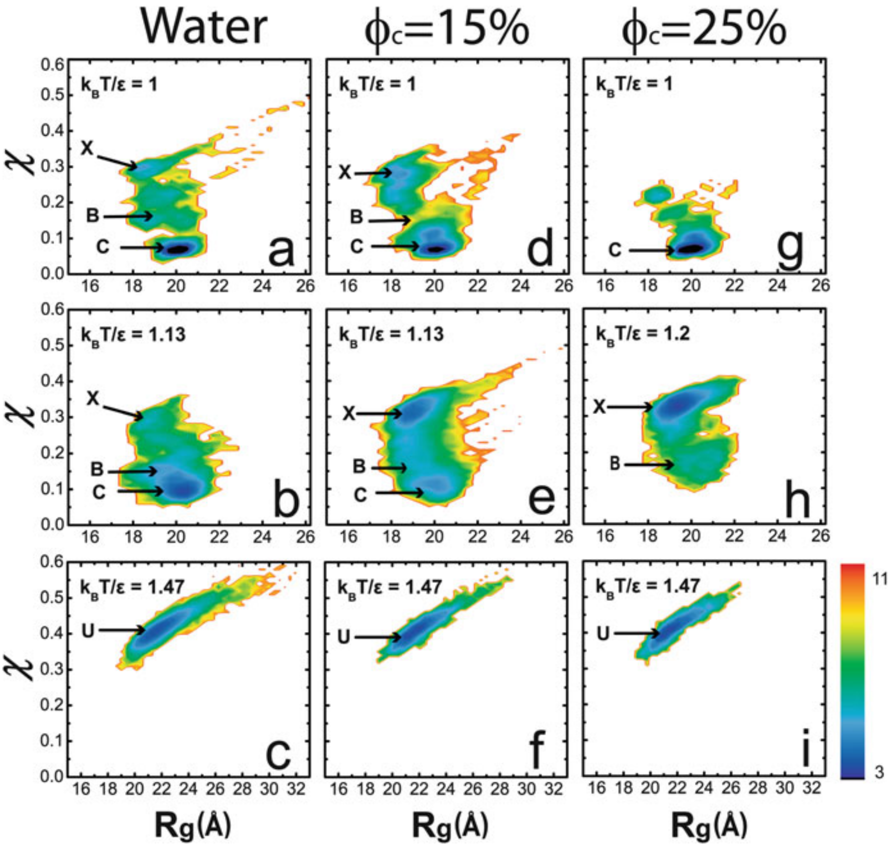
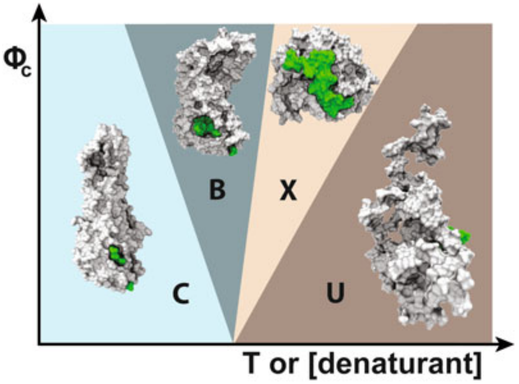


Fig. 8.5 Free-energy diagram as a function of radius of gyration R_g and the overlap function (χ) for $\varphi_c = 0$ % (water) (**a**, **b**, and **c**), 15 % (**d**, **e**, and **f**), 25 % (**g**, **h**, and **i**) at various temperatures expressed in k_BT/ε. χ measures the deviation from crystal structure (χ=0). The *color* is scaled by k_BT. The native *football-shaped* species is labeled *C*, the *bean structure* is labeled *B*, the *spherical state* is named *X*, and the unfolded state is indicated by *U*

Fig. 8.6 A schematic phase diagram of VlsE conformations in the φ_c–T (or urea) plane. The antigenic IR6 sequence is shown in *green* for all representative states *C*, *B*, *X*, and *U*



