Coarse-Grained Protein Model with Residue Orientation Energies Derived from Atomic Force Fields

Marcos R. Betancourt*

Department of Physics, Indiana University Purdue University Indianapolis, 402 North Blackford Street LD156-J, Indianapolis, Indiana 46202

Received: July 15, 2009; Revised Manuscript Received: September 13, 2009

Coarse-grained models for protein simulations can potentially access longer time scales in larger protein systems than atomic-level models. Here, a coarse-grained residue pair potential, with distance and orientation dependency, is derived from equilibrium ensembles of residue pairs generated by molecular dynamics (MD). In particular, the Boltzmann inversion method is used to determine the energies. The residue pair potential is combined with local dihedral angle potentials for the backbone and side chains and used in the folding simulations of six small proteins, (28−67 residues) containing a variety of secondary structures. For the proteins tested, folding simulations by Monte Carlo methods generate structures similar to the native ones. However, these native-like structures were among the lowest in energy for α helical proteins but not for proteins containing extended β structures. It is also found that a careful balance between local and nonlocal interactions is essential. Possibilities for improving coarse-grained models derived from atomic force fields are discussed.

1. Introduction

Many important biophysical and biochemical processes involving proteins occur on time scales that are too difficult to access with simulations involving atomic force fields. For example, in protein folding applications, some of the latest, more accurate, and efficient techniques are carried out in proteins with less than ∼100 residues and require large computer resources and simulations that can last for several months.[1−3] A popular solution is to use coarse-grained models, which replace a group of atoms by an effective atom, hence reducing the number of degrees of freedom. Coarse-grained models have been used to model proteins and protein folding since the pioneering work of Levitt and Warshel.[4] Despite significant progress and a large variety of models introduced during the last several decades, the energy parametrization of these models remains a significant challenge. The reason is that in problems such as the determination of protein thermodynamics, its native structure, and protein binding, the results can be sensitive to errors in the energy−structure relationships.

Traditionally, the energy parametrization of coarse-grained models has been achieved either from empirical potentials based on physicochemical models[5−9] or from knowledge-based potentials (KBPs) obtained from residue conformational probabilities in protein native structures.[10−13] The latter approach has been one of the most useful in the protein structure prediction problem,[2,14,15] providing significant structural accuracy. Despite their advantages, KBPs are less physical than empirical potentials and are generally restricted to physiological conditions.[16] On the other hand, the parametrization of empirical potentials involves many approximations and choices that are often inadequate and less systematic.

A third approach has been introduced more recently, and it is based on extracting the coarse-grained energy parameters from the atomic-level simulations of residues (or short peptides) in thermodynamic equilibrium.[17−21] This approach has the advantage of being more physical than KBPs yet systematically derivable and can be obtained for different thermodynamical states and solvent conditions. There are several variations of this method. One involves extracting potentials of mean force (PMF) using the force matching (FM) method.[19,22] Another averages and integrates the atomic forces to parametrize analytical models of the PMF.[17] A third approach involves the Boltzmann Inversion (BI) method, which is often used in KBPs and involves expressing the potential as a function of the probabilities of observing a group of atoms interacting in different conformations.

In a recent work, a residue pairwise potential for nonbonded interactions was derived using atomic coarse-graining and the BI method for all 210 amino acid pairs.[21] It was shown that, despite being distance-dependent only, this potential has a reasonable ability to identify a native structure among decoys. However, it is well known that pairwise residue distance potentials are insufficient to describe protein structure and folding.[23,24] In the present work, the objective is to add residue orientation dependencies to the nonbonded potential and to use it in protein folding simulations of small proteins. For the simulations, the potential is complemented with a local potential (including bonded interactions) for the backbone and side chains and with excluded volume interactions.

In the following sections, the coarse-grained potential is described. The method for obtaining the potential from molecular dynamic simulations is reviewed. Finally, the results of folding a small group of proteins is presented and discussed.

2. Theory and Computational Methods

Nonlocal Energies. In general, statistical potentials are derived from the probability $P(\mathcal{C})$ for a polypeptide to be in a given conformation $\mathcal{C}$. If the polypeptide is in thermodynamic equilibrium at temperature T, then its potential energy $E(\mathcal{C})$ can be obtained with the BI method, that is

* Tel.: (317) 274-6910. Fax: (317) 274-2393. E-mail: mrbetanc@iupui.edu.

10.1021/jp906710c CCC: $40.75
Published on Web 10/09/2009

$$E(\mathcal{O}) = -\ln[P(\mathcal{O})/P^0(\mathcal{O})] \quad (1)$$

where $P^0(\mathcal{O})$ is some reference probability that defines the zero reference energy, with the energy in units of k_BT.

