JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

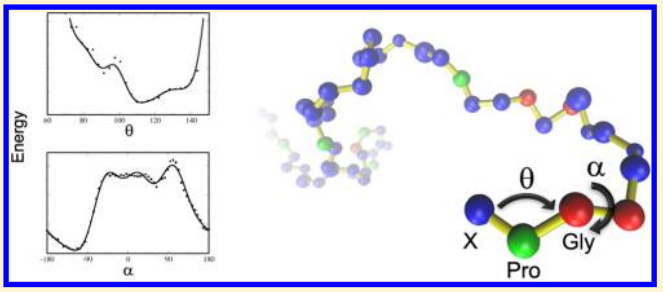
Coarse-Grained Potentials for Local Interactions in Unfolded Proteins

Ali Ghavami,[†] Erik van der Giessen,[†] and Patrick R. Onck*[,†]

[†]Micromechanics of Materials, Zernike Institute for Advanced Materials, University of Groningen, 9747 AG Groningen, The Netherlands

Supporting Information

ABSTRACT: Recent studies have revealed the key role of natively unfolded proteins in many important biological processes. In order to study the conformational changes of these proteins, a one-bead-per-amino-acid coarse grained (CG) model is developed, and a method is proposed to extract the potential functions for the local interactions between CG beads. Experimentally obtained Ramachandran data for the coil regions of proteins are converted into distributions of pseudo-bond and pseudo-dihedral angles between neighboring alpha-carbons in the polypeptide chain. These are then used to derive bending and torsion potentials, which are residue and sequence specific. The validity of the developed model is testified by studying the radius of gyration as well as the hydrodynamic properties of chemically denatured proteins.



1. INTRODUCTION

Despite the classical view that a protein can attain its biological function only upon folding into a unique structure, there is increasing evidence that unfolded proteins play a key role in many important biological functions.[1,2] The basic functions of this class of proteins exploit the absence of a stable secondary structure in their polypeptide chain. Studies of the role of unfolded proteins in different molecular processes suggest that these proteins can be classified into four functional groups, involved in molecular recognition and signaling, protein modification, molecular assembly, and entropic chain activities (such as linkers/spacers, bristles, springs).[2−4] Rapid increase in our knowledge on structures and functions of unfolded proteins has disclosed that many proteins and protein domains are intrinsically unstructured, and, more interestingly, their relative amount increases in more evolved and complex organisms.[5] All of these findings indicate the importance of unfolded proteins and the necessity of developing new approaches to study them in more detail.

Although atomic-level molecular dynamics simulations have achieved a high level of reliability and accuracy, they are still not able to reach biologically interesting time and length scales due to the limitations in computational resources. The situation is even worse for unfolded proteins since by their nature they are more dynamic, and therefore longer simulations are required in order to obtain statistically meaningful results. These limitations have drawn the interest of researchers toward the development of coarse-grained (CG) models to reduce the degrees of freedom and, thus, to increase the spatial and temporal domains of interest.

Available CG models can be categorized into different classes according to the level of coarse-graining, the treatment of the solvent environment, and the method used for the force field parametrization.[6] Among them, the CG models which describe each amino acid with one or a few beads have been successful in representing structural detail of proteins while maintaining computational efficiency.[7] Coarser models have not gained much interest mainly because the secondary structure cannot easily be described and the force field cannot be informed from experimental data in a straightforward manner.[8] The solvent environment is of considerable importance in molecular simulations because many of the molecular driving forces like hydrophobicity, electrostatic interactions, and hydrogen bonding are directly or indirectly affected by the properties of the solvent.[9] In order to decrease the computational time, many CG models do not explicitly include solvent molecules into the simulations.[10−13] Instead, the solvent effects are implicitly included in their force field.[14] In all CG models, there is a trade-off between predictive power and accuracy which, to a great extent, is related to the parametrization strategy used in development of the force field. Depending on the functional form of the force field and available data for parametrization, several approaches such as elastic network,[15] Boltzmann inversion,[12] force matching,[16] or a combination of these methods could be employed to determine the force field.

A short literature review on the available CG models and their features shows that there has not been much effort on developing CG models to study the disordered state of proteins. Elastic network and Go models are simple, but their force fields are completely biased to a unique reference structure. In more complex CG models like MARTINI,[17] the Head-Gordon model,[18] and the model developed by Korkut and Hendrickson,[11] a priori knowledge of the local secondary structure of the proteins is required to perform the simulations. On the other hand, CG models that can predict the folded structure of proteins to some extent, like the force fields developed by

Received: May 2, 2012
Published: November 5, 2012

Tozzini et al.[19,20] and Bereau and Deserno,[21] are parametrized using databases of folded protein structures which do not guarantee to give the correct ensembles for unfolded proteins.