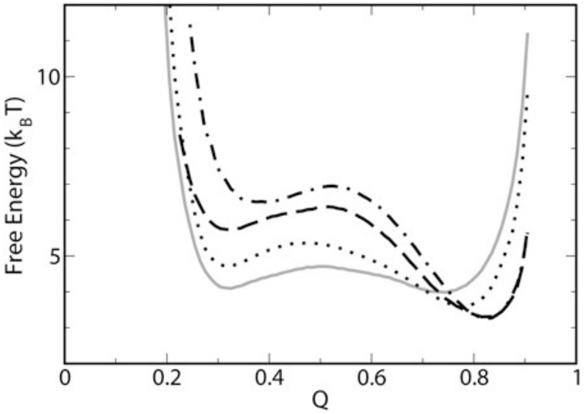


Fig. 8.7 Free energy profiles are plotted as a function of Q (the fraction of native contact formation) at different crowding conditions at 360 K. φ_c (water) = 0, *solid line*; φ_c (Ficoll70) = 25 %, *dotted line*; φ_c (Ficoll70) = 40 %, *dashed line*; and φ_c (dumbbell) = 40 %, *dot-dashed line*

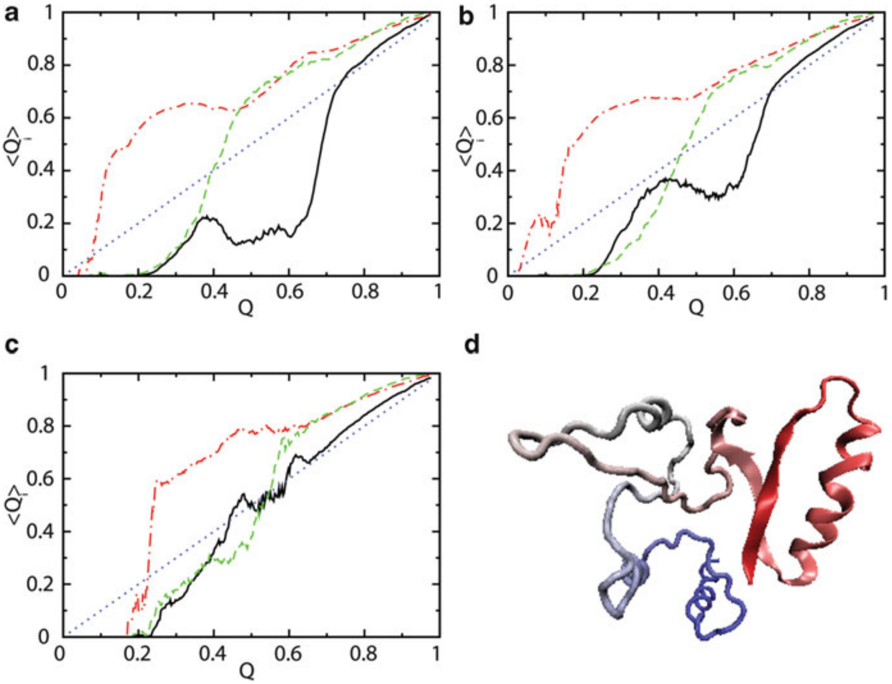


Fig. 8.8 Probability of select native contact formation $<Q>_i$ at the *i*th region of a protein in the evolution of protein folding. Contact formation of the first β-strand (*black*), the first α-helix (*red*), and the third β-strand (*green*) is plotted as a function of Q in (**a**) water, (**b**) φ_c =40 %, Ficoll70, and (**c**) φ_c =40 %, dumbbell-like crowding agent, respectively. (**d**) A conformation in the unfolded state with some contacts formed about β_1 in early Q that causes topological frustrations in the folding landscape. The *diagonal line* is provided as a visual guidance for a mean-field like behavior

8.4.2 Multi-scale Simulation of Protein Folding with Chemical Interference

8.4.2.1 Protein Folding in Urea

The mutliscale simulations using MultiSCAAL were used in order to investigate the effect of urea on the folding landscape of Trp-cage [16]. In this approach, CG simulations of Trp-cage in urea were performed first and then structures fished from these simulations are fed into AA simulations in order to zoom in on important details. In order to perform the CG simulation, statistical potential maps of amino acid LJ parameters were created for different concentrations of urea. These maps were created using the Boltzmann Inversion technique presented above.