The BI approach assumes that the polypeptide energy can be separated as a sum of individual terms, each related to independent probabilities by the Boltzmann relation. These probabilities are a function of a few conformation variables that define the interaction between particular elements of the polypeptide such as the residues. In the atomic force field coarse-graining approach, one or more of each of these elements are simulated to generate equilibrium canonical ensembles from which the statistical potential is obtained. The assumption is once again that when these energies are added, the energy for the entire polypeptide is recovered (the additivity assumption). Selecting these polypeptide elements is somewhat arbitrary, which can be anywhere from a united atom (heavy atom plus hydrogens) to a group of residues. In this process, many internal degrees of freedom are averaged out, presumably corresponding to coordinates less relevant to the time scales and spatial resolutions of interest. In addition, the solvent degrees of freedom are also averaged out, adding their entropic and enthalpic effects to the coarse-grained potential. The potential is, in effect, a renormalized free energy. To maximally satisfy the additivity assumption, the polypeptide subunits and coordinates have to be strategically selected to minimize the redundancy of interactions and to capture the correlations between coordinates as much as possible (i.e., not all coordinates can be assumed to be independent).

In this work, the focus is in coarse-graining the nonbonded interactions between entire residues, with the coordinates being given by their relative distance and orientations. Models using orientations between entire residues,[25] atom groups,[26,27] atoms,[28] and even nearest-residue side chains in local interactions[29] have been previously used in the context of KBPs. These nonbonded (or nonlocal) interactions alone cannot fully describe the polypeptide conformations. Therefore, local interactions, consisting of interactions between bonded atoms and nonbonded interactions between adjacent residues obtained by the BI method, are included. The backbone and side-chain dihedral angles are used as local interaction coordinates. These interactions are adapted from other derivations and are described further below in this article. Because of the use of a combination of coarse-grained nonlocal interactions and atomic resolution local interactions, the present model is considered hybrid (or multiscale).

Given that there is a local potential that depends on the backbone dihedral angles ϕ, ψ, and ω and the set of side-chain angles $\{\chi\}$, a general form for a coarse-grained nonbonded potential between a pair of residues can be expressed in terms of their relative orientation coordinates. If A and B are the composition of two residues with indices i and j, their equilibrium probability to be in a given configuration can be generally expressed as

$$P'_{ij} = P_{AB}(r, \tau, \sigma_i, \theta_i, \sigma_j, \theta_j | \phi_i, \psi_i, \{\chi\}_i, \phi_j, \psi_j, \{\chi\}_j) \quad (2)$$

where r, τ, σ_i, θ_i, σ_j, and θ_j are the relative orientation angles depicted in Figure 1 and the probability depends conditionally on the backbone and side-chain angles (neglecting ω). Note that σ_i, θ_i and σ_j, θ_j are spherical polar angles for each residue in relation to the other, and r and τ are the distance and torsion angle between them. Also notice that the residue polar axes are

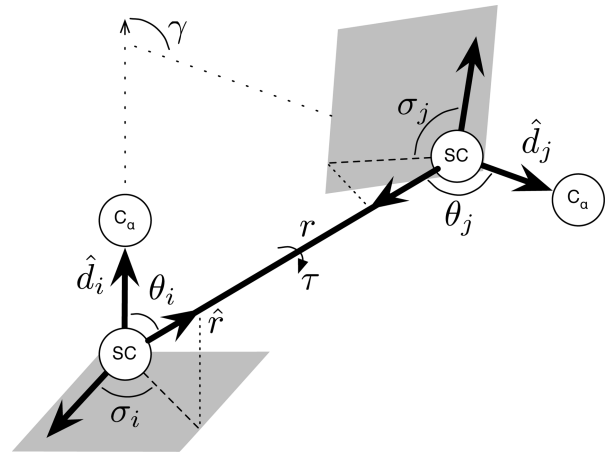

Figure 1. Definition of general orientation angles and coarse-grained vectors between two residues with indices i and j. SC stands for the side chain, with a reference point located at a predefined atom close to the center of mass of the side chain.[21] The σ angles are not used in the present model and are only shown for reference. These can be defined between the projections (dashed lines) of $\hat{r}$ in the plane perpendicular to the $\hat{d}$ vector and a side-chain reference vector in the plane. The current model uses the distance r, the angles θ_i and θ_j, and the angle γ formed between the two $\hat{d}$ vectors. The torsion angle τ, defined between the $\hat{d}_i$, $\hat{r}$, and $\hat{d}_j$ vectors, is replaced by γ.

