The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules

Jeffrey K. Noel and José N. Onuchic

1 Introduction

Structural biology techniques, such as nuclear magnetic resonance (NMR), x-ray crystallography, and cryogenic electron microscopy (cryo-EM), have provided extraordinary insights into the details of the functional configurations of biomolecular systems. Recent advances in x-ray crystallography and cryo-EM have allowed for structural characterization of large molecular machines such as the ribosome, proteasome, and spliceosome. This deluge of structural data has been complemented by experimental techniques capable of probing dynamic information, such as Förster resonance energy transfer (FRET) and stopped flow spectrometry. While these experimental studies have provided tremendous insights into the dynamics of biomolecular systems, it is often difficult to combine the low resolution dynamical data with the high-resolution structural data into a consistent picture. Computer simulation of these biomolecular systems bridges static structural data with dynamic experiments at atomic resolution (Fig. 1).

Since the first molecular dynamics simulations of bovine pancreatic trypsin inhibitor 35 years ago [38], molecular simulations have become indispensable tools in biophysics. Molecular dynamics simulations of biomolecules treat the molecule as a collection of classical particles interacting through a potential energy function called a force field [1]. The molecule's dynamics are propagated through time by

J.K. Noel
Department of Physics and Center for Theoretical Biological Physics,
University of California, La Jolla, CA 92093, USA
e-mail: jknoel@ucsd.edu

J.N. Onuchic (✉)
Department of Physics and Center for Theoretical Biological Physics,
Rice University, Houston, TX 77005, USA
e-mail: jonuchic@rice.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_2,

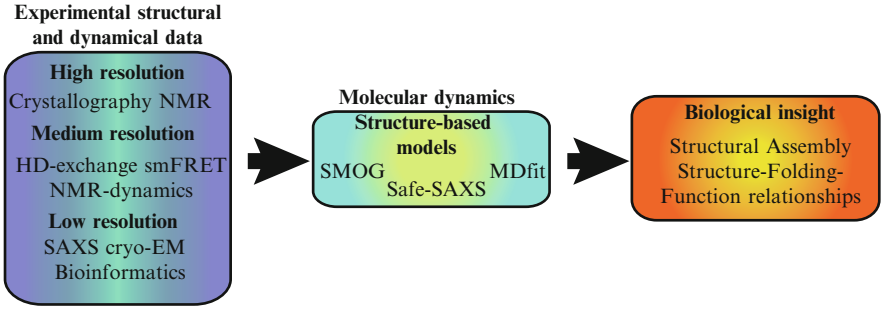

Fig. 1 Structure-based models bridge static high-resolution structural data with lower resolution dynamical and structural data at the all-atom level. Many experimental inputs can be combined to form a coherent picture of a biological process

numerical integration of Hamilton's equations resulting in a molecular trajectory. This trajectory can be used to gain a kinetic and thermodynamic understanding of the system. Simulations can be performed using empirically parameterized force fields that include explicit solvent. In principle, the chemistry-based representation should reproduce the structure and dynamics of a biomolecular system without requiring input from experimental structural data. In practice, making contact with experimental observables poses harsh challenges for these force fields both due to the level of accuracy required and the long time scales needed [54, 66]. In order to integrate experimental data in a consistent manner, biomolecular models with robust potential energy functions able to access long time scales are necessary. The energy landscape theory of protein folding provides the theoretical underpinning for *structure-based models* (SBM) [47]. These models impose a *native bias* by explicitly including structural data in the Hamiltonian. The structural data is derived from experimental techniques that are able to discern a representative structure of a molecule in a deep free energy basin, e.g., a protein native state. The native bias dramatically reduces the complexity of the resulting force field. These simplifications allow for a clear physical understanding of a system and open up biologically relevant timescales while retaining the essential dynamical features. SBM have been validated by their application to protein dynamics, such as folding, stretching, oligomerization, and functional transitions. Multiple experimental inputs can be naturally included, e.g., by extending the single native bias to include information from multiple conformers to explore conformational transitions. Fueled by the introduction of an all-atom (AA) SBM, prospective new applications for SBM are being explored in areas such as RNA folding, molecular machines, and prediction of protein–protein interactions. This chapter will present the basics of SBM and explain how a publicly available SBM, SMOG (Structure-based MOdels in GROMACS http://smog.ucsd.edu), has been used to explore the dynamics of systems as disparate as folding knots in proteins and accommodation in the ribosome.

2 Structure-Based Models

2.1 Foundations in Energy Landscape Theory

The inclusion of a native bias, the hallmark of a SBM, has a rigorous footing in the energy landscape theory of protein folding [8, 33, 47]. Protein folding is a self-organizing process whereby a protein transitions from a highly disordered ensemble (unfolded) to a structured ensemble (folded/native state). The relatively short timescale with which the folded state is reached implies that any competing nonnative states (traps) are shallow compared with the overall energy bias to folding. If these traps are sufficiently shallow, the nonnative interactions can be grouped into an effective diffusion [9, 17]. In addition, the uniqueness of the folded state implies that it corresponds to the global minimum in the free-energy landscape. The *principle of minimal frustration* states that evolution has achieved this folding robustness by selecting for sequences where the interactions present in the native structure are mutually supportive, i.e., attractive. The interactions are minimally frustrated or, in other words, maximally consistent. This organization leads to the protein folding on a *funneled landscape* where the energy on average decreases as it forms structures similar to the native structure.

Minimal frustration and the funneled energy landscape give the theoretical foundation for SBMs. A structure-based potential dramatically reduces the biomolecular Hamiltonian's complexity by stabilizing interactions that are spatially close in the native configuration. While real protein funnels have residual energetic frustration caused by nonnative interactions, the SBMs discussed here are "perfectly funneled" models, since in the force field *all* interactions stabilize the native structure. Nonnative interactions are strictly repulsive. In such a framework, any barriers to folding must be free energy barriers arising from the various ways energy and entropy compensate during folding. The ability of perfectly funneled models to reproduce experimental folding trends and mechanisms shows that geometrical effects like chain connectivity have an enormous influence on protein dynamics [5, 11, 47]. Since the precise energetics are secondary to the geometry of the protein molecule, this idea leads to the commonly held notion that geometry determines the folding mechanism.

Even though SBMs were formulated in the context of protein folding, their applications are widespread. Folding is only a first step in the lives of proteins which go on to perform a myriad of functions in the cell. The funneled energy landscape upon which the protein folds is the same landscape that controls functional protein motions. Multiple functional conformational states captured by experiment can be naturally included by extending the funneled landscape to have multiple basins. Structured RNAs must also have evolutionary pressure to reduce the level of frustration or they would encounter their own "Levinthal's paradox." The robust dynamics of large molecular complexes such as the ribosome and proteasome must

depend even less on the precise atomic energetic details and more on emergent properties controlled by the geometry of their constituents. While all these systems will have residual levels of frustration, the use of SBMs as a baseline is crucial to partition the global properties, those largely dependent on structure, from the details dependent on specific energetics.

2.2 Structure-Based Model as a Baseline

Simplified models have a long history of elucidating the organizing principles governing complex systems. A key question is how sensitive a model is to its underlying parameters. Determining the correct value for a parameter is often equally important as understanding the sensitivity to perturbations in that parameter. Since molecular geometry has a central influence on the motions leading to molecular function, simplified models based on low free energy structures are a natural starting point. The simplest models look at the normal modes of an energy landscape created by replacing all short range interactions in a native structure by Hookean springs [61]. These models can capture relevant rigid body motions. SBMs provide an important generalization by allowing the possibility for "cracking," [24, 25, 40, 68] allowing interactions to break and reform, since the springs are replaced by short range potentials. Thus, SBM can capture motion on all scales from native basin dynamics to unfolding.

The straightforward formulation of a structure-based potential allows for sensitivity analysis of the force field parameters [69] and their simplicity makes them extremely fast to compute. The force field is readily extensible allowing the introduction of complicated effects to be explored parametrically. For example, the effects of electrostatics can be explored by perturbative addition of Coulomb interactions [4,14,35], or the effects of solvent probed by the perturbative addition of desolvation barriers [12]. A crucial question in the protein folding field has been how proteins manage to achieve such smooth energy landscapes, or equivalently, why do AA empirical force fields and structure prediction schemes have difficulty achieving the level of specificity seen in proteins? Using structure-based potentials with AA geometries, we can begin to address this question. These models completely partition energetic effects from geometric effects, and through careful investigation, may discern to what extent energetics contribute to the apparent native specificity in protein structure, folding, and function. While processes like the formation of nonnative intermediates during folding [18, 53, 60] and protein misfolding are clearly cases that perfectly funneled SBM will be unable to fully describe, through adding complexity in a piecemeal fashion to a robust baseline model, a more complete understanding of the interplay between geometry and energy in even these complicated systems will result.