In this work, we propose a one-bead-per-amino-acid model for unfolded proteins. In order to obtain the local interaction potentials, experimentally obtained Ramachandran plots for the coil regions of proteins are converted into distributions of pseudo-bond and pseudo-dihedral angles between neighboring α-carbons in the polypeptide chain. These distributions are used to derive bending and torsion potentials that are residue and sequence specific. Potentials with different levels of accuracy are proposed, and an assessment study is then performed on the generated potential functions to select the best set of potentials in terms of simplicity and accuracy. Finally the model is used to study chemically denatured proteins and the effect of sequence composition on them.

2. EXTRACTION METHOD TO OBTAIN COARSE-GRAINED POTENTIALS

2.1. Mapping Backbone Internal Degrees of Freedom (ϕ, ψ) to Pseudo-Bending and Torsion Angles (θ, α). A geometrical representation of a coarse-grained polypeptide chain together with the CG degrees of freedom is shown in Figure 1(b). In the all-atom representation of the backbone (Figure 1(a)), the

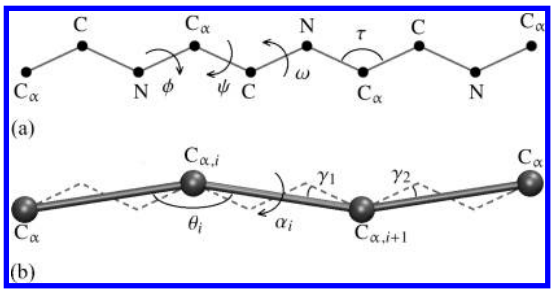


Figure 1. All atom schematic of a polypeptide chain (a) and coarse-grained representation (b) of the backbone with pseudo-bending and torsion angles. In (b) the dashed lines represent the polypeptide chain, and the solid lines are the pseudo-bonds between C_α carbons which represent the coarse-grained geometry.

bond lengths and bond angles display only a small variation from their average value[22] so they are assumed to remain fixed in the present work. The average bond lengths of C_α–N, C_α–C, and C–N are 0.145 nm, 0.152 nm, 0.133 nm, respectively, with the average bond angles C_α–C–N = 116°, C–N–C_α = 122°, and N–C_α–C = τ = 111°.[23] A *trans* conformation is presumed for the peptide bond (ω = 180°), and the rare possibility of *cis* conformation is ignored. The stated assumptions imply that the dihedral angles ϕ and ψ (see Figure 1(a)) are the only degrees of freedom of the all-atom backbone.

Figure 1(b) demonstrates the CG representation of the polypeptide chain by connecting the α-carbons through pseudo-bonds. With the assumptions above, the pseudo-bond lengths between subsequent C_α's remain fixed at a distance of 0.38 nm as defined by the geometry. The pseudo-bending angle θ and pseudo-dihedral angles α for the CG chain are defined between three and four consecutive C_α's. The relationship between the CG (θ,α) and all-atom (ϕ,ψ) degrees of freedom can be established by explicit geometrical expressions.[19,24] The pseudo-bending angle θ_i of bead i of the coarse-grained chain is thus found to be given by

$$\begin{aligned}\cos\theta_i = {} & \cos\tau(\cos\gamma_1\cos\gamma_2 - \sin\gamma_1\sin\gamma_2\cos\phi_i\cos\psi_i) \\ & - \sin\gamma_1\sin\gamma_2\sin\phi_i\sin\psi_i \\ & + \sin\tau(\cos\psi_i\sin\gamma_1\cos\gamma_2 + \cos\phi_i\cos\gamma_1\sin\gamma_2)\end{aligned} \quad (1)$$

where γ_1 = 20.7°, γ_2 = 14.7°, and τ = 111° are constant angles (see Figure 1(b)). The following approximate formula has been suggested for the pseudo-torsion angle:[19]

$$\alpha_i = 180^\circ + \psi_i + \phi_{i+1} + \gamma_1\sin\psi_{i+1} + \gamma_2\sin\phi_i \quad (2)$$

In contrast to the pseudo-bending angle, which depends only on one set of backbone dihedral angles (ϕ_i,ψ_i), the pseudo-torsion angle is a function of two consecutive sets of backbone dihedral angles (ϕ_i,ψ_i,ϕ_{i+1},ψ_{i+1}). It is important to note that in the force fields developed specifically for well-defined secondary structures,[19,24] the simplifying assumption $\phi_i = \phi_{i+1}$,$\psi_i = \psi_{i+1}$ is made for mapping α. However, for proteins without any regular structures, this assumption is not justified. This will be discussed in more detail in the following sections regarding the extraction of torsion potentials.

Besides the approximate formula discussed above, an exact method has been adopted to calculate the pseudo-torsion angle from the backbone dihedral angles. In this method, a small all-atom chain is built using four consecutive dihedral angles with the geometrical data mentioned earlier. The pseudo-torsion angles are then measured directly from the coordinates of the α-carbons of the generated chain by means of simple vector algebra (see Flory[25] for details). This will allow for more precise extraction of pseudo-torsion potentials from all-atom experimental data and will be used subsequently to verify the accuracy of the use of eq 2.