defined from the side chain to the backbone C_α. This model averages over bond lengths, bond angles, and the solvent configurations. The main problem with this model is its high dimensionality, making its numerical sampling impractical. A reasonable assumption is to ignore the side-chain degrees of freedom $\{\chi\}$, which allows the representation of the residues as rigid bodies, except for the backbone hydrogen and oxygen, which are determined by the ϕ,ψ angles. Even without the side-chain degrees of freedom, it is still difficult to model. Therefore a more drastic approximation is made, where the probabilities are assumed to be independent of the ϕ, ψ angles alltogether, becoming

$$P''_{ij} = P_{AB}(r, \tau, \sigma_i, \theta_i, \sigma_j, \theta_j) \quad (3)$$

which is the rigid body model. With this model, the contributions of the backbone hydrogen bonds have to be considered separately. A six-dimensional probability space for each amino acid pair is still significantly large; therefore, one last approximation is made. Contributions of the axial angles σ_i and σ_j are eliminated by integration, resulting in the probability

$$P_{ij} = P_{AB}(r, \tau, \theta_i, \theta_j) \quad (4)$$

From the definitions in Figure 1, the θ angles are given by

$$\cos\theta_i = \hat{d}_i \cdot \hat{r} \quad (5)$$

$$\cos\theta_j = -\hat{d}_j \cdot \hat{r} \quad (6)$$

The angle τ is related to the angle γ between the two side-chain unitary vectors $\hat{d}$ by the relation

$$\cos\gamma = \hat{d}_i \cdot \hat{d}_j = \cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j \cos\tau \quad (7)$$

Because the relation between the cosines of τ and γ is linear, it is convenient to replace τ by γ, which is simpler to compute. With this, the final form of the probability becomes

$$P_{ij} = P_{AB}(r, \gamma, \theta_i, \theta_j) \tag{8}$$

Note that a description in terms of one distance and three angles was also used by Yang and Zhou[28] to derive an atomic-level KBP and by Mukherjee et al.[26] to derive side-chain ellipsoid KBPs, although they each treated the angles independent of each other for practical reasons. The reference-state probability is chosen as

$$P^0_{ij} = P_{AB}(r_0, \gamma, \theta_i, \theta_j)(r/r_0)^2 \tag{9}$$

where r_0 is a distance at and above which the interactions can be neglected and the r^2 term accounts for the volume element. Note that the derivation above could have also been described in terms of the pair distribution function $P_{AB}(r, \gamma, \theta_i, \theta_j)/r^2$.

The pairwise energy between two residues with indices i and j is therefore

$$\begin{aligned} E_{ij} &= E_{AB}(r, \gamma, \theta_i, \theta_j) \\ &= -\ln(P_{ij}/P^0_{ij}) \end{aligned} \tag{10}$$

This energy describes the interaction between two residues with cylindrical symmetry. Because residues are not cylindrically symmetric, some corrections are required. These are introduced in the form of excluded volume energies at atomic resolution. The use of atomic resolution for nonbonded interactions goes against the justification of coarse-grained models. However, in this initial test, it helps in reducing the number of approximations, and in fact, its calculation is not too costly if ones takes advantage of the excluded volume short-range nature. Therefore, to the energy given by eq 10, an arbitrarily high energy value is added if any of the heavy atoms between the two residues overlap.

Notice that similar to the present work but in the context of KBPs, Buchete et al.[25] derived orientation-dependent probabilities between entire residues. However, they used a different approximation to reduce the six-dimensional orientational probability space given by eq 3. First, they averaged over the torsion angle (τ) and split the probability into two terms, where each residue was represented as a point particle in spherical coordinates with respect to the other, that is, $P_{AB}(r, \tau, \sigma_i, \theta_i, \sigma_j, \theta_j) \approx P_{AB}(r, \sigma_i, \theta_i)P_{BA}(r, \sigma_j, \theta_j)$. It is not evident whether this approximation is better or worse than the present one, except that the probabilities given by eq 8 are in four dimensions, involving fewer angles and higher correlations, and that in the Buchete approximation, the distance coordinate r appears twice, potentially increasing the nonadditivity, although this could be changed by the use of conditional probabilities.

Molecular Dynamics Simulations. Each of the 210 probabilities given by eq 8 are modeled using molecular dynamics by generating equilibrium ensembles for every pair of amino acid residues in water. The ensembles used are the same ones used in our previous work.[21] Their generation is summarized as follows (see original reference for details). Molecular dynamics simulations were carried out using GROMACS and the Gromos G43a1 united-atom force field. A residue pair was placed in a dodecahedron box with 25 Å diameter and solvated using SPC water. Equilibrium ensembles were obtained at 300 K from 50 independent 11 ns trajectories, with the first nanosecond discarded, and integrated in time steps of 2 fs. Configurations were stored every ps. Equilibration was tested from the convergence of the radial distribution function (RDF), judged by the deviations among the RDFs computed for each trajectory. It was assumed that the orientational-dependent distribution function also converged. On average, the simulations yielded nearly 100 points per bin, although the bins were more populated at larger distances and concentrated around equilibrium positions. It is quite likely that the densities involve significant noise fluctuations around low-probability regions. Nevertheless, it is shown in the Results and Discussion section that the level of noise is not too high, judging by the data.