The results obtained from the MultiSCAAL simulations were compared with those of AA simulations performed in the same study. The AA atom simulation of Trp-cage utilized the enhanced sampling technique of Replica Exchange Method (REM). AA-REM and MultiSCAAL simulations were performed in aqueous and 8 M urea solvent conditions. MultiSCAAL were shown to be more accurate and more efficient that AA-REM.

In terms of accuracy, MultiSCAAL samples a broader energy landscape, with a wide distribution of ensemble structures as can be seen in Fig. 8.9. Interestingly, in the case of 8 M urea the dominant structure sampled by MultiSCAAL matches better with interatomic distances obtained by NMR experiments [66]. By using a reduced representation in side-chain beads in the CG model, without explicit solvent molecules, the protein can explore different side-chain orientations faster. This allows the indole group of Trp 6 to exit the hydrophobic core of the protein and this structural feature can account for the shorter distances between Trp 6 and other amino acids.

In terms of efficiency, MultiSCAAL simulation was shown to provide a considerably enhanced sampling efficiency and lower computational cost than the standard AA-REMD simulations with the total simulation length being $\sim$25 times greater in less computational hours ($<$1/2).

8.4.2.2 Protein Folding and Ionic Concentration

Calmodulin (CaM) is the smallest known functional protein and plays an important role in regulating intercellular signaling. CaM possesses a great conformational flexibility as it can bind over 300 targets when fully saturated with calcium [67]. SCM based coarse-grained simulations were used to study the crowding effects on the conformational states of apoCaM [68]. In addition, these calculations were extended using a multi-scale approach to include electrostatics in studying the conformational states of both apoCaM and holoCaM at different salt concentrations

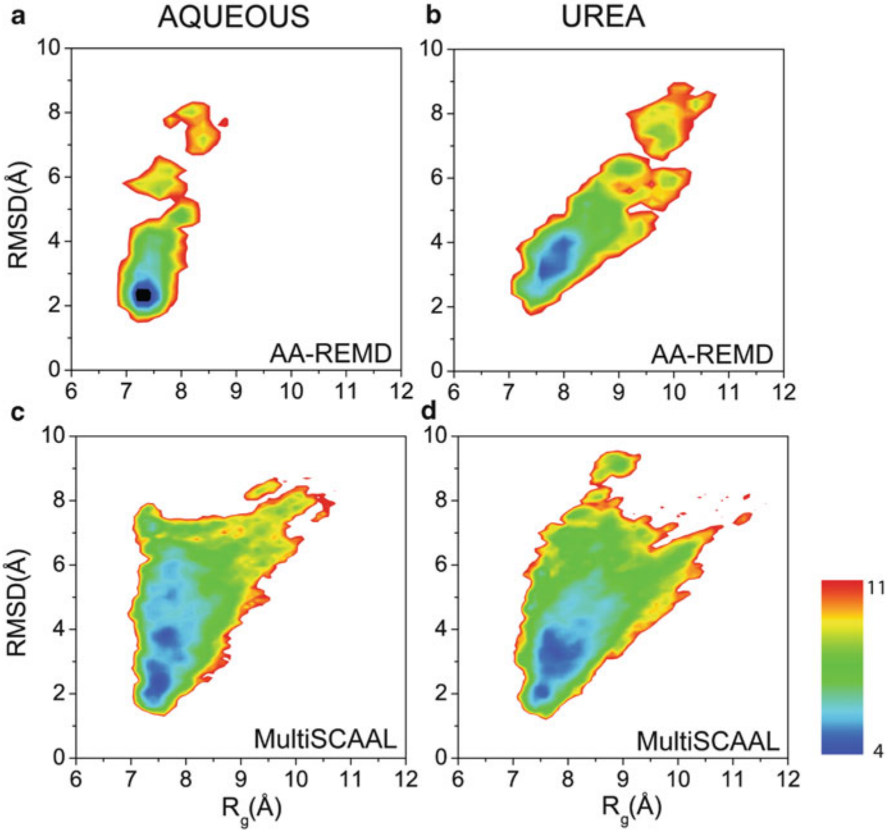


Fig. 8.9 Two–dimensional free energy landscape for Trp-cage as a function of the radius of gyration (R_g) and the root-mean-square-deviation (RMSD) under (**a**, **c**) aqueous and (**b**, **d**) urea conditions based on two different simulation schemes at 300 K: (**a**, **b**) simulations using AA-REMD; (**c**, **d**) simulations using MultiSCAAL. The free energy is *colored* by $k_B T$