2.2. Coil Library. The backbone ϕ and ψ values of proteins are often presented in two-dimensional density plots, called Ramachandran plots. It turns out that the Ramachandran space [−180°,180°) → [−180°,180°) is divided into several regions, each referring to a specific secondary structure. The empty regions refer to unfavorable conformations which are mainly caused by the steric clash between neighboring side chains or steric hindrance to the formation of hydrogen bonds between peptide groups and water molecules.[26] We use these density plots to generate the mean force potentials for local interactions in the unfolded state. For this purpose, we adopt the Boltzmann inversion method

$$U(q) = -k_B T\ln(P(q)) \quad (3)$$

where q is any desired degree of freedom, $P(q)$ is the probability distribution for q, T is the temperature, and k_B is the Boltzmann constant.

Special care must be taken while choosing the appropriate Ramachandran plots. The required density plots should not include data related to any secondary structure, and also since we are interested in local interactions, long-range effects must be absent or have a negligible impact on the density plots. The data that satisfy these conditions best are from protein coil regions. Coil regions are those parts of proteins that cannot be classified to any kind of known secondary structure. This implies that their backbone conformations are not biased to any secondary structure. It has been shown that the intrinsic backbone preferences of dipeptides are strikingly similar to the backbone conformations of coil regions of proteins[26] confirming the assumption that long-range hydrophobic or electrostatic

dx.doi.org/10.1021/ct300684j | J. Chem. Theory Comput. 2013, 9, 432−440

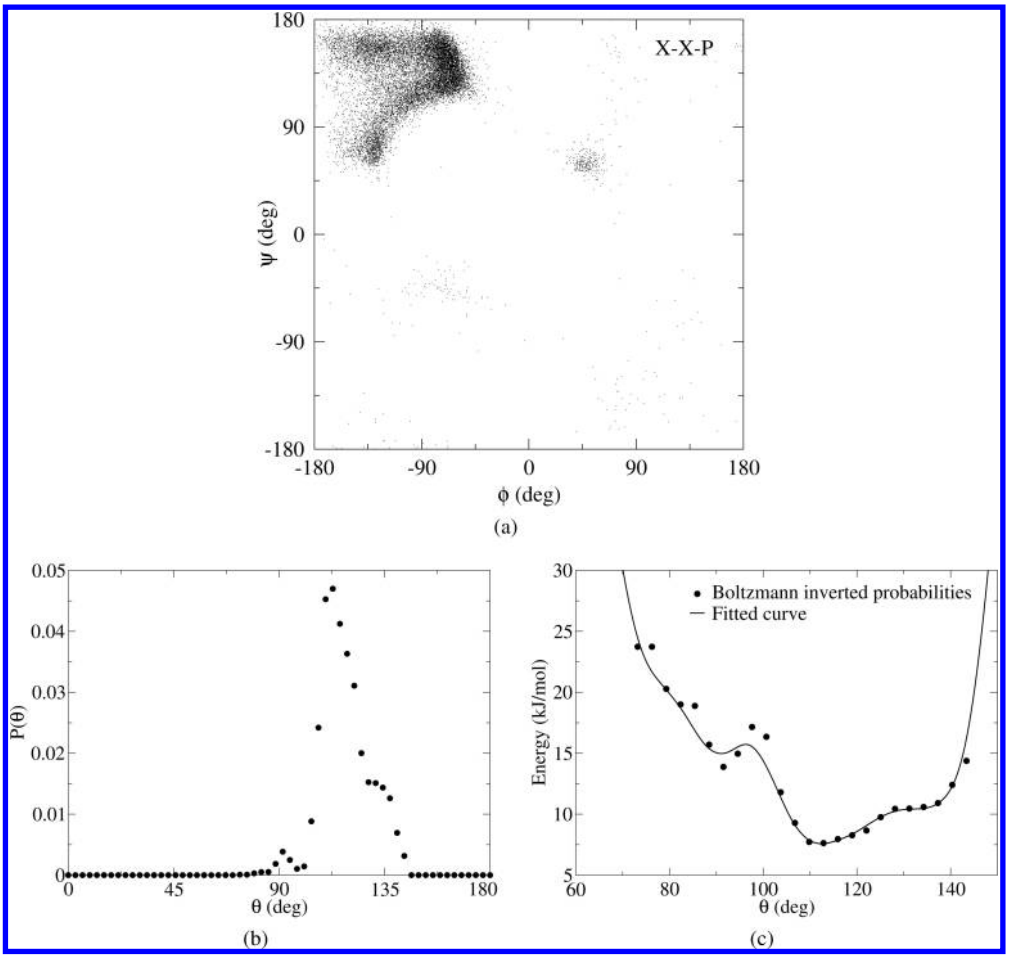


Figure 2. Extraction procedure for the bending potential of X-X-P combinations. (a) Ramachandran data for X-X-P combinations extracted form the coil library. This plot for X-X-P contains 3559 points. (b) Normalized distribution of the bending angle θ, which is obtained by mapping all the Ramachandran data from (a) to the pseudo-bending angle θ through eq 1 and counting the frequencies in each bin (number of bins m = 60).[29] (c) Obtained bending potential, $U(\theta)$ after applying the Boltzmann-inversion method on the probability distribution $P(\theta)$ presented in (b) and fitted by a piecewise polynomial.