To model the residues, the backbone bond angles of the amide hydrogen ($H-N-C_\alpha$) and the carbonyl oxygen ($C_\alpha -C-O$) were fixed to their average values found in peptide chains. Counterions (Cl^- or a Na^+) were added for every residue with positive or negative charge to maintain a total charge neutrality. Histidine was modeled with zero net charge, and disulfide bonds were not allowed. After the simulation, the atoms C_α, C, and O were added to the amide terminus, and the atoms N, H, and C_α were added to the carbonyl terminus of each residue to filter and eliminate configurations that occupied excluded volume sites. This was done to simulate the effects of the regular atoms of near-neighbor residues in a polypeptide. Between 2×10^5 and 5×10^5, configurations were obtained for each amino acid pair.

Local Energies. The local energy contribution is divided into two, one for the backbone and another for the side chains. While it would be desirable and convenient for consistency to use a local coarse-grained potential derived from atomic force fields, KBP-based dihedral potentials are used instead, one reason being that current atomic force fields still need some improvement in describing dihedral angle energies.[30]

KBPs for side-chain structure prediction given the backbone coordinates are readily available in terms of rotamer libraries.[31] Here, a KBP based on the side-chain dihedral angles χ with higher resolution is used.[32] It was shown that in side-chain prediction tests, the KBP performed better than potentials derived from atomic force fields. This suggests that the same could be the case for backbone dihedral angle potentials. The side-chain KBP depends only on the residue composition, the side-chain dihedral angles χ, and, conditionally, the backbone dihedral angles ϕ and ψ. It also includes correlations between consecutive dihedral angles.

The backbone potential consists of a KBP that was obtained for all amino acid triplets and includes correlations between the angles of adjacent residues.[33] In the original version, only residues not involved in hydrogen bonds were used. The idea was to remove the effects of the hydrogen bonds from the local potential. Here instead, a version that includes all residues, regardless of whether they were involved in a hydrogen bond or not, is used. As the results will show, this gives better agreement with the experimental data. The reasons for this will be discussed in the next section.

3. Results and Discussion

The nonlocal potential energies are expressed as discrete functions, with uniform intervals for each distance and orientation coordinate. Despite the large number of conformations obtained for each pair of amino acids, the sampling of configurations obtained for particular orientations was limited, which can lead to significant errors. For this reason and to limit the histogram size in computer memory, the number of

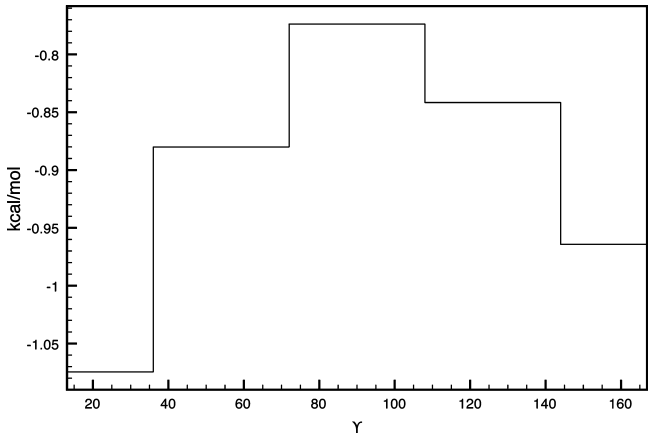

Figure 2. Phe−Phe potential as a function of the angle γ for $r = 4.5$ Å, integrated over the cos θ_i and cos θ_j coordinates.

coordinate intervals was kept relatively small. In particular, the distance r was divided into 0.5 Å intervals from 0 to 12.5 Å, while γ was divided into five intervals of 36° from 0 to 180°. It is more convenient to express the θ angles in terms of their cosines to make the random distribution of relative orientations more uniform and to simplify the calculations. Both cos θ_i and cos θ_j were divided in 0.4 size intervals from −1 to 1. The reference state distance of negligible interactions (eq 9) was set to $r_0 = 12.5$ Å.

Figure 2 shows an example of the potential between two phenylalanine residues at $r = 4.5$ Å as a function of γ and integrated over both cos θ_i and cos θ_j. Note that the distance is defined between the $C_{\delta 1}$ atoms. Phenylalanine residues attract through the entire γ range, but the plot shows a preference for the side chains to be oriented parallel (and in the same direction) rather than perpendicular. This result is in qualitative agreement to what has been observed from the analysis of structures in the Protein Data Bank (PDB).[34,35] According to Figure 2, the combined energy of the parallel and antiparallel orientations (first and last bin only) is approximately −0.5 kcal/mol more stable than the perpendicular orientation (middle bin). This value is in agreement with other reported values (−0.5 kcal/mol or lower), despite the differences in definitions (such as choosing $r = 4.5$ Å).