3 Implementation of Structure-Based Models

SBMs have a long history in the protein folding field. The folding dynamics of minimally frustrated sequences were first tested in lattice models. Bryngelson et al. [10] and Socci et al. [56] investigated a minimally frustrated lattice model with three types of beads. They found that the dynamics could be well described by diffusion along a small number of collective coordinates on an effective free energy surface defined by those coordinates. As the structural correspondence between cubic lattices and actual proteins is low, Nymeyer et al. implemented an off-lattice, coarse-grained model of a protein-like structure. They compared the folding dynamics of an energetically frustrated [62] versus a completely unfrustrated β-barrel [45]. They showed that the completely unfrustrated model, effectively a SBM, exhibited the characteristics of a good folder, specifically, having exponential folding kinetics on a funnel-shaped landscape that is robust to reasonable perturbations. Following these successes, Clementi et al. [15] introduced the popular "C_α model," which also had a coarse-grained representation of the protein. This model reproduced the transition-state ensembles (TSE) of several small two- and three-state proteins. The C_α model has since been adopted by several investigators to explore myriad topics in protein folding (see these references for some highlights [2, 11, 12, 22, 26, 28, 29, 52, 59]). The off-lattice geometry allowed clear representation of protein structures, making comparisons to experimentally determined dynamics possible. In order to capture geometric effects like side chain packing, Whitford et al. introduced an AA SBM [69]. This model is being used to represent proteins [69], RNA/DNA [64] and ligands in a consistent fashion for both dynamics [42, 43, 66] and molecular modeling [27, 50, 51]. These two models, AA and C_α, are currently in wide use and are available on the SMOG web server [44].

Before the two available models are described in detail, we review the key components common to any SBM. The defining characteristic is that the parameters are determined from a native structure. The potential V is composed of three contributions,

$$V = \underbrace{V^{\text{Bonded}} + V^{\text{Repulsive}}}_{\text{Maintain geometry}} + \underbrace{V^{\text{Attractive}}}_{\text{Tertiary structure}} . \tag{1}$$

V^{Bonded} includes interactions that maintain the covalently bonded structure and torsional angles. This term also ensures correct chirality. $V^{\text{Repulsive}}$ contains spherically symmetric hard wall repulsions that enforce excluded volume and prevent chain crossings. Collectively, these two terms maintain the protein's structure and allowed conformational diversity. $V^{\text{Attractive}}$ contains short range, attractive interactions between atoms (or residues if coarse graining) close in the native state. These interactions are the core of the SBM and are discussed in the next section.

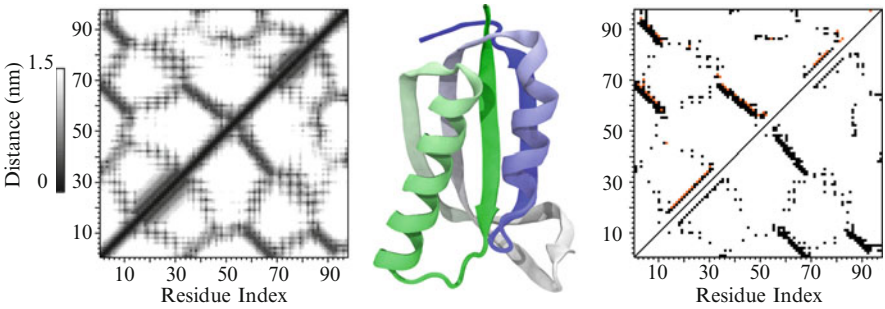

Fig. 2 Native contact map of ribosomal protein S6 (PDB code: 1RIS). Structure of the α/β protein S6 is shown with the N-terminus (residue 1) colored *green*. *Left panel* shows the proximity of the nearest atomic contact for each residue pair up to a maximum of 1.5 nm. *Right panel* compares two coarse-grained native contact maps. A pair of residues are considered a native contact if they share a native atom–atom contact. *Top triangle*: 6Å cutoff. *Bottom triangle*: a 6Å cutoff with geometric occlusion using Shadow [44]. The contacts which are excluded by Shadow are colored *orange*

3.1 Native Contact Map

Atoms that are spatially near in the native state are considered *contacts* and together the set of all contacts composes a *native contact map* (Fig. 2). A contact map is a binary symmetric matrix that encodes which atom pairs ij are given attractive interactions in the SBM potential. In the context of a SBM, the native contact map should approximate the distribution of stabilizing enthalpy in the native state that is provided by short range interactions like van der Waals forces, hydrogen bonding, and salt bridges. Any long range interactions or nonlocal effects are taken into account in a mean field way through the native bias. For example, the hydrophobic effect is encoded by the density of native contacts being larger on the interior of the protein than on the surface.

Methods for constructing contact maps are based on the heavy atom distances in the native structure. There are three widely used techniques: heavy atom cutoffs [16], van der Waals radii overlaps [15, 55, 58], and geometric occlusions [44, 71]. Heavy atom cutoff maps define a cutoff distance R_C, typically 4–6.5Å, and consider all heavy atoms within R_C of each other in contact. van der Waals radii cutoff maps increase the radii of all the heavy atoms by either a multiplicative constant ($\sim$1.25) or an additive constant ($\sim$1.4Å). Any atoms that then overlap are considered to be in contact. The rationale for the multiplicative constant comes from overlapping electron clouds, or "soft spheres." The additive constant represents the size of one water molecule. Half the diameter of water is added to each atomic radius, and if atoms then overlap it means that a water cannot be placed between them. The set of atom pairs excluding water from each other are presumed to interact, and thus considered contacts (the software package CSU [55] uses this approach). Geometric occlusion maps take the output of a heavy atom cutoff contact map, $R_C \gtrsim 6$Å, and then remove any contacts that are geometrically obstructed. $R_C > 4.5$Å introduces

many unphysical or "occluded" contacts where atoms are interacting through an intervening atom. Since these interactions are mostly induced dipole interactions, electron screening effects should dampen the occluded interactions. van der Waals radii overlaps and geometric occlusions both provide the short range, first layer of atomic contacts. Geometric occlusion maps add longer range water- or cofactor-mediated contacts up to the cutoff distance. The advantage of geometric occlusion is that atoms separated by voids, or those coordinated by water and metals not explicitly included in a protein simulation, can be accounted for without introducing spurious occluded contacts.

van der Waals radii overlaps and geometric occlusions provide contact maps that behave similarly in protein-folding simulations. Simulations with these maps consistently predict cooperative, protein-like transitions for globular proteins. They also reproduce thermodynamic folding intermediates for proteins with known intermediates [15]. In the authors' experience, heavy atom cutoff maps are not robust in protein-folding simulations. Short-range cutoffs miss longer range contacts, leaving the contact map sensitive to the precise packing of the native state, and thus overweight regions of the contact map. This reduces the cooperativity of the transition, leading to spurious thermodynamic intermediates. Longer cutoffs reduce the sensitivity to packing by adding larger numbers of contacts, but this introduces many unphysical contacts where atoms are interacting through an intervening atom. This overabundance of contacts, by reducing the relative strength of each individual contact, also tends to decrease cooperativity. SMOG uses a geometric occlusion contact map called Shadow [44] for proteins. On the SMOG server, the default for RNA/DNA systems is a 4Å heavy atom cutoff, but there are indications that Shadow is also sensible for RNA folding.

Single bead per residue coarse-grained contact maps are generally derived from the corresponding atomic structure. Coarse-grained contact maps could conceivably be generated from the coarse-grained structure using C_α–C_α distance cutoffs (generally 7–12Å). Since the coarse-grained structure ignores side chain packing, this metric poorly predicts the enthalpic contributions to the native state [39]. For the C_α model, SMOG considers two residues in contact if they share at least one atomic contact.

3.2 SBM Potential

The SMOG structure-based forcefield is available in two grainings, a coarse-grained (C_α) model [15] and AA model [64, 69].