interactions are absent or have a very small effect on the Ramachandran data obtained from this class of residues. The DASSD library is used to extract Ramachandran plots of the coiled regions of proteins. We extract the dihedral angles of the central residues of short fragments (of lengths 1, 3, and 5) from the coiled regions, which gives the possibility to extract meaningful potentials by considering the effect of neighboring residues on the obtained potentials. The database is generated from 5,227 nonredundant high resolution (less than 2 Å) protein structures.[27]

2.3. 3-Letter Amino Acid Model. The current size of the coil library is not large enough to extract CG bending and torsion potentials for all 20 amino acids accounting for all possible neighbors. This would give a huge number of different potentials which is not in accordance with the aim of coarse graining and data reduction. To reduce the number of different potentials, the amino acids are categorized into several subgroups based on the similarities between their Ramachandran data. Four basic types of Ramachandran plots have been reported in the literature depending on the stereochemistry of the amino acids: glycine, proline, "generic" (which refers to the remaining 18 amino acids), and "pre-proline" (which refers to residues preceding a proline).[28] Since the Ramachandran plots are the main input to generate the CG potentials, different potential functions are expected for glycine (G), proline (P), and the rest of the amino acids (X) depending on their neighbor residues.

2.4. Bending Potentials. Pseudo-bending potentials are obtained by Boltzmann-inversion of the θ probability distribution. Initially, ϕ and ψ dihedral angles for the central residue of different triple combinations of P, G, and X are extracted from the 3-residue-fragments in the coil library. A table with the number of data points for each combination is shown in the Supporting Information (Table S1). The extraction procedure is depicted schematically in Figure 2 for X-X-P combinations. In Figure 2(a) the ϕ and ψ values are plotted for all X amino acids (i.e., those amino acids that are not P or G) that have an X preceding it and a P following it. In the next step, each data point in the Ramachandran space is mapped to θ-space using eq 1. Collecting all data points in data bins (here we use m = 60 data bins[29]) gives the θ probability distribution (Figure 2(b)) which is then directly converted into the bending potential by eq 3 (Figure 2(c)).

In order to study the neighbor-residue effect on the obtained bending potentials, three levels of accuracy have been investigated. The level-1 bending potentials are obtained by picking out Ramachandran data for all 27 combinations of G, P, and X and then extracting bending potentials by the procedure