Figure 3 shows the probability ratio $\exp\{-E_{AB}(r, \gamma, \theta_i, \theta_j)\}$ for two phenylalanine residues as a function cos θ_i and cos θ_j for the various γ values at $r = 4.5$ Å. Darker bins are more populated. The average number of data points per bin for the corresponding histograms at this distance is 65, ranging from 0 to a maximum of 323. In terms of relative energies, the populated bins cover a range of $2.2k_BT$ (or from 0 to −1.3 kcal/mol). Some angle combinations are geometrically forbidden, resulting in empty bins. The higher-probability bins are consistent with packing configurations of the two residues. For example, when the residues are nearly parallel, $\langle\gamma\rangle = 18°$, there is a peference for the aromatic end of one residue to be located near the other ($\theta_i \approx 45°$, $\theta_j \approx -45°$, or vice versa). When they are nearly perpendicular, $\langle\gamma\rangle = 90°$, one of the preferences is also for the aromatic ends to be close ($\theta_i \approx 90°$, $\theta_j \approx 90°$). Some noise is apparent from the density fluctuations and grows for low-probability angles and shorter distances. At longer distances, the distributions become more uniform, and the energy fluctuations for low-probability angles are also significant. Note that when using two residues of the same kind, the θ angle distribution is symmetric, and twice the data are available for most bins. However, the distributions for two distinct residues are not symmetric and may contain larger fluctuations. Despite the noise, it is evident that the amount of data used to build the four dimensional probability densities provides enough signal content, in particular, at residue contact distances.

The local and nonlocal potentials were used in a coarse-grained model to simulate several small proteins by the Monte Carlo method. The method generates a series of concerted motions involving any number of atoms that efficiently generate new conformations without violating bond distance or angle constraints.[33] In tests using Gō (native biased) interactions, this method was able to fold 50−200 residue proteins in simulations lasting from 20 min to 6 days on an 81 CPU computer cluster.

Here, six proteins were used to test the potential. Their names and properties are shown in Table 1. Except for the SH3 domain, these proteins have been used in atomistic simulations elsewhere.[3,36,37] For each protein, low-energy structures were generated by simulated annealing starting from random conformations. Several hundred trajectories were simulated, each ending in a local energy minimum. For each protein, the minimum-energy structure with the lowest coordinate rmsd from native, along with its native structure, is shown in Figure 4. The lowest rmsd structure of proteins rich in helical content, that is, 1unc, 1vii, 1bdd, and 1lq7, turned out to be one of the lowest in energy as well. On the other hand, proteins with β structures generate minimum-energy structures somewhat similar to the native state but not near the lowest energies. These difficulties in folding β structures are typical in the simulations of many protein models. In the present one, extended β structures are generated but are higher in energies than helical structures. The backbone KBPs used in these simulations are likely biased toward helices.

The results presented in Figure 4 were obtained after a few trial and error variations of the potential. Originally, the model attempted to capture the effects of the hydrogen bonds in the nonlocal potential by adding orientational terms involving the interactions between the backbone hydrogen atoms in the N−H direction with the backbone oxygen atoms in the C−O direction. In addition, the effects of the hydrogen bonds were suppressed from the backbone dihedral angle potential by eliminating the residues involved in hydrogen bonds from their derivation. However, this model resulted in low-energy structures with geometries further away from native. It is intriguing that the current model performs much better, despite including the directional effects of the hydrogen bond mainly in the backbone potential. More tests are required to analyze and verify these observations. Nevertheless, the possibility remains that while hydrogen bonds are described by nonbonded interactions in the atomic force field, their effect could be better described by local interactions in coarse-grained residue pair potentials derived using the BI method.

Another variation was the selection of a backbone potential weight (i.e., multiplicative factor). The energies from the nonlocal and local side-chain potentials were added, each with a weight of 1, while the backbone potential was added with a weight of 2.5. This weight was necessary to avoid having either the nonlocal or the backbone potential dominate during the minimization, resulting in numeorus non-native energy minima. Note that the side-chain potential weight was not tested in this respect and did not seem to play a significant role, although there are simulation results suggesting that they may be important in the search of the native structure.[38] The sensitivity to these weights indicates the existence of a delicate balance between nonlocal and local interactions in coarse-grained models. From this result, it is not surprising that an improper