in crowded environment [69]. This was done by developing a unique multi-scale solution of charges computed from quantum chemistry, together with SCAAL protein reconstruction, SCM coarse-grained molecular simulations, and statistical physics, to represent the charge distribution in the transition from apoCaM to holoCaM upon calcium binding. The simulations were performed at different salt concentrations, different volume fraction of crowding agents, and a combination of both. These simulations showed that increased levels of macromolecular crowding, in addition to calcium binding and ionic strength typical of that found inside cells, can impact the conformation, secondary structure and the EF hand orientation of CaM [69].

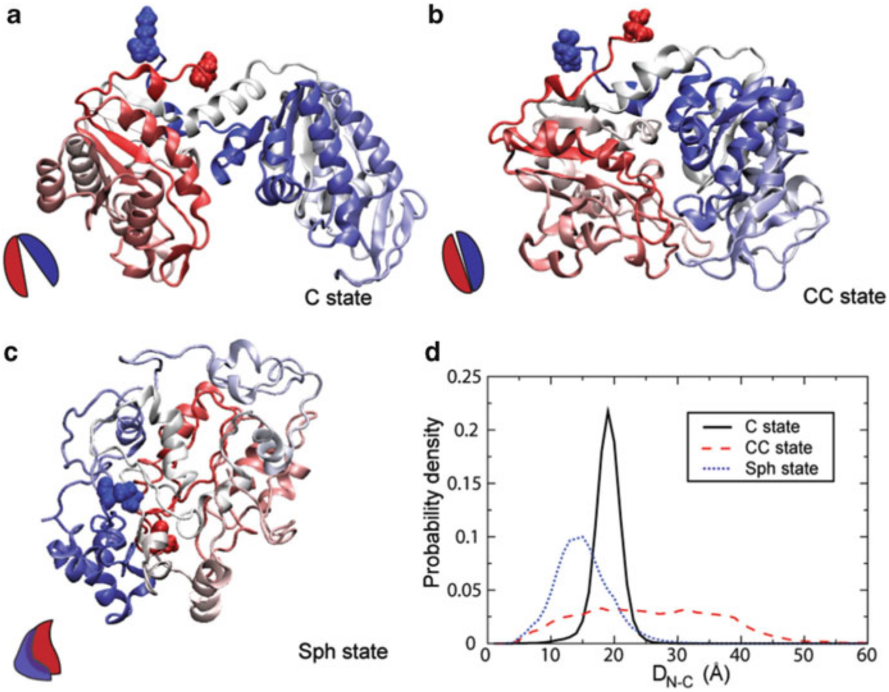


Fig. 8.10 Structural characteristics of the dominant compact ensemble structures of PGK in cartoon representation. (**a**) Crystal state C, (**b**) Collapsed crystal state CC and (**c**) Spherical state Sph. The coloring of each protein model ranges from N-terminus (*red*) to C-terminus (*blue*). The -N and -C termini are represented with Van der Waals spheres. The schematic representation at the *bottom left* of each panel is to address a simplistic view of the arrangement of the N- and C-lobes in each conformation. **(d)** The probability distribution of the distance between N- and C- termini of the three dominant structures of PGK under the condition when each prevails in the simulations. C state (*solid black*), Collapsed Crystal CC (*dashed red*) and Spherical Sph (*dotted blue*)

8.4.3 Other Applications

SCM coarse-grained molecular dynamics simulations were used to investigate the effect of macromolecular crowding on the folding and enzymatic activity of phosphoglycerate kinase (PGK) [70]. Experiments suggested that PGK in a crowded medium adopts conformations that are not seen in dilute conditions. In addition, crowding was shown to enhance the enzymatic activities of PGK by more than 15 times. In the SCM coarse-grained molecular simulations, three possible compact ensembles of PGK were identified as shown in Fig. 8.10. These results suggest that rather than undergoing a hinge motion, the ADP and substrate sites at the inner parts of two domains of PGK are already located in proximity in compact form under crowded or even *in vivo*.