dx.doi.org/10.1021/ct300684j | J. Chem. Theory Comput. 2013, 9, 432−440

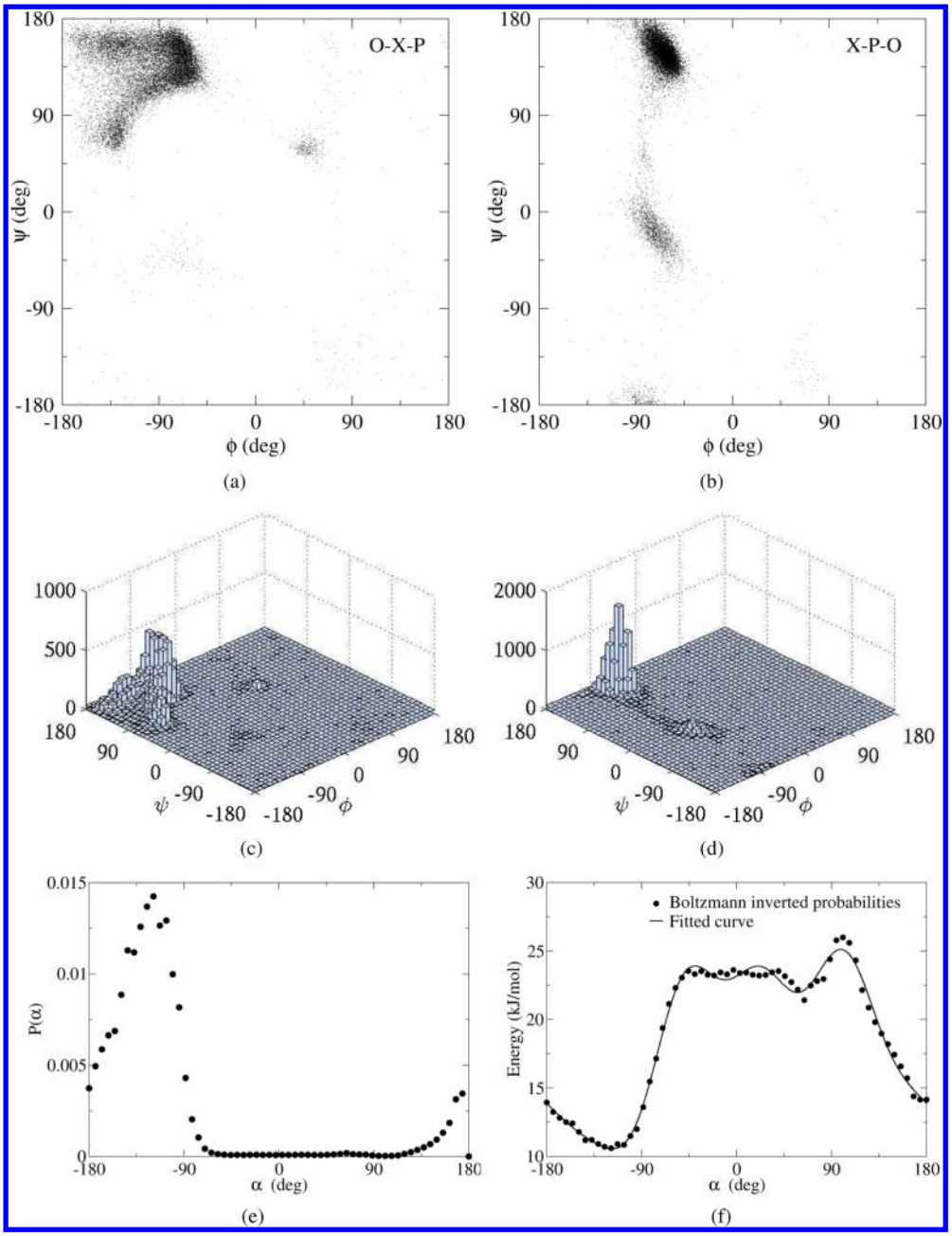


Figure 3. Extraction procedure for the torsion potentials for X-P combinations. (a) Ramachandran data for O-X-P combinations in the coil library, i.e. ϕ and ψ values for the central residue X that has an O (any residue) preceding it and a proline following it. (b) Ramachandran data for X−P-O combinations. (c), (d) Frequency plots for Ramachandran data of O-X-P and X-P-O combinations in the coil library using a $n^2 = 40 \times 40$ grid. (e) Normalized probability distribution for the torsion angle α after mapping Ramachandran data from (c) and (d) to the pseudo-torsion angles and counting the frequencies in each bin (number of bins $m = 60$). (f) Obtained torsion potential after applying Boltzmann inversion on the probability distribution presented in (e).

explained in Figure 2. The level-1 bending potentials are shown in the Supporting Information, Figures S1 to S3. For special combinations the number of available data points is not sufficient to generate statistically meaningful potentials, so that the data points from the closest potential from the other combinations were added to construct the potentials. [In some special cases (GGG, GGP, PGG, PGP, GPG, GPP, PPG, and PPP) the number of Ramachandran data points is not sufficient. However, even for these limited number of data points, the obtained potential functions show a local minimum. In these cases, among the other potentials in the family with the same central residue, the data points from the one with the same local minimum (XGG, XGP, GGX, XGP, XPG, XPP, PPX, and XPP, respectively) were added to make a combined data set in order to generate the potentials.] Level-2 potentials consist of a reduced version of level-1 in which we distinguish those central residues that precede and do not precede a proline (P), resulting in 6 different potentials (O-*Z*-Y and O-*Z*-P, with $Z \in${G, P, X}, Y=X+G, O=X+G+P). The level-2 bending potentials are shown in Figure 9 of the Appendix. For the level-3 potentials no neighbor-residue effect is included, which means that the Ramachandran data are extracted for each residue regardless of the neighboring residues, resulting in three separate bending potentials for G, P, and X (see Supporting Information, Figure S4).