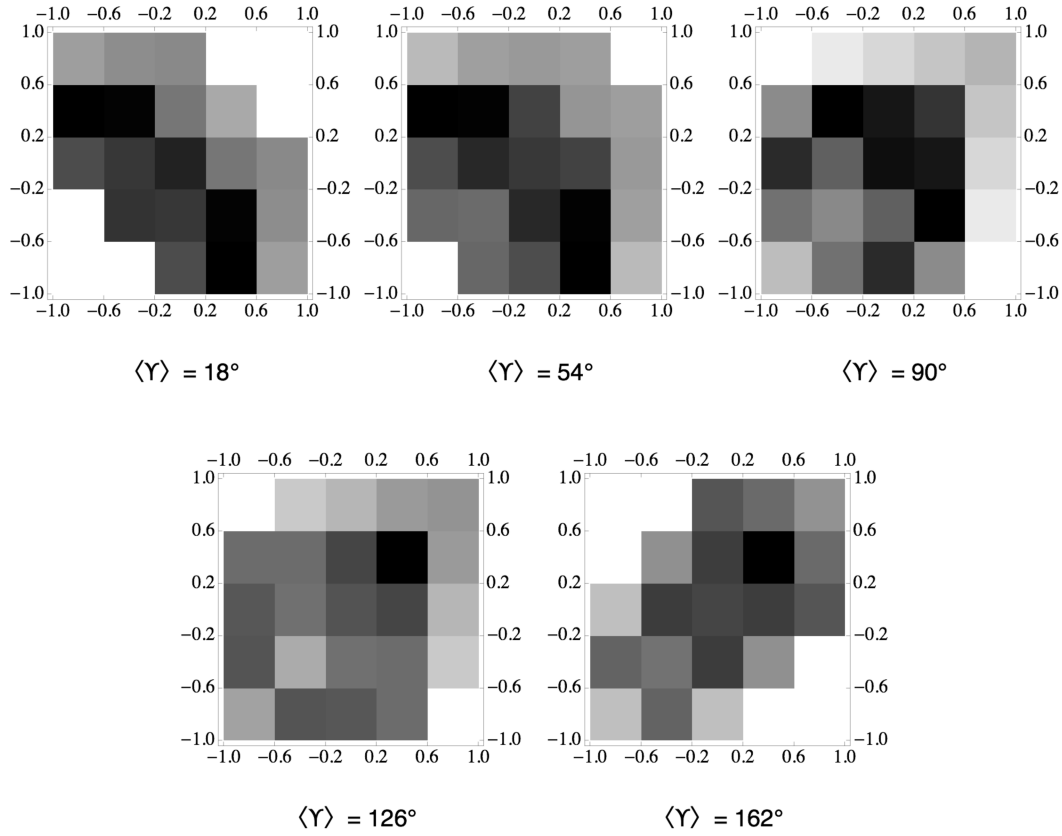

Figure 3. Phe−Phe potential as a function of cos θ_i and cos θ_j for $r = 4.5$ Å and different γ values. The gray scale is proportional to $\exp\{-E_{AB}(r, \gamma, \theta_i, \theta_j\}$, darker for more populated bins.

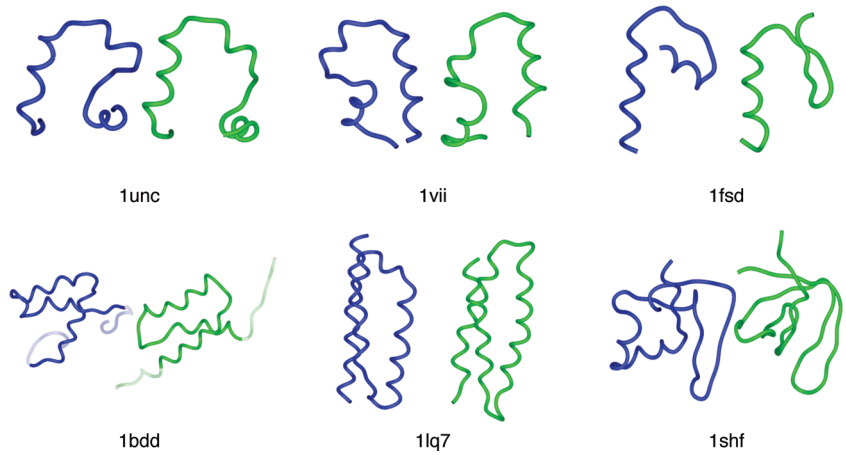

Figure 4. Folded structures (blue or left) and their corresponding native structures (green or right). The coordinate rmsd between the folded and native structures are 1unc, 2.82 Å; 1vii, 3.79 Å; 1fsd, 3.89 Å; 1bdd, 3.39 Å(9−54); 1lq7, 4.06 Å; and 1shf, 6.27 Å.