SCM coarse-grained simulations were also used to investigate the competing effects of crowding and urea on the folding of protein Trp-cage [71]. This study shows that crowding enhancement of folding rates of Trp-cage is most pronounced for extended conformations of Trp-cage in the presence of high concentrations of urea.

Finally, a new algorithm was recently added to SCM in order to extend its capabilities to deal with more realistic crowded conditions [72]. This self-assembled clustering algorithm (CGCYTO) was used to produce a polydisperse (PD) coarse-grained model for *E. coli* cytoplasm. It is shown by SCM coarse-grained molecular simulations that the folding temperature of a test protein apoazurin in a PD cytoplasm model is ∼5° greater than that in a Ficoll 70 model [72].

8.5 Conclusion

This chapter presented some of the recent developments in coarse-grained (CG) molecular dynamics techniques when it applies to the problem of protein folding in varying crowding and solvent conditions. We mainly focused on the evolving (Side-chain-C_α Model, [15]) SCM-based techniques. SCM molecular simulations were used to study the protein folding dynamics in crowded conditions that mimic the highly condensed cellular cytoplasm. In these studies, the computational efficiency of simulations based on a minimalist model is utilized to incorporate the additional crowding particles. Several studies have used the SCM simulations to model different types, shapes, and concentration of crowders. SCM simulations achieved a great success in explaining and predicting the behavior of protein folding dynamics in crowded medium as can be seen in the example studies discussed in this chapter.

Additional techniques can extend the capabilities of a CG model to address different types of environmental conditions such as solvent, denaturants, and ions. Several examples of these techniques were presented in this chapter in addition to some applications of SCM-based simulations. A growing trend now in computational studies is to design a multi-scale approach to simulate biophysical systems. This approach tries to combine the advantages of both the more detailed atomic simulations with the efficiency of coarse-grained ones. The chapter presented an example of these multi-scale approaches, MultiSCAAL. MultiSCAAL uses CG simulations in order to speed up and expand the sampling of the all-atom protein folding landscape. All the techniques and the examples discussed here show that well-designed coarse-grained molecular simulations can be a great tool in addressing complicated problems such as protein folding. With the new emerging techniques and with the help of coarse-grained models we can achieve significant progress in understanding complicated systems, especially when they are coupled with experimental methods or with higher resolution (All-atom or Quantum) simulations.