dx.doi.org/10.1021/ct300684j | J. Chem. Theory Comput. 2013, 9, 432−440

2.5. Torsion Potentials. A similar methodology as described above is applied to derive the pseudo-torsion potentials. The main difference with the bending procedure is that two separate sets of Ramachandran data are required to convert the all-atom dihedral angles ϕ and ψ to the coarse-grained dihedral angle α according to eq 2. As a result, torsion potentials are derived for all possible double-combinations of amino acids, giving nine different torsion potentials for the 4-residue fragments O-Z-Z′-O with $\{Z,Z'\} \in \{X,G,P\}$. This allows to sample the first set of all-atom dihedral angles from the Ramachandran data of the first residue Z and sample the second set from the Ramachandran data of the next residue Z' (see Figure 3(a) and (b)). However, sampling out of two separate Ramachandran plots results in an enormous number of possible combinations, because each Ramachandran plot may contain more than 10^5 data points. Therefore, we first divide each Ramachandran plot into a limited number of cells using an $n \times n$ grid (see Figure 3(c) and (d)) and characterize each cell by the coordinate of the cell center, e.g. $(\phi,\psi)_I$ and a weight factor N_I representing the number of data points in that cell. Then a pseudo-torsion angle α_k is obtained by picking two sets of dihedral angles (e.g., $(\phi,\psi)_I$ from the first Ramachandran plot and $(\phi,\psi)_J$ from the second Ramachandran plot where $I,J = 1,...,n^2$) giving n^4 different pseudo-torsion angles α_k in the range $(-180°,180°)$. Each α_k value is assigned a weight factor $N_k = N_I \times N_J$. In order to find the probability distributions, the computed α_k's are distributed over m bins ranging from $-180°$ to $180°$, and the frequency of occurrence of α_p in each bin $p \in \{1,m\}$ is obtained by $N_p = \sum N_k$. The resulting normalized distribution function P (Figure 3(e)) is then Boltzmann inverted using eq 3, and the torsion potential is determined by fitting a goniometric polynomial (Figure 3(f)) (The polynomial and coefficients can be found in eq S1 and Table S2 of the Supporting Information.).

Torsion potentials have also been developed using two levels of accuracy: For level-1 torsion potentials, two sets of Ramachandran data have been extracted, taking into account the effect of sequence. For example, in order to obtain the torsion potential for the X-P combination, the first set of Ramachandran data corresponds to the central residue (X) of the 3-residue segment O-X-P and the second one corresponds to P from X-P-O (see Figure 3 (a) and (b)). It should be noted that the size of the current coil library is insufficient to include the effect of other neighbors (e.g., subdividing O into G, P, and X and generating potentials for different combinations). Torsion potentials of level-2 are less accurate by construction, as two sets of Ramachandran data are selected regardless of the sequence. For example, to obtain the torsion potential for the X-P combination, the first set of Ramachandran data corresponds to the X residue from O-X-O segments and the second to the P residue from O-P-O, also resulting in 9 torsion potentials.

In order to study the effect of the accuracy of the mapping scheme of eq 2 on the results, the level-1 potentials have been extracted using both the approximate eq 2 and the exact vector algebra method (see the last paragraph of subsection 2.1). A summary of the developed bending and torsion potentials is given in Table 1.

3. ASSESSMENT OF THE POTENTIAL FUNCTIONS

In this section, we will report on molecular dynamics (MD) simulations performed to study the effect of the different levels of accuracy for the developed potentials on the radius of gyration R_g of arbitrary CG chains. The simulations are performed with

Table 1. Bending and Torsion Potentials Developed with Different Levels of Accuracy

bending		description
level-1	Ben-1	27 potentials for all possible combinations of G, P, and X
level-2	Ben-2	6 potentials for G, P, and X and their preproline variants
level-3	Ben-3	3 potentials for G, P, and X, regardless of the neighbor residues
torsion		**description**
level-1	Tor-1	The neighbor residue effect is included, and mapping is done by eq 2.
level-1	Tor-1-ex	The neighbor residue effect is included, and mapping is done by vector algebra.
level-2	Tor-2	The neighbor residue effect is not included, and mapping is done by eq 2.

GROMACS version 4.0.7[30] in canonical ensemble at a temperature of T = 300 K using a Langevin thermostat. The equations of motion are integrated with a time-step of 0.02 ps for at least 10^7 MD steps, and the first 10^6 steps are not considered in the averaged R_g and R_s values reported in this study (see Figures S5 and S6 for a convergence study and time-step size selection). The translation of the center of mass of the chain is removed at each time-step, but the chain is allowed to freely rotate. All CG beads have a mass of 124 amu (i.e., the average mass of all amino acids), and the bond lengths are kept fixed by a stiff harmonic potential of the form $U_{bond} = K(r-b)^2$ with K = 8038 kJ/nm^2/mol[31] and b = 0.38 nm and r is the distance between two consecutive beads. The excluded volume effect is incorporated using a purely repulsive potential with a repulsion radius of R_{rep} = 0.38 nm. This is chosen based on the proposed values in the literature for short charge-free polypeptides denatured in urea and GdmCl.[32]

The simulations have been carried out for chains composed of 100 CG beads with different amounts of G, P, and X residues. The results are presented in terms of diagrams (see Figure 4 and Figure 5), in which each curve represents the R_g of a two-phase chain characterized by the pair (phase 1, phase 2) of the number of beads in phase 1 and phase 2, respectively. When the chains are composed of mixed phases, the residues are randomly distributed along the chains. Each data point shows the average of 8 simulations in which the specific sequence is chosen at random.