3.2.1 C_α Model

The C_α model coarse grains the protein as single bead of unit mass per residue located at the position of the α-carbon. $\vec{\mathbf{x}}_\mathbf{0}$ denotes the coordinates (usually obtained

from the Protein Data Bank http://www.rcsb.org) of the native state and any subscript 0 signifies a value taken from the native state. The potential is given by

$$V_{C\alpha}(\vec{\mathbf{x}}, \vec{\mathbf{x_0}}) = \sum_{\text{bonds}} \epsilon_r (r - r_0)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{\text{backbone}} \epsilon_{\text{D}} F_{\text{D}}(\phi - \phi_0) + \sum_{\text{contacts}} \epsilon_{\text{C}} C(r_{ij}, r_0^{ij}) + \sum_{\text{non-contacts}} \epsilon_{\text{NC}} \left(\frac{\sigma_{\text{NC}}}{r_{ij}} \right)^{12}, \tag{2}$$

where the dihedral potential F_{D} is,

$$F_{\text{D}}(\phi) = [1 - \cos(\phi)] + \frac{1}{2}[1 - \cos(3\phi)]. \tag{3}$$

The coordinates $\vec{\mathbf{x}}$ describe a configuration of the α-carbons, with the bond lengths to nearest neighbors r, three body angles θ, four body dihedrals ϕ, and distance between atoms i and j given by r_{ij}. C denotes the contact potentials given to the native contacts (see Sect. 3.2.3). Protein contacts that are separated by less than 3 residues are neglected. Excluded volume is maintained by a hard wall interaction giving the residues an apparent radius of $\sigma_{\text{NC}} = 4$Å. The native bias is provided by using the parameters from the native state $\vec{\mathbf{x}}_{\mathbf{0}}$. Setting the energy scale $\epsilon \equiv k_{\text{B}} T^* = 1$, the coefficients are given the homogeneous values: $\epsilon_r = 100\epsilon$, $\epsilon_\theta = 40\epsilon$, $\epsilon_{\text{D}} = \epsilon_{\text{C}} = \epsilon_{\text{NC}} = \epsilon$.

3.2.2 All-Atom Model

The AA potential is quite similar to the C_α potential, although representing the AA geometry requires some additional terms. In the AA model, all heavy (nonhydrogen) atoms are explicitly represented as beads of unit mass, so each interaction is now between atoms as opposed to residues. Bonds, angles, and dihedrals therefore have their traditional chemical meanings. In each residue, there is an additional backbone dihedral and, except for glycine, many side chain dihedrals. Improper dihedrals maintain backbone chirality and, when necessary, side chain planarity. The AA potential V_{AA} is

$$V_{\text{AA}}(\vec{\mathbf{x}}, \vec{\mathbf{x_0}}) = \sum_{\text{bonds}} \epsilon_r (r - r_0)^2 + \sum_{\text{angles}} \epsilon_\theta (\theta - \theta_0)^2 + \sum_{\text{impropers/planar}} \epsilon_\chi (\chi - \chi_0)^2 + \sum_{\text{backbone}} \epsilon_{\text{BB}} F_{\text{D}}(\phi - \phi_0) + \sum_{\text{sidechains}} \epsilon_{\text{SC}} F_{\text{D}}(\phi - \phi_0) + \sum_{\text{contacts}} \epsilon_{\text{C}} C(r_{ij}, r_0^{ij}) + \sum_{\text{non-contacts}} \epsilon_{\text{NC}} \left(\frac{\sigma_{\text{NC}}}{r_{ij}} \right)^{12}. \tag{4}$$

As in the C_α model, the coefficients are given homogeneous values: $\epsilon_r = 100\epsilon$, $\epsilon_\theta = 20\epsilon$, $\epsilon_\chi = 40\epsilon$, $\epsilon_{\rm NC} = 0.01\epsilon$, and $\sigma_{\rm NC} = 2.5$Å. The effective repulsive size for the atoms becomes $\sigma_{\rm eff} = (0.01)^{1/12}\sigma_{\rm NC} \approx 1.7$Å. Again, protein contacts that are separated by less than 3 residues are neglected. A technical issue is normalizing the dihedral energy around each bond. When assigning dihedral strengths, we first group dihedral angles that share the middle two atoms. For example, in a protein backbone, one can define up to four dihedral angles that possess the same C−C_α covalent bond as the central bond. Each dihedral in the group is scaled by $1/N_{\rm D}$, where $N_{\rm D}$ is the number of dihedral angles in the group.

Two ratios determine the distribution of dihedral and contact energies, contact to dihedral ratio $R_{C/D}$ and backbone to side chain ratio $R_{BB/SC}$. In proteins $R_{BB/SC} = \epsilon_{BB}/\epsilon_{SC} = 2$ [69] and in RNA/DNA $R_{BB/SC} = \epsilon_{BB}/\epsilon_{SC} = 1$ [64]. The contacts and dihedrals are scaled relative to their total contributions, $R_{C/D} = \frac{\sum \epsilon_C}{\sum \epsilon_{BB} + \sum \epsilon_{SC}} = 2$. Lastly, the total contact and dihedral energy is set equal to the number of atoms $\epsilon N_{\rm atoms} = \sum \epsilon_C + \sum \epsilon_{BB} + \sum \epsilon_{SC}$. This choice gives folding temperatures near 1 in reduced units ensuring a consistent parameterization.

Notice that every term is based on the native structure except the noncontact excluded volume term. In the C_α model, all the residues have a homogeneous shape, but in the AA model each residue has its unique geometry explicitly represented. This gives the AA model structure independent sequence information that adds heterogeneity to the model. This heterogeneity adds geometric frustration to the model, since interactions can only be satisfied if the side chains are correctly oriented [43]. A question of current interest is whether this sequence-dependent information adds constraints to the folding dynamics, allowing the native bias to be relaxed [3, 69].

3.2.3 Contact Potential

All of the pair interactions defined in the native contact map interact through a short range, attractive potential, denoted in the SBM potential by $C(r_{ij}, r_0^{ij})$. The contact potential has a minimum at r_0^{ij}, the distance between the pair in the native state. Traditionally, a contact is defined through a Lennard–Jones (LJ) type potential, since the LJ shape is readily available in molecular dynamics packages. In the C_α model a "10–12" LJ potential is used for contacts with the minimum set at the separation between the C_α pair in the native state r_0^{ij},

$$C_{\rm CA}(r_{ij}, r_0^{ij}) = 5\left(\frac{r_0^{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{r_0^{ij}}{r_{ij}}\right)^{10}, \tag{5}$$

and in the AA model a "6–12" LJ potential with the minimum set at the separation between a contacting atomic pair in the native state,

$$C_{\rm AA}(r_{ij}, r_0^{ij}) = \left(\frac{r_0^{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r_0^{ij}}{r_{ij}}\right)^{6}. \tag{6}$$

Different LJ potentials are used because the native contact distances r_0^{ij} can be much longer in the C_α model. The contacts are coarse-grained to be between the C_α atoms, which can be as distant as 14Å. The r^{-6} is much broader than the r^{-10} and can lead to unphysical structures in unfolded states as native pairs interact at long distances.

The LJ potentials are well tested and work for many systems, but they have limitations for certain applications because the LJ potential has an excluded volume that moves with the minimum. The effective size of two atoms in contact grows with r_0^{ij}. This additional excluded volume has little effect on the entropy of unfolded conformations since mostly noncontacts are interacting, but has a large effect on the entropy of the folded ensemble where most contacts are formed. In cases where the user wants to control the excluded volume term [32,43], an attractive Gaussian well coupled with a fixed hard wall-excluded volume is used,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) = \left(1 + \left(\frac{\sigma_{\mathrm{NC}}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij}, r_0^{ij})\right) - 1, \tag{7}$$

where

$$G(r_{ij}, r_0^{ij}) = -\exp\left[-(r_{ij} - r_0^{ij})^2/(2\sigma^2)\right]. \tag{8}$$

This unusual construction anchors the depth of the minimum at -1. The width of the Gaussian well σ is determined to model the variable width of the LJ potential. $C_{\mathrm{AA}}(1.2r_0^{ij}, r_0^{ij}) \sim -1/2$ so σ is defined such that $G(1.2r_0^{ij}, r_0^{ij}) = -1/2$ giving $\sigma^2 = (r_0^{ij})^2/(50 \ln 2)$. If σ_{NC} is significantly smaller than r_0^{ij}, (7) reduces to the more pedagogical form,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) \to \left(\frac{\sigma_{\mathrm{NC}}}{r_{ij}}\right)^{12} + G(r_{ij}, r_0^{ij}) \quad \text{for } \sigma_{\mathrm{NC}} \ll r_0^{ij}. \tag{9}$$

The flexibility of the Gaussian approach also allows for multiple basin contact potentials for energy landscapes with multiple minima (see Sect. 4.3). Using multiple LJ potentials with different locations of the minima is not a viable option because the longest LJ potential would occlude the others with its excluded volume term. A multibasin Gaussian potential C_{MB} for minima taken from two structures $\vec{\mathbf{x}}_\alpha$ and $\vec{\mathbf{x}}_\beta$ is given by [32],

$$C_{\mathrm{MB}}(r_{ij}, r_\alpha^{ij}, r_\beta^{ij}) = \left(1 + \left(\frac{\sigma_{\mathrm{NC}}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij}, r_\alpha^{ij})\right)\left(1 + G(r_{ij}, r_\beta^{ij})\right) - 1. \tag{10}$$

Analogous to (7), this construction fixes the depth of both minima at -1.