TABLE 1: Proteins Used in Folding Tests

description	PDB ID	length	structure
villin headpiece	1unc	35	α,β
HP-36	1vii	36	α
protein A	1bdd	60	α
de novo	1fsd	28	α,β
de novo	1lq7	67	α
SH3 domain	1shf	59	β

TABLE 2: Folding Results for the Orientation Models[a]

	lowest rmsd			lowest energy			
PDB ID	rmsd	rank	Z score	rmsd	rank	Z score	total
1unc	2.82	24	−0.93	3.80	4	−2.40	1280
1vii	3.79	76	−1.50	4.68	23	−2.05	1376
1bdd	3.39	15	−1.35	4.40	5	−2.80	1568
1fsd	3.89	50	−0.47	5.80	67	−2.09	1234
1lq7	4.06	43	−1.68	4.14	2	−2.17	1837
1shf	6.27	33	−0.72	8.92	128	−2.88	1434

[a] The lowest-energy structure rank is based on the rmsd, and the lowest rmsd structure rank is based on energy. Total is the total number of minimized structures.

balance between helical and extended backbone conformations can lead to improper folding in protein models.[30]

Table 2 shows some of the properties of the energy-minimized structures with the orientation-dependent potential. The table shows the coordinate rmsd, the rank, and the Z score for the lowest rmsd structure and for the lowest-energy structure. The coordinate rmsd is measured with respect to the native structure identified by the PDB ID. The rank for the lowest rmsd structure

TABLE 3: Folding Results for Distance-Dependent-Only Models[a]

	lowest rmsd			lowest energy			
PDB ID	rmsd	rank	Z score	rmsd	rank	Z score	total
1unc	2.82	86	−0.35	8.26	987	−0.81	1280
1vii	3.79	51	−0.51	7.61	902	−0.84	1376
1bdd	3.39	84	−0.46	12.24	1233	−0.98	1568
1fsd	3.89	20	−0.44	5.79	65	−0.71	1234
1lq7	4.06	39	−0.52	10.13	235	−1.13	1839
1shf	6.27	33	−0.75	10.60	763	−0.95	1434

[a] The lowest-energy structure rank is based on the rmsd, and the lowest rmsd structure rank is based on energy. Total is the total number of minimized structures. The structures were obtained by minimizing the orientational-dependent model and evaluated by the distance-dependent-only energies.

is in relation to the structures sorted by energy and, for the lowest-energy structure, is in relation to the structures sorted by rmsd. The Z score is defined as $(E - \langle E \rangle)/\sigma_E$, where $\langle E \rangle$ is the average energy of the energy-minimized structures, σ_E is the standard deviation, and E the energy of either the lowest rmsd or the lowest-energy structure. The total number of structures is also shown. Overall, the lowest-energy structures were ranked significantly high in terms of their rmsd. The ranking was lower for the β-rich structures than that for the others. On the other hand, the rank of the lowest rmsd structures was not that different between the β- and α-rich structures. That is, native-like structures are some of the lowest in energy, but not all low-energy structures are native-like.

It should be mentioned that the use of distance-dependent nonbonded energies without orientations gives somewhat inferior results than those with orientations. For example, folding tests performed with a villin headpiece and the distance-dependent potential without orientations[21] resulted in an optimal structure with a rmsd of 3.33 Å from native, compared to 2.82 Å using orientations, and a lowest-energy structure of 4.65 Å rmsd compared to 3.80 Å rmsd with orientations. This is a small difference but indicates that the current implementation of the orientations can improve the nonbonded residue pair potential. Larger improvements may be possible by improving the sampling resolution and ensemble sizes.

Another test was made by evaluating the energy without orientation dependencies on the conformations that resulted from the minimization of the orientation-dependent potential. For comparison, the distance partition was also selected as 0.5 Å intervals. These results are shown in Table 3. With a few exceptions (such as 1 fsd), the ranks, Z scores, and rmsd values are better with orientations than without them. However, the rankings of the lowest rmsd structures are comparable to the orientation-dependent ones. This implies that the orientation dependencies mainly help in reducing the number of low-energy non-native structures. Note that this is not a totally fair test given that the structures were minimized using the orientation-dependent model. As a results the distance-dependent energy distributions are highly skewed toward higher energies resulting poor Z scores.