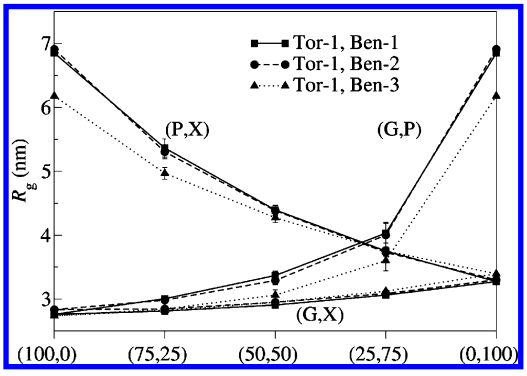


Figure 4. The average radius of gyration R_g of proteins with different combinations of X, P, and G residues (P,X), (G,X), and (G,P) is used to evaluate the accuracy of the obtained bending potentials. The horizontal axis of the diagram shows the number of residues of each phase. The error bars are the standard deviation from the mean values of 8 simulations.

In Figure 4, the bending potentials of level-1, -2, and -3 are evaluated for the level-1 torsion potential (indicated by Tor-1,

dx.doi.org/10.1021/ct300684j | J. Chem. Theory Comput. 2013, 9, 432−440

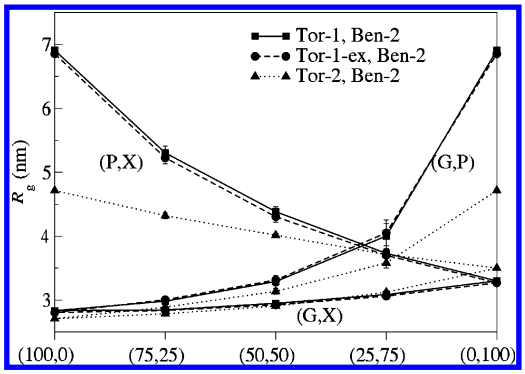


Figure 5. The effect of different torsion potentials on the obtained R_g has been studied for different chain compositions. It is clear that level-2 torsion potentials deviate considerably from the level-1 potentials. In addition, the results show that using the approximate method eq 2 for mapping backbone ϕ and ψ angles to the pseudo-dihedral angles induces a negligible error in the overall dimensions of the chain. The error bars indicate the standard deviation from the mean values of 8 simulations.

see Table 1). As can be seen from the three combinations of phases (P,X), (G,X), and (G,P), a higher value of R_g is predicted as the amount of P residues in the chain is increased, while increasing the amount of G residues always results in a lower R_g. This is due to the fact that the P residues induce a higher local stiffness and the G residues induce a lower local stiffness than the rest of the amino acids,[33] which is captured in the developed potentials. The results for the level-2 bending potentials (Ben-2) show only a slight difference compared to the level-1 potentials (Ben-1), while the level-3 potentials profoundly underestimate the radius of gyration. This clearly demonstrates that accounting for the pre-proline variants is essential, whereas accounting for all possible G, P, and X combinations does not add additional accuracy.

A similar procedure is pursued for evaluating the torsion potentials. Figure 5 shows the same diagram for the level-1 and level-2 torsion potentials (Tor-1 and Tor-2), while keeping the bending potentials fixed (level-2, Ben-2). An additional set of simulations is carried out using the level-1 torsion potentials generated by the exact mapping scheme described earlier (Tor-1-ex, see Table 1). The graph clearly depicts the big effect of neighboring residues on the overall dimensions of the CG chain. It is interesting to see that the level-2 torsion potentials predict a lower R_g compared to the more detailed level-1 torsion potentials for chains composed of P and G residues but, on the other hand, predict a larger R_g for chains mainly composed of X residues. The reason could be explained in the context of pre-proline residues which have a stiffer and extended backbone conformation.[33] Since no neighbor effect is accounted for in extracting the Tor-2 potentials, the pre-proline conformations are also included in potentials with central residue X (resulting in a higher R_g for purely X-containing chains). On the other hand, the stiffness of the potentials for pre-proline residues is reduced by including other less-stiff conformations (resulting in lower R_g for purely P-containing chains). The close agreement between Tor-1 and Tor-1-ex also demonstrates that eq 2 gives a good approximation for the pseudo-torsion angles (α) as a function of the all-atom dihedral angles. To estimate the accuracy of the level-1 torsion potentials, relative to level-2, we have investigated an even more accurate torsion potential for one point of Figure 5, namely (P,X) = (100,0) for which the difference between level-1 and level-2 is the largest. The generated torsion potential (using two sets of PPP Ramachandran data instead of using OPP and PPO) results in R_g = 6.80 nm. The predicted radius of gyrations using To1-1, Tor-1-ex, and Tor-2 are R_g = 6.91, R_g = 6.85, and R_g = 4.72, respectively, clearly showing the enhanced accuracy of Tor-1 compared to Tor-2.