All of the various potential shapes are presented in Fig. 3. It should be noted that the folding temperature (defined in Sect. 4.1.1) is typically 0.2–0.3 reduced units higher for the Gaussian potential as compared to LJ because the extra excluded volume in the LJ potential destabilizes the native state.

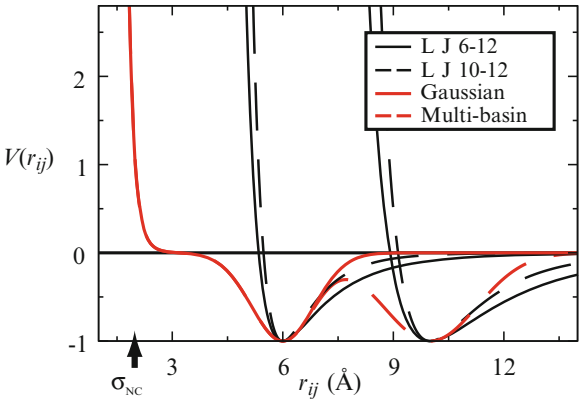

Fig. 3 Comparison of Lennard–Jones and Gaussian contact potentials. *Black curves* show LJ contact potentials with minima at 6Å and 10Å. The Gaussian contact potential shown in *red* has an excluded volume σ_{NC} that can be set independently of the location of the minimum. The *dotted red line* shows how the Gaussian contact would change as another minimum at 10Å is added

3.3 Molecular Dynamics with SBM

Molecular dynamics uses Newtonian mechanics to evolve the motions of atoms in time. The interactions defined in the SBM potential define the various forces on the atoms since force is given by the negative gradient of the potential energy. The system is evolved through time in discrete steps. The NVT canonical ensemble is maintained using a thermostat. Thermostats including a drag term, such as stochastic dynamics or Langevin dynamics are used for implicit solvent systems like SBMs. Velocity-rescaling thermostats can introduce heating artifacts when not coupled to an explicit solvent [41]. Langevin dynamics has been used to model the viscosity of a solvent [25, 57]. The output of a molecular dynamics simulation is a trajectory, a time-ordered series of snapshots of the atomic coordinates. The trajectory can be analyzed as a function of time to uncover kinetic properties or, by application of the ergodic theorem, as an ensemble to compute thermodynamic properties.

A molecular dynamics trajectory contains the coordinates of all the atoms in the system, a massive amount of information. Therefore, the trajectory is reduced down to one or a few reaction coordinates that monitor the progress of the dynamics under investigation. For protein folding, a useful reaction coordinate would differentiate between the unfolded ensemble, folding intermediates, and the folded ensemble. A reaction coordinate for studying a conformational transition would differentiate the various conformers. A natural reaction coordinate for SBMs is Q, the fraction of native contacts formed. A contact between the native pair ij is considered formed if it satisfies $r_{ij} < \gamma r_0^{ij}$, where $\gamma \approx 1.2$–1.4. In protein folding, low Q would correspond to the unfolded ensemble, medium Q would contain the transition state ensemble (TSE) and any intermediates, and high Q the folded ensemble. To investigate a conformational transition between two structures A and B, monitoring switching between high Q_A and high Q_B would indicate transitions. Other possible reaction coordinates are root mean square deviation from a reference structure or radius of gyration. An exciting possibility is to monitor the position of an explicitly represented FRET probe in order to compare with experimental data [66].

After the choice of reaction coordinate is made, the value of the coordinate during the trajectory (or several concatenated trajectories) can be histogrammed to obtain a potential of mean force (PMF) along the reaction coordinate. If the chosen coordinate adequately separates two basins, it can be used to identify the transition state at the peak on the free energy landscape. Q has been shown to be a suitable coordinate for protein transitions and thus the peaks in $F(Q)$ can be identified as TSEs [13] (see Fig. 5). Great care must be exercised when making quantitative predictions of thermodynamic and kinetic quantities from simplified models. The kinetics of the system are not simply determined by the free energy landscape, but are highly dependent on diffusion rates. Diffusion rates vary for different molecular systems and must be calibrated separately. For discussion of diffusion in SBM see [30,46,66]. Secondly, the precise values of free energy barriers and thermal stability are a fine balance and depend on the details of the SBM potential. This said, given a constant parameterization, kinetic and thermodynamic quantities tend to scale in a consistent fashion. Fast-folding proteins will consistently have smaller free energy barriers than slow-folding proteins [11, 69]. Some quantities are robust to perturbations, in particular the TSE and other so-called geometrical features of the energy landscape [32, 69].

3.4 SMOG: Automated Generation of SBM

Molecular dynamics simulations have benefited from years of research on computer algorithms constructed with one goal in mind: speed. Molecular dynamics suites like GROMACS [23], NAMD [49] and Desmond [7], package all the necessary algorithms to run stable molecular dynamics and the ability to scale the calculations to thousands of processors. These packages have made homegrown molecular dynamics codes built to run SBMs obsolete. SMOG, Structure-based Models in GROMACS, is a publicly available web server located at http://smog.ucsd.edu [44]. Any PDB structure consisting of standard amino acids, RNA, DNA, and common ligands, can be uploaded to SMOG, which outputs the necessary coordinate, topology, and parameter files to run a SBM in GROMACS. This provides the flexibility necessary to implement efficient and highly scalable SBMs. SMOG in conjunction with GROMACS version 4.5 scales easily to 128 processors when simulating a ribosome, $\sim$150,000 atoms. Protein-folding simulations of much smaller systems scale to $\sim$100 atoms per core on a single motherboard.

3.5 Choosing a Graining: C_α or All-Atom

The C_α and AA model are both able to describe the properties of the molecular scaffold's geometry. When comparing the two models, C_α and AA, the main advantage of C_α is its speed. Because the AA model has roughly eight times more

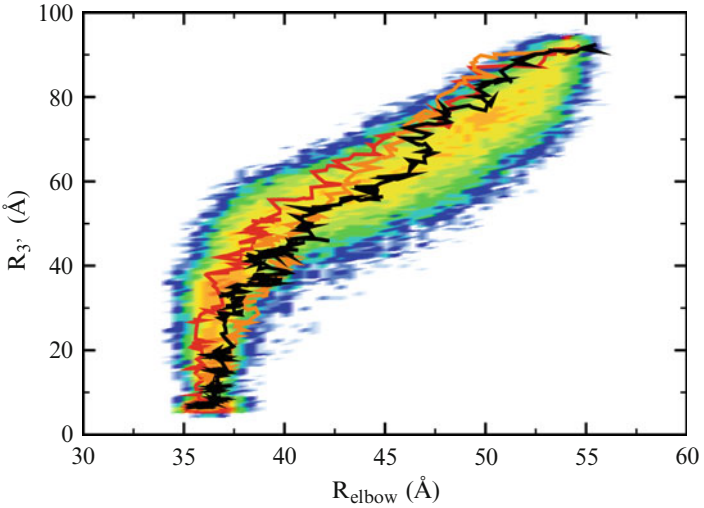

Fig. 4 Comparison of SBM and explicit solvent simulations of tRNA accommodation in the ribosome. Trajectories of three 4 ns explicit solvent-targeted molecular dynamics (TMD) overlay the probability distribution of 704μs structure-based TMD runs. With such a short sampling time, the explicit solvent TMD is dominated by steric interactions between the ribosome and the tRNA. The SBM naturally captures the sterics and is consistent with the detailed model. $R_{3'}$ and R_{elbow} monitor the position of the tRNA along the accommodation pathway. Simulations were started from the A/T state (high $R_{3'}$ and R_{elbow}) and stopped at the accommodated (A/A) state. See [66] for details

atoms and has slower diffusion due to side chain interactions, the C_α model runs significantly faster than AA. This speed is important for studying processes with large barriers, like folding and oligomerization. AA can narrow the speed gap with parallelization, but not close it completely. Nonetheless, AA has been used to fold small single domain proteins [69] and even proteins with complex topologies [43]. Many processes without large activation barriers, e.g., native basin dynamics, have energy landscapes that are easily sampled, and thus the performance hit of AA is of no consequence.