4. Conclusion

It has been demonstrated that the sampling of residue pairs in equilibrium conformations using atomic force fields and the BI method can be used to parametrize coarse-grained interactions between residues with orientation dependencies. In particular, the combination of nonlocal interactions that include residue pairwise distances and orientations, with the addition of local KBPs for backbone and side-chain dihedral angles, allowed the generation of low-energy conformations close to the native structure. Finer-grained excluded volume energies were also necessary to avoid low-energy nonphysical conformations. The nonlocal interactions defined at the residue level reduces the computational cost over atomic-level models and still gives relatively good agreement with experimental results. Despite this, it is evident that the model needs significant improvements. The amount of data used to generate the coarse-grained potential was marginal at best, and large ensembles or better sampling are likely required to improve the quality of the coarse-grained energy function. Native structures rich in β sheets are not identified as the lowest-energy structures. Improvements in the local dihedral angle potentials are necessary, and their derivation from atomic simulations will be required for consistency. In addition, a finer angle partition of the directional dependencies in the nonlocal potential and the inclusion of other neglected directional terms accounting for the hydrogen bonding effects may be necessary to improve the model.

Unlike models based on KBPs, the atomic force-field-derived coarse-grained potential has the advantage of being adaptable to various solvent conditions, temperatures, and pressures. Despite the limitations in the quality of atomic force fields and current approximations, the model showed its potential to model the protein structure efficiently.

References and Notes

(1) Snow, C. D.; Nguyen, H.; Pande, V. S.; Grueble, M. *Nature* **2002**, *420*, 102–106.
(2) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Annu. Rev. Biophys.* **2008**, *37*, 289–316.
(3) Meinke, J. H.; Hansmann, U. H. E. *J. Comput. Chem.* **2009**, *30*, 1642–1648.
(4) Levitt, M.; Warshel, A. *Nature (London)* **1975**, *253*, 694–698.
(5) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59–107.
(6) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849–873.
(7) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874–887.
(8) van Giessen, A. E.; Straub, J. E. *J. Chem. Theory Comput.* **2006**, *2*, 674–684.
(9) Bereau, T.; Deserno, J. *J. Chem. Phys.* **2009**, *130*, 23510/6–23510/15.
(10) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945–950.
(11) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859–883.
(12) Skolnick, J.; Jaroszewski, L.; Kolinski, A.; Godzik, A. *Protein Sci.* **1997**, *6*, 676–688.
(13) Miyazawa, A.; Jernigan, R. L. *Proteins* **1999**, *34*, 49–68.
(14) Regan, L.; Woolfson, D. N. *Curr. Opin. Struct. Biol.* **2008**, *18*, 475–476.
(15) Kryshtafovych, A.; Fidelis, K.; Moult, J. *Proteins* **2009**, doi: 10.1002/prot.22562.
(16) Betancourt, M. R. *Proteins* **2009**, *76*, 72–85.
(17) Basdevant, N.; Borgis, D.; Ha-Duong, T. *J. Phys. Chem. B* **2007**, *111*, 9390–9399.
(18) Ayton, G. S.; Noid, W. G.; Voth, G. A. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.
(19) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289–4303.
(20) Sherwood, P.; Brooks, B. R.; Sansom, M. S. P. *Curr. Opin. Struct. Biol.* **2008**, *18*, 630–640.
(21) Betancourt, M. R.; Omovie, S. J. *J. Chem. Phys.* **2009**, *130*, 195103.
(22) Thorpe, I. F.; Zhou, J.; Voth, G. A. *J. Phys. Chem. B* **2008**, *112*, 13079–13090.
(23) Tobi, D.; Shafran, G.; Linial, N.; Elber, R. *Proteins* **2000**, *40*, 71–85.
(24) Vendruscolo, M.; Domany, E. *J. Chem. Phys.* **1998**, *109*, 11101–11108.
(25) Buchete, N. V.; Straub, J. E.; Thirumalai, D. *J. Mol. Graph. Model.* **2004**, *22*, 441–450.
(26) Mukherjee, A.; Bhimalapuram, P.; Bagchi, B. *J. Chem. Phys.* **2005**, *123*, 014901.
(27) Lu, M.; Dousis, A. D.; Ma, J. *J. Mol. Biol.* **2008**, *376*, 288–301.
(28) Yang, Y.; Zhou, Y. *Proteins* **2008**, *72*, 793–803.

(29) Fang, Q.; Shortle, D. *Proteins* **2005**, *60*, 90–96.
(30) Jang, S.; Kim, E.; Pak, Y. *Proteins* **2007**, *66*, 53–60.
(31) Dunbrack, R. L., Jr; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661–1681.
(32) Betancourt, M. R. Submitted.
(33) Betancourt, M. R. *J. Phys. Chem. B* **2008**, *112*, 5058–5069.
(34) McGaughey, G. B.; Gagné, M.; Rappé, A. K. *J. Biol. Chem.* **1998**, *273*, 15458–15463.
(35) Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210–1250.
(36) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740–744.
(37) Kim, S. Y.; Lee, J.; Lee, J. *Biophys. Chem.* **2005**, *115*, 195–200.
(38) Wei, Y.; Nadler, W.; Hansmann, U. H E. *J. Chem. Phys.* **2008**, *128*, 025105.

JP906710C