References

1. Rahman A (1964) Correlations in the motion of atoms in liquid argon. Phys Rev 136:405–411
2. McCammon A, Gelin B, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–590
3. Buchner GS, Murphy RD, Buchete NV, Kubelka J (2011) Dynamics of protein folding: probing the kinetic network of folding-unfolding transitions with experiment and theory. Biochim Biophys Acta-Proteins Proteomics 1814:1001–1020
4. Scheraga HA, Khalili M, Liwo A (2007) Protein-folding dynamics: overview of molecular simulation techniques. Annu Rev Phys Chem 58:57–83
5. Dror RO, Dirks RM, Grossman JP, Xu HF, Shaw DE (2012) Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys 41, D. C. Rees, Ed., ed Palo Alto: Annual Reviews, pp 429–452
6. Stagg L, Zhang SQ, Cheung MS, Wittung-Stafshede P (2007) Molecular crowding enhances native structure and stability of alpha/beta protein flavodoxin. Proc Natl Acad Sci USA 104: 18976–18981
7. Homouz D, Perham M, Samiotakis A, Cheung MS, Wittung-Stafshede P (2008) Crowded, cell-like environment induces shape changes in aspherical protein. Proc Natl Acad Sci USA 105: 11754–11759
8. van den Berg B, Wain R, Dobson CM, Ellis RJ (2000) Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. EMBO J 19:3870–3875
9. Ai X, Zhou Z, Bai Y, Choy W-Y (2006) 15N NMR spin relaxation dispersion study of the molecular crowding effects on protein folding under native conditions. J Am Chem Soc 128:3916–3917
10. Charlton LM, Barnes CO, Li C, Orans J, Young GB, Pielak GJ (2008) Residue-level interrogation of macromolecular crowding effects on protein stability. J Am Chem Soc 130:6826–6830
11. Sasahara K, McPhie P, Minton AP (2003) Effect of dextran on protein stability and conformation attributed to macromolecular crowding. J Mol Biol 326:1227–1237
12. Kozer N, Kuttner YY, Haran G, Schreiber G (2007) Protein-protein association in polymer solutions: from dilute to semidilute to concentrated. Biophys J 92:2139
13. Snoussi K, Halle B (2005) Protein self-association induced by macromolecular crowding: a quantitative analysis by magnetic relaxation dispersion. Biophys J 88:2855–2866
14. Rivas G, Fernández JA, Minton AP (2001) Direct observation of the enhancement of noncooperative protein self-assembly by macromolecular crowding: indefinite linear self-association of bacterial cell division protein FtsZ. Proc Natl Acad Sci USA 98:3150–3155
15. Cheung MS, Finke JM, Callahan B, Onuchic JN (2003) Exploring the interplay between topology and secondary structural formation in the protein folding problem. J Phys Chem B 107:11193–11200
16. Samiotakis A, Homouz D, Cheung MS (2010) Multiscale investigation of chemical interference in proteins. J Chem Phys 132:175101
17. Anfinsen CB, Haber E, Sela M, White FH (1961) Kinetics of formation of native ribonuclease during oxidation of reduced polypeptide chain. Proc Natl Acad Sci USA 47:1309–1314
18. Levinthal C (1968) Are there pathways for protein folding? J Chim Phys Phys-Chim Biol 65:44
19. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289
20. Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 48:545–600
21. Leopold PE, Onuchic JN, Montal M (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci USA 89:8271–8275
22. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4:10–19
23. Bryngelson JD, Wolynes PG (1987) Spin-glasses and statistical mechanics of protein folding. Proc Natl Acad Sci USA 84:7524–7528
24. Go N (1983) Theoretical studies of protein folding. Annu Rev Biophys Bioeng 12:183–210

25. Baker D (2000) A surprising simplicity to protein folding. Nature 405:39–42
26. Socci ND, Onuchic JN (1994) Folding kinetics of protein like heteropolymers. J Chem Phys 101:1519–1528
27. Taketomi H, Ueda Y, Gō N (1975) Studies on protein folding, unfolding and fluctuations by computer simulations. Int J Pept Protein Res 7:445–459
28. Dill KA (1985) Theory for the folding and stability of globular proteins. Biochemistry 24:1501–1509
29. Ueda Y, Taketomi H, Gō N (1978) Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme. Biopolymers 17:1531–1548
30. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 298:937–953
31. Thirumalai D, Guo Z (1995) Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labeling experiments. Biopolymers 35:137–140
32. Wolynes PG, Onuchic JN, Thirumalai D (1995) Navigating the folding routes. Science 267:1619–1620
33. Shea J-E, Onuchic JN, Brooks CL III (1999) Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. Proc Natl Acad Sci USA 96:12512–12517
34. Betancourt MR, Thirumalai D (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 8:361–369
35. Miyazawa M, Jernigan RL (1985) Estimation of interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552
36. Cheung MS, Klimov D, Thirumalai D (2005) Molecular crowding enhances native state stability and refolding rates of globular proteins. Proc Natl Acad Sci USA 102:4753–4758
37. Cheung MS, Thirumalai D (2006) Nanopore-protein interactions dramatically alter stability and yield of the native state in restricted spaces. J Mol Biol 357:632–643
38. Cheung MS, Chavez L, Onuchic JN (2004) The energy landscape for protein folding and possible connections to functions. Polymer 45:547–555
39. Hyeon C, Thirumalai D (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models. Nat Commun 2:487
40. Veitshans T, Klimov D, Thirumalai D (1997) Protein folding kinetics: timescales, pathways, and energy landscapes in terms of sequence-dependent properties. Fold Des 2:1–22
41. Kolinski A, Skolnick J (1992) Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. J Chem Phys 97:9412–9426
42. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213:859–883
43. Sope AK (1996) Empirical potential Monte Carlo simulation of fluid structure. Chem Phys 202:295–306
44. Reith D, Putz M, Muller-Plathe F (2003) Deriving effective mesoscale potentials from atomistic simulations. J Comput Chem 24:1624–1636
45. Betancourt MR, Omovie SJ (2009) Pairwise energies for polypeptide coarse-grained models derived from atomic models. J Chem Phys 130:195103
46. Makowski M, Liwo A, Makowska MKJ, Scheraga HA (2007) Simple physics-based analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water. 2. Tests with simple spherical systems. J Phys Chem B 111:2917–2924
47. Debye P, Hückel E (1923) The theory of electrolytes. I. Lowering of freezing point and related phenomena. Physikalische Zeitschrift 24:185–206
48. Gront D, Kmiecik S, Kolinski A (2007) Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J Comput Chem 28:1593–1597