The explicit representation of atomic coordinates is advantageous, even for simplified models like SBM. A clear benefit is acting as a bridge between minimalist models and empirical force fields. Any conformations realized during a simulation of an AA SBM can be compared with, and used as input for, empirical force fields with an explicit solvent. Since the sterics are correct, any process that is dominated by large-scale structural fluctuations should be well represented by an AA SBM [42, 66]. Figure 4 shows targeted molecular dynamics (TMD) simulations of the tRNA accommodation process in the ribosome, a massive ribonucleoprotein molecular machine ($\sim$2.4 MDa). The trajectories from explicit solvent simulations overlay the AA SBM trajectories. On a smaller scale, the AA geometry opens the door to studying side chain degrees of freedom during folding and binding

simulations. Constricted conformations like polypeptide slipknots, found in coarse-grained models, are shown to be sterically possible with the AA geometry [43]. Lastly, the AA geometry allows a clear way to add perturbative nonnative chemical effects like hydrogen bonding [3] and partial charges.

4 Applications

SBMs are being applied to diverse problems, and in the remaining sections we describe a representative sample of how perfectly funneled SBMs are currently in use. In each case, the SBM can be constructed and implemented using SMOG and GROMACS. In Sects. 4.1–4.3, molecular dynamics is used to describe a system at thermodynamic equilibrium. In this case, it is necessary to adequately sample configuration space until the quantities of interest have converged. Finally, in Sect. 4.4 molecular dynamics is used to find deep energetic minima in perturbed structure-based potentials for molecular modeling applications.

4.1 Folding

4.1.1 Protein Folding

The most-established application of SBM is to the study of protein folding. Determining the TSE, the shape and size of free energy barriers, and the existence of folding intermediates are all topics of interest. Figure 5 shows the result of AA SBM folding simulations for two of the most thoroughly studied proteins, chymotrypsin inhibitor-2 (CI2) and the SH3 domain. These two proteins are two-state folders, meaning the protein only populates two basins spanned by a cooperative transition.

Figure 5a,d shows representative traces of Q versus time during constant temperature molecular dynamics near folding temperature T_F. T_F is the temperature such that the folding and unfolding basins are equally populated. Simulations are performed at T_F because it maximizes the sampling rate of the folding transition. T_F is determined by running simulations at high and low temperatures, and iteratively converging on a temperature where both folding and unfolding is observed. Q is defined as the fraction of native residue pairs with at least one atom–atom contact within 1.2 times its native separation. Alternative definition of Q, such as the fraction of atom–atom contacts formed, may shift the locations of basins in the resulting free energy landscape, but will preserve the heights of any barriers.

Q traces from long molecular dynamics trajectories at various temperatures can be combined using weighted histogram analysis (WHAM) [31], to obtain an optimal density of states. The density of states can then be used to extrapolate $F(Q)$ at any temperature (Fig. 5b, e). Always, care must be taken to ensure that the trajectories

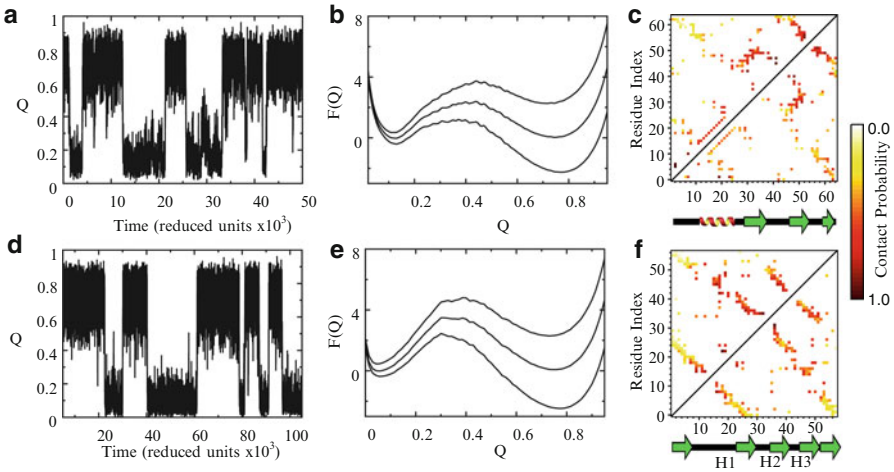

Fig. 5 All-atom structure-based simulations of two state-folding proteins CI2 (*top*) and SH3 domain (*bottom*). PDB codes: 1FMK, 1YPA. (**a,d**) The reaction coordinate Q plotted as a function of time for a typical simulation near T_F. Both proteins exhibit transitions between a folded ensemble at $Q \sim 0.8$ and an unfolded ensemble at $Q \sim 0.1$. (**b,e**) Free energy $F(Q)$ for temperatures $0.98T_F$, T_F, and $1.02T_F$ calculated by weighted histogram analysis of long constant temperature MD trajectories. A set of "long" trajectories typically contain 30 folded to unfolded transitions. (**c,f**) Transition state ensemble (TSE) for the two proteins. Contact formation probabilities are calculated by an unweighed average of all configurations $0.40 < Q < 0.45$. The *upper triangle* shows results from the C_α model and the *lower triangle* shows the AA model. Secondary structure is denoted below the contact maps as are the positions of the three hairpin turns in SH3. CI2 has a diffuse TSE that resembles the native state. The contact probability is largely predicted by sequence separation. SH3 has a more polarized TSE with contacts from the first ten residues largely absent. For both proteins, the introduction of energetic and structural heterogeneity through the AA geometry creates a more specific and less diffuse TSE. The simulations were prepared using SMOG v1.0.6 [44] with default parameters

reflect equilibrium. One easy method is to chop all trajectories in half and verify that $F(Q)$ and the TSE are the same for both halves. The TSE is the ensemble of structures that compose the bottleneck to folding. CI2 and SH3 each have a single TSE that connects the unfolded state to the folded state defined by the structure populating the top of $F(Q)$. Figure 5c,f shows the average contact maps of the structures with $0.4 < Q < 0.45$. The contact formation probabilities can be connected to Φ-value analysis, an experimental technique that estimates the contribution of a particular residue's contacts to the TSE [19]. In simulation, Φ_i is given by

$$\Phi_i = \frac{P_i^{\mathrm{TSE}} - P_i^{\mathrm{U}}}{P_i^{\mathrm{F}} - P_i^{\mathrm{U}}}, \tag{11}$$

where P_i is the probability that residue i forms its contacts and U/F refers to the unfolded/folded ensembles [36]. Φ_i near 1 means that residue i is very native-like in the TSE and a Φ_i near 0 means that residue i is still unfolded in the TSE.

Since the TSE is a simple average over structures, it can hold hidden complexity. For some proteins, the TSE is composed of multiple routes through the TSE [6, 22]. Consider SH3; its TSE could be composed of two routes, a major route where hairpin 2 and hairpin 3 form first and a minor route where hairpin 1 and hairpin 2 form first (Fig. 5f). Multiple routes can be identified by clustering the contact maps of TSE structures using the number of shared contacts as a similarity measure [6]. These routes represent entropically viable routes through the TSE. Thus, two real proteins that fold to the same structure may follow seemingly very different paths due to minor energetic differences.

4.1.2 Multimeric Folding and Binding

Many important biological processes are regulated by the homo- or hetero-oligomers that are formed when proteins bind [70]. A large survey of protein dimers showed that the binding mechanisms found in experiments were reproduced by SBMs [36], which gives strong evidence that protein binding is controlled by protein geometry. The energy landscapes of these proteins exhibited a rich variety of folding routes and binding mechanisms. The interplay of folding and binding can be explored in SBMs by introducing interface contacts into the native contact map. The contact map of crystallographic structures of protein dimers are analyzed in the same way as for monomers, atoms spatially close between the protomers are considered native contacts. Folding trajectories of protomers A and B will have three relevant order parameters, Q_A, Q_B, and Q_{AB}. Note that when analyzing the TSE and folding routes of homo-oligomeric proteins, clustering the TSE is crucial [6]. This is because the structural symmetry is broken by the requirement of labeling the protomers, i.e., protomer A folds then binds protomer B is the same route as B folds then binds A.

Observing binding in simulations is complicated by the entropy loss of binding. In order to observe binding events, the effective concentration of monomers is often much higher than in vivo. The concentration of monomers is imposed either by a linker between the monomers [36], periodic boundary conditions [64], or an umbrella potential [6, 43, 52] (all available in GROMACS). The umbrella potential would be implemented as a harmonic center of mass constraint, making the simulated potential

$$V_{\text{dimer}} = V_{\text{AA}} + k\left(r_{\text{CM}} - r_0^{\text{CM}}\right)^2, \tag{12}$$

where r_{CM} is the distance between the centers of mass and r_0^{CM} is the distance in the native state. k is calibrated to be as weak as possible while still observing binding. Varying k can model varying protomer concentration. The stability of the dimer versus the monomers can be controlled by scaling the strength of the interface contacts.

4.2 Native Basin Dynamics

Entropically driven motions accessible via thermal fluctuations are important components of functional protein dynamics [20]. These motions are difficult or impossible to intuit from rigid crystallographic structure analysis [42]. Analysis of small-angle x-ray scattering (SAXS) on C-terminal Src Kinase (Csk) indicated that Csk occupies extended conformations in solution, whereas the crystal structure showed a compact arrangement of Csk's SH2, SH3, and kinase domains [27]. Typically, a candidate structure for the protein structure is determined by fitting a rigid body model to the SAXS data, but this presumes that Csk assumes a relatively static structure in solution. In order to characterize the Csk solution structure, constant temperature molecular dynamics simulations of the Csk native basin were performed using the AA SBM. Theoretical scattering curves were computed from the resulting native ensemble and compared with the experimental scattering data. Jamros et al. [27] showed that in all cases, theoretical scattering curves generated from mixed populations of Csk structures fit the empirical SAXS data better than any rigid model. This suggests that Csk populates a broad ensemble of structures in solution, adopting conformations not observed in the crystal structure. More pertinently, an SBM is able to suggest a solution ensemble of structures for Csk using only information from the crystal structure. This procedure, termed Safe-SAXS, should be widely applicable to analyzing solution structures of biological macromolecules.

4.3 Multiple Basin Models

When a protein is able to be crystallized in substantially different conformations, it implies the energy landscape has multiple minima. This behavior can be seen in systems with a high degree of structural symmetry. A dual basin-funneled landscape solved the mystery of the Rop dimer, a dimer of two helix bundles that switched from a parallel arrangement to an antiparallel arrangement upon optimization of the hydrophobic core [21, 34, 52]. An SBM was used that combined the two crystal structure contact maps into a single native contact map. Thermodynamic sampling of the landscape showed that the parallel and antiparallel structures were of similar stability, so small experimental perturbations could tip the balance between the structures [52].

Combining multiple structures into a single landscape has also been used to study conformational transitions in adenylate kinase (AKE) [26, 67, 68]. AKE has two domains, LID and NMP, that must undergo large conformational changes during its enzymatic function (Fig. 6). The conformational change is captured by two crystal structures, one in the open state and the other in a closed state, with native contact maps M_{O} and M_{C}, respectively. The contacts that are in both maps is given by

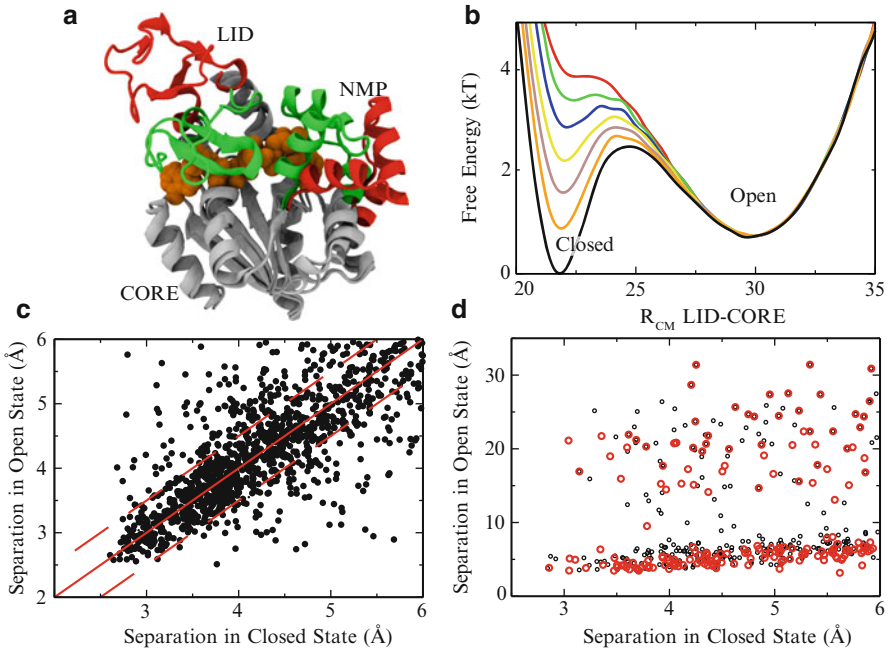

Fig. 6 Modeling conformational transitions in adenylate kinase (AKE). (**a**) AKE contains two domains, NMP and LID, that undergo >25Å motions between open (*red*) and closed (*green*) states. These motions are coupled with ligand (shown as *orange spheres*) binding as it catalyzes ATP+AMP⇆2ADP. The model is built using structures with PDB codes 1AKE and 4AKE. (**b**) The relative occupation of the closed and open states can be tuned to experimental data by varying the strength of the subset of contacts only existing in the closed state M_C. M_{BB} is scaled by 0.6 (*red*) to 1.2 (*black*) relative to the open contacts. (**c**) The subset of atomic contacts existing in both states M_{same}. The *dotted lines* designate a deviation of less than 0.5Å between states. Contacts that have significant shifts between structures may impart strain on the protein and can be included with double minima Gaussian potentials. (**d**) The subset of atomic contacts existing only in the closed state. Black circles show contacts of atoms in the LID domain and red circles show contacts of atoms in the NMP domain. See [68] for details

$M_{same} = M_O \bigcap M_C$ and the complement of M_{same} are the contacts that are in either map but not both M_{diff}. Results from a SBM with native contact map M_{diff} is shown in Fig. 6b. The relative stabilities of the two states can be easily tuned in the SBM. The distance between contacts that exist in both states (Fig. 6c) may change between structures and can be included with double minima Gaussian potentials (Sect. 3.2.3). How to handle multiple dihedral angle values is less obvious. Whitford et al. [68] simply used the dihedrals from the open state, viewing the closed state as an excitation of the open state. Similar methods have been used to look at conformational changes in protein kinase A [24] and kinesin [25].

4.4 Molecular Modeling

SBM are structurally robust, which makes them ideal candidates for molecular modeling applications. During molecular dynamics the native bias maintains a native-like configuration but all interactions are malleable. Under molecular dynamics, a system populates the lowest free energy basins, and coupled with simulated annealing can even search for the lowest potential energy minima [63]. Through the introduction of external biasing potentials, AA SBMs built from high-resolution structures can reveal candidate AA structures from low resolution experimental data.

In a recent study of the ribosomal elongation cycle, Ratje et al. [50] used multiparticle cryoelectron microscopy analysis to capture subpopulations of EF-G-ribosome complexes at subnanometer resolution. While this resolution is not fine enough to achieve atomic details, the known crystallographic structure can be used to obtain atomic models of the microscopy data with a procedure termed MDFIT [65]. MDFIT biases the AA SBM with an energetic term developed in Orzechowski and Tama [48], which uses the correlation between the simulated and experimental electron density. The overall potential function therefore becomes

$$V_{\text{model}} = V_{\text{AA}} + V_{\text{map}} = V_{\text{AA}} + W \sum_{ijk} \rho_{ijk}^{\text{sim}} \rho_{ijk}^{\text{exp}}, \tag{13}$$

where W is an overall weight and ρ_{ijk}^{sim} and ρ_{ijk}^{exp} are the normalized electron densities at voxel (i, j, k) and V_{AA} is the AA SBM potential. A molecular dynamics simulation initialized at the crystallographic structure will distort to maximize the overlap between the simulated structure and the experimental electron density. The structure-based potential naturally maintains tertiary contacts present in the crystal structure without the need for ad hoc restraints.