49. Canutescu A, Shelenkov A, Dunbrack R (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12:2001–2014
50. Rotkiewicz P, Skolnick J (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. J Comput Chem 29:1460–1465
51. Heath AP, Kavraki LE, Clementi C (2007) From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. Proteins Struct Funct Bioinform 68:646–661
52. Frenkel D, Smit B (2001) Understanding molecular simulation: from algorithms to applications. Academic Press, San Diego, CA
53. van den Berg B, Ellis RJ, Dobson CM (1999) Effects of macromolecular crowding on protein folding and aggregation. EMBO J 18:6927–6933
54. Rivas G, Ferrone F, Herzfeld J (2004) Life in a crowded world. EMBO Rep 5:23–27
55. Record MT, Courtenay ES, Cayley S, Guttman HJ (1998) Biophysical compensation mechanisms buffering E. coli protein-nucleic acid interactions against changing environments. Trends Biochem Sci 23:190–194
56. Ellis RJ, Minton AP (2003) Cell biology – join the crowd. Nature 425:27–28
57. Minton AP (2005) Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited. Biophys J 88:971–985
58. Zhou HX, Dill KA (2001) Stabilization of proteins in confined spaces. Biochemistry 40: 11289–11293
59. Perham M, Stagg L, Wittung-Stafshede P (2007) Macromolecular crowding increases structural content of folded proteins. FEBS Lett 581:5065–5069
60. Eicken C, Sharma V, Klabunde T, Lawrenz MB, Hardham JM, Norris SJ et al (2002) Crystal structure of Lyme disease variable surface antigen VlsE of *Borrelia burgdorferi*. J Biol Chem 277:21691–21696
61. Sanbonmatsu KY, Garcia AE (2002) Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. Proteins 46:225–234
62. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
63. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM (1992) The weighted histogram analysis method for free-energy calculations on biomolecules I. The method. J Comput Chem 13:1011–1021
64. Chodera JD, Swope WC, Pitera JW, Seok C, Dill KA (2007) Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. J Chem Theory Comput 3:26–41
65. Homouz D, Stagg L, Wittung-Stafshede P, Cheung MS (2009) Macromolecular crowding modulates folding mechanism of alpha/beta protein apoflavodoxin. Biophys J 96:671–680
66. Mok KH, Kuhn LT, Goez M, Day I, Lin J, Andersen NH et al (2007) A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein. Nature 447:106–109
67. Means AR, Dedman JR (1980) Calmodulin – an intracellular calcium receptor. Nature 285:73–77
68. Homouz D, Sanabria H, Waxham MN, Cheung MS (2009) Modulation of calmodulin plasticity by the effect of macromolecular crowding. J Mol Biol 391:933–943
69. Wang Q, Liang K-C, Czader A, Waxham MN, Cheung MS (2011) The effect of macromolecular crowding, ionic strength and calcium binding on calmodulin dynamics. PLoS Comput Biol 7:e1002114
70. Dhar A, Samiotakis A, Ebbinghaus S, Nienhaus L, Homouz D, Gruebele M et al (2010) Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding. Proc Natl Acad Sci USA 107:17586–17591
71. Samiotakis A, Cheung MS (2011) Folding dynamics of Trp-cage in the presence of chemical interference and macromolecular crowding. I. J Chem Phys 135:175101-175101-16
72. Wang Q, Cheung M (2012) A physics-based approach of coarse-graining the cytoplasm of E. coli Biophys J 102:2353–2361