The electron density map works well as a global bias, but local biases can also be introduced. Candidate structures for protein–protein complexes can be derived by introducing interprotein contacts from bioinformatic analysis and minimizing the resulting structure-based potential with molecular dynamics. Schug et al. [51] were able to predict the structure of the Spo0B/Spo0F two-component signal transduction (TCS) complex within 2.5Å of an existing crystal structure. TCS is ruled by transient interactions, posing harsh challenges to obtain atomic resolution structures. These transient interactions though have bioinformatic signatures, which provide the external biasing potential needed for modeling. Short-range contact potentials were introduced between correlated residues and the resulting potential

$$V_{\text{model}} = V_{\text{AA}} + k(r_{\text{CM}})^2 + \sum_{\{i,j\}} C_{\text{AA}}(r_{ij}, \bar{r}), \tag{14}$$

where r_{CM} is the distance between the proteins' centers of mass, $\{i, j\}$ denotes the correlated residues, C_{AA} is Eqn. 6, r_{ij} the distance between those residues' C_α atoms and $\overline{r} = 7$ Å. A weak center of mass constraint, as with multimeric folding (see Sect. 4.1.2), is a common method of encouraging two molecules to dock. The resulting structure from the AA SBM simulations can be directly used as input to an AA empirical force field for additional minimization.

5 Concluding Remarks

The principle of minimal frustration and the funneled landscape provide the theoretical framework for SBMs. We have presented numerous applications of SBMs, including protein folding and oligomerization, structure–function relationships in protein conformational transitions and structural modeling of protein–protein and ribonucleoprotein complexes. These models are publicly available at SMOG http://smog.ucsd.edu. Recent technical improvements in computer hardware for molecular dynamics simulations should allow for a new level of collaboration between simplified protein models and explicit solvent models. Protein folding simulations on the millisecond timescale will enable quantitative characterization of the roughness of the folding energy landscape [37, 54]. As experimentalists continue pushing boundaries in the characterization of molecular machines at the single molecule level, further theoretical investigation is needed to assess how the interplay of global properties with specific energetic details shapes the dynamics of these large macromolecular complexes [66]. We expect the importance of large-scale structural fluctuations, largely controlled by geometry, to be a central theme in the discussion of molecular machines in the years to come.

Acknowledgments JKN would like to thank Joanna Sułkowska for many helpful discussions and Paul Whitford and Ryan Hayes for a careful reading of the chapter. This work was supported by the Center for Theoretical Biological Physics sponsored by the national science foundation (NSF) (Grant PHY-0822283) and NSF Grant NSF-MCB-1051438.

References

1. Adcock, S.A., McCammon, J.A.: Molecular dynamics: survey of methods for simulating the activity of proteins. Chem. Rev. **106**(5), 1589–1615 (2006)
2. Andrews, B.T., Gosavi, S., Finke, J.M., Onuchic, J.N., Jennings, P.A.: The dual-basin landscape in gfp folding. Proc. Nat. Acad. Sci. USA **105**(34), 12283–12288 (2008)
3. de Araujo, A.F.P., Onuchic, J.N.: A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. Proc. Nat. Acad. Sci. USA **106**(45), 19001–19004 (2009)
4. Azia, A., Levy, Y.: Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. J. Mol. Biol. **393**(2), 527–542 (2009)

5. Baker, D.: A surprising simplicity to protein folding. Nature **405**(6782), 39–42 (2000)
6. Baxter, E.L., Jennings, P.A., Onuchic, J.N.: Interdomain communication revealed in the diabetes drug target mitoneet. Proc. Nat. Acad. Sci. USA **108**(13), 5266–5271 (2011)
7. Bowers, K.J., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., Gregersen, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M.A., Sacerdoti, F.D., Salmon, J.K., Shan, Y., Shaw, D.E.: Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of ACM/IEEE, p. 43 (2006)
8. Bryngelson, J., Wolynes, P.: Spin glasses and the statistical mechanics of protein folding. Proc. Nat. Acad. Sci. USA **84**, 7524 (1987)
9. Bryngelson, J., Wolynes, P.: Intermediates and barrier crossing in a random energy model (with applications to protein folding). J. Phys. Chem. **93**, 6902–6915 (1989)
10. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G.: Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins: Struct. Funct. Bioinf. **21**(3), 167–195 (1995)
11. Chavez, L.L., Onuchic, J.N., Clementi, C.: Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. J. Am. Chem. Soc. **126**(27), 8426–8432 (2004)
12. Cheung, M.S., García, A.E., Onuchic, J.N.: Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. Proc. Nat. Acad. Sci. USA **99**(2), 685–690 (2002)
13. Cho, S., Levy, Y., Wolynes, P.G.: P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. Proc. Nat. Acad. Sci. USA **103**(3), 586–591 (2006)
14. Cho, S.S., Weinkam, P., Wolynes, P.G.: Origins of barriers and barrierless folding in bbl. Proc. Nat. Acad. Sci. USA **105**(1), 118–123 (2008)
15. Clementi, C., Nymeyer, H., Onuchic, J.N.: 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**(5), 937–953 (2000)
16. Clementi, C., García, A.E., Onuchic, J.N.: Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein l. J. Mol. Biol. **326**(3), 933–954 (2003)
17. Clementi, C., Plotkin, S.S.: The effects of nonnative interactions on protein folding rates: theory and simulation. Protein Sci. **13**(7), 1750–1766 (2004)
18. Ferguson, N., Schartau, P.J., Sharpe, T.D., Sato, S., Fersht, A.R.: One-state downhill versus conventional protein folding. J. Mol. Biol. **344**(2), 295–301 (2004)
19. Fersht, A.R.: Characterizing transition-states in protein-folding - an essential step in the puzzle. Curr. Opin. Struct. Biol. **5**(1), 79–84 (1995)
20. Frauenfelder, H., Sligar, S.G., Wolynes, P.G.: The energy landscapes and motions of proteins. Science **254**(5038), 1598–1603 (1991)
21. Gambin, Y., Schug, A., Lemke, E.A., Lavinder, J.J., Ferreon, A.C.M., Magliery, T.J., Onuchic, J.N., Deniz, A.A.: Direct single-molecule observation of a protein living in two opposed native structures. Proc. Nat. Acad. Sci. USA **106**(25), 10153–10158 (2009)
22. Gosavi, S., Chavez, L.L., Jennings, P.A., Onuchic, J.N.: Topological frustration and the folding of interleukin-1 beta. J. Mol. Biol. **357**(3), 986–996 (2006)
23. Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E.: Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theo. Comput. **4**(3), 435–447 (2008)
24. Hyeon, C., Jennings, P.A., Adams, J.A., Onuchic, J.N.: Ligand-induced global transitions in the catalytic domain of protein kinase A. Proc. Nat. Acad. Sci. USA **106**(9), 3023–3028 (2009)
25. Hyeon, C., Onuchic, J.N.: Mechanical control of the directional stepping dynamics of the kinesin motor. Proc. Nat. Acad. Sci. USA **104**(44), 17382–17387 (2007)
26. Okazaki, K., Koga, N., Takada, S., Onuchic, J.N., Wolynes, P.G.: Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: struc-based molecular dynamics simulations. Proc. Nat. Acad. Sci. USA **103**(32), 11844–11849 (2006)

27. Jamros, M.A., Oliveira, L.C., Whitford, P.C., Onuchic, J.N., Adams, J.A., Blumenthal, D.K., Jennings, P.A.: Proteins at work: a combined small angle x-ray scattering and theoretical determination of the multiple structures involved on the protein kinase functional landscape. J. Biol. Chem. **285**(46), 36121–36128 (2010)
28. Kaya, H., Chan, H.S.: Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? J. Mol. Biol. **326**(3), 911–931 (2003)
29. Koga, N., Takada, S.: Roles of native topology and chain-length scaling in protein folding: a simulation study with a go-like model. J. Mol. Biol. **313**(1), 171–180 (2001)
30. Kouza, M., Li, M.S., O'brien, E.P., Hu, C.-K., Thirumalai, D.: Effect of finite size on cooperativity and rates of protein folding. J. Phys. Chem. A **110**(2), 671–676 (2006)
31. Kumar, S., Rosenberg, J., Bouzida, D., Swendsen, R.H.: The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem. **13**(8), 1011 (1992)
32. Lammert, H., Schug, A., Onuchic, J.N.: Robustness and generalization of structure-based models for protein folding and function. Proteins: Struct. Funct. Bioinf. **77**(4), 881–891 (2009)
33. Leopold, P.E., Montal, M., Onuchic, J.N.: Protein folding funnels - a kinetic approach to the sequence structure relationship. Proc. Nat. Acad. Sci. USA **89**(18), 8721–8725 (1992)
34. Levy, Y., Cho, S.S., Shen, T., Onuchic, J.N., Wolynes, P.G.: Symmetry and frustration in protein energy landscapes: a near degeneracy resolves the rop dimer-folding mystery. Proc. Nat. Acad. Sci. USA **102**(7), 2373–2378 (2005)
35. Levy, Y., Onuchic, J.N., Wolynes, P.G.: Fly-casting in protein-dna binding: frustration between protein folding and electrostatics facilitates target recognition. J. Am. Chem. Soc. **129**(4), 738–739 (2007)
36. Levy, Y., Wolynes, P.G., Onuchic, J.N.: Protein topology determines binding mechanism. Proc. Nat. Acad. Sci. USA **101**(2), 511–516 (2004)
37. Lindorff-Larsen, K., Piana, S., Dror, R.O., Shaw, D.E.: How fast-folding proteins fold. Science 334, 517–520 (2011)
38. McCammon, J.A., Gelin, B.R., Karplus, M.: Dynamics of folded proteins. Nature **267**(5612), 585–590 (1977)
39. Mittal, A., Jayaram, B.: Backbones of folded proteins reveal novel invariant amino acid neighborhoods. J. Biomol. Struct. Dyn. **28**(4), 443–454 (2011)
40. Miyashita, O., Onuchic, J.N., Wolynes, P.G.: Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc. Nat. Acad. Sci. USA **100**(22), 12570–12575 (2003)
41. Mor, A., Ziv, G., Levy, Y.: Simulations of proteins with inhomogeneous degrees of freedom: the effect of thermostats. J. Comput. Chem. **29**(12), 1992–1998 (2008)
42. Nechushtai, R., Lammert, H., Michaeli, D., Eisenberg-Domovich, Y., Zuris, J.A., Luca, M.A., Capraro, D.T., Fish, A., Shimshon, O., Roy, M., Schug, A., Whitford, P.C., Livnah, O., Onuchic, J.N., Jennings, P.A.: Allostery in the ferredoxin protein motif does not involve a conformational switch. Proc. Nat. Acad. Sci. USA **108**(6), 2240–2245 (2011)
43. Noel, J.K., Sulkowska, J.I., Onuchic, J.N.: Slipknotting upon native-like loop formation in a trefoil knot protein. Proc. Nat. Acad. Sci. USA **107**(35), 15403–15408 (2010)
44. Noel, J.K., Whitford, P.C., Sanbonmatsu, K.Y., Onuchic, J.N.: Smog@ctbp: simplified deployment of structure-based models in gromacs. Nucleic Acids Res. 38, W657 (2010)
45. Nymeyer, H., García, A.E., Onuchic, J.N.: Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc. Nat. Acad. Sci. USA **95**(11), 5921–5928 (1998)
46. Oliveira, R.J., Whitford, P.C., Chahine, J., Wang, J., Onuchic, J.N., Leite, V.B.P.: The origin of nonmonotonic complex behavior and the effects of nonnative interactions on the diffusive properties of protein folding. Biophys. J. **99**(2), 600–608 (2010)
47. Onuchic, J.N., Wolynes, P.G.: Theory of protein folding. Curr. Opin. Struct. Biol. **14**(1), 70–75 (2004)

48. Orzechowski, M., Tama, F.: Flexible fitting of high-resolution x-ray structures into cryo-electron microscopy maps using biased molecular dynamics simulations. Biophys. J. **95**(12), 5692–5705 (2008)
49. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., Schulten, K.: Scalable molecular dynamics with namd. J. Comput. Chem. **26**(16), 1781–1802 (2005)
50. Ratje, A.H., Loerke, J., Mikolajka, A., Brünner, M., Hildebrand, P.W., Starosta, A.L., Dönhöfer, A., Connell, S.R., Fucini, P., Mielke, T., Whitford, P.C., Onuchic, J.N., Yu, Y., Sanbonmatsu, K.Y., Hartmann, R.K., Penczek, P.A., Wilson, D.N., Spahn, C.M.T.: Head swivel on the ribosome facilitates translocation by means of intra-subunit trna hybrid sites. Nature **468**(7324), 713–716 (2010)
51. Schug, A., Weigt, M., Onuchic, J.N., Hwa, T., Szurmant, H.: High-resolution protein complexes from integrating genomic information with molecular simulation. Proc. Nat. Acad. Sci. USA **106**(52), 22124–22129 (2009)
52. Schug, A., Whitford, P.C., Levy, Y., Onuchic, J.N.: Mutations as trapdoors to two competing native conformations of the rop-dimer. Proc. Nat. Acad. Sci. USA **104**(45), 17674–17679 (2007)
53. Scott, K.A., Batey, S., Hooton, K.A., Clarke, J.: The folding of spectrin domains i: wild-type domains have the same stability but very different kinetic properties. J. Mol. Biol. **344**(1), 195–205 (2004)
54. Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., Wriggers, W.: Atomic-level characterization of the structural dynamics of proteins. Science **330**(6002), 341–346 (2010)
55. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M.: Automated analysis of interatomic contacts in proteins. Bioinformatics **15**(4), 327–332 (1999)
56. Socci, N.D., Onuchic, J.N., Wolynes, P.G.: Diffusive dynamics of the reaction coordinate for protein folding funnels. J. Chem. Phys. **104**(15), 5860–5868 (1996)
57. Sułkowska, J., Sułkowski, P., Szymczak, P., Cieplak, M.: Tightening of knots in proteins. Phys. Rev. Lett. **100**(5), 058106 (2008)
58. Sułkowska, J.I., Cieplak, M.: Selection of optimal variants of gō-like models of proteins through studies of stretching. Biophys. J. **95**(7), 3174–3191 (2008)
59. Sułkowska, J.I., Sułkowski, P., Onuchic, J.: Dodging the crisis of folding proteins with knots. Proc. Nat. Acad. Sci. USA **106**(9), 3119–3124 (2009)
60. Sutto, L., Lätzer, J., Hegler, J.A., Ferreiro, D.U., Wolynes, P.G.: Consequences of localized frustration for the folding mechanism of the im7 protein. Proc. Nat. Acad. Sci. USA **104**(50), 19825–19830 (2007)
61. Tirion, M.: Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. **77**(9), 1905–1908 (1996)
62. Veitshans, T., Klimov, D., Thirumalai, D.: Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. Folding and Design **2**(1), 1–22 (1997)
63. Wales, D.J.: Energy Landscapes. Cambridge University Press, Cambridge (2003)
64. Whitford, P., Schug, A., Saunders, J., Hennelly, S., Onuchic, J., Sanbonmatsu, K.: Supplementary-nonlocal helix formation is key to understanding s-adenosylmethionine-1 riboswitch function. Biophys. J. **96**(2), L7–L9 (2009)
65. Whitford, P.C., Ahmed, A., Yu, Y., Hennelly, S.P., Tama, F., Spahn, C.M.T., Onuchic, J., Sanbonmatsu, K.Y.: Excited states of ribosome translocation revealed through integrative molecular modeling. Proc. Nat. Acad. Sci. USA **108**(47), 18943–18948 (2011)
66. Whitford, P.C., Geggier, P., Altman, R.B., Blanchard, S.C., Onuchic, J.N., Sanbonmatsu, K.Y.: Accommodation of aminoacyl-trna into the ribosome involves reversible excursions along multiple pathways. RNA **16**(6), 1196–1204 (2010)
67. Whitford, P.C., Gosavi, S., Onuchic, J.N.: Conformational transitions in adenylate kinase. Allosteric communication reduces misligation. J. Biol. Chem. **283**(4), 2042–2048 (2008)

68. Whitford, P.C., Miyashita, O., Levy, Y., Onuchic, J.N.: Conformational transitions of adenylate kinase: switching by cracking. J. Mol. Biol. **366**(5), 1661–1671 (2007)
69. Whitford, P.C., Noel, J.K., Gosavi, S., Schug, A., Sanbonmatsu, K.Y., Onuchic, J.N.: An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. Proteins: Struct. Funct. Bioinf. **75**(2), 430–441 (2009)
70. Wodak, S.J., Janin, J.: Structural basis of macromolecular recognition. Adv. Protein Chem. **61**, 9–73 (2002)
71. Wu, L., Zhang, J., Qin, M., Liu, F., Wang, W.: Folding of proteins with an all-atom go-model. J. Chem. Phys. **128**(23), 235103 (2008)