The assessment carried out in this section indicates that the level-2 bending potentials together with the level-1 torsion potentials can be successfully used to describe the essential local interactions in our CG model for unfolded proteins. The level-2 bending potentials and the level-1 torsion potentials are presented in the Appendix (Figures 9 and 10, respectively).

To investigate the relative influence of R_{rep} and the developed potentials on R_g, we carried out some test simulations on a N = 100 chain (all X residues). By using R_{rep} = 0.38 nm and the developed torsion and bending potentials, the calculations predict R_g = 3.57 nm, while R_g = 2.59 nm is predicted for a freely jointed chain (i.e., same length, but without the bending and torsion potentials). This clearly reflects the effect of the developed potentials on the value of R_g. Also for the same N = 100 chain, changing R_{rep} from 0.1 to 1.0 nm only leads to a 1.0 nm change in R_g, while for a freely jointed chain it results in a 6.0 nm change in R_g. Therefore, it can be inferred that in the presence of local bending and torsion potentials, R_g only has a weak dependence on R_{rep}.

4. APPLICATION TO DENATURED PROTEINS

When the native state of a protein breaks down due to high temperature, pressure or the presence of a denaturing agent, it will turn to a dynamic set of complex conformations which is called the denatured state of a protein.[34] The addition of denaturants weakens the hydrophobic forces and disrupts the native hydrogen bonds in the protein,[35−37] and, as a result, only local interactions that restrict the polypeptide backbone to limited regions of conformation space are the dominant source of structure in the denatured state of the proteins.[38] Henceforth, we make use of the radius of gyration and Stokes radius for denatured proteins to benchmark the potential functions for bonded interactions developed in this study.

4.1. The Radius of Gyration R_g of Denatured Proteins.

Experimental studies have revealed that the ensemble average radius of gyration of denatured proteins follows a power law scaling of the type

$$R_g = R_0 N^{\nu} \tag{4}$$

where N is the number of residues, R_0 is a constant related to the persistence length of the polymer, and ν is a scaling exponent. More specifically, it has been found that for cross-link-free low-charge chemically unfolded proteins with sizes ranging from 16 to 549 residues, R_0 = 0.202 ± 0.041 nm and ν = 0.588 ± 0.037.[39]

We have performed MD simulations using the exact amino acid sequence of 20 denatured proteins as extracted from the protein data bank and using the developed bending and torsion potentials of our 3-letter amino acid model (a list of the simulated proteins is provided as Supporting Information, see Table S3). Repulsive interactions between non-neighboring beads are included to take into account the excluded volume effect using R_{rep} = 0.38 nm.

The predicted values of R_g for 20 denatured proteins are shown in Figure 6. For proper sampling, each protein sequence has been simulated for 10^7 MD steps, and each data point

dx.doi.org/10.1021/ct300684j | J. Chem. Theory Comput. 2013, 9, 432−440

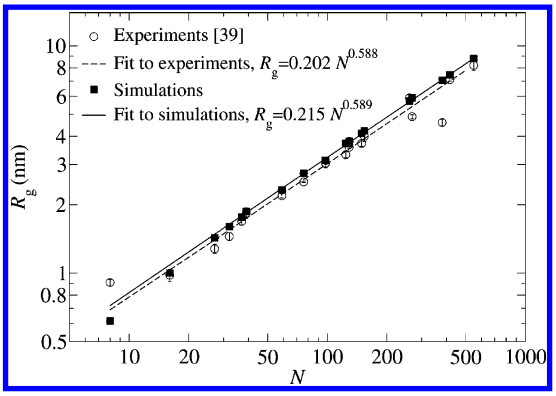


Figure 6. Comparison between the simulated and experimentally obtained R_g for 20 different noncross-linked chemically denatured proteins.[39] Experimental error bars are taken from Kohn et al.,[39] and the numerical error bars represent the standard deviation of 8 independent simulations. The studied proteins are listed in Table S3 of the Supporting Information.

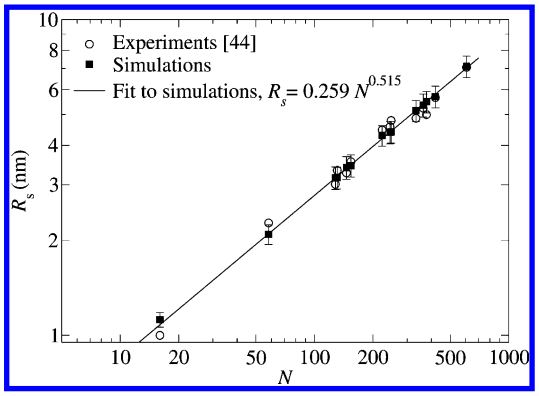


Figure 7. Comparison of the simulated and experimental Stokes radius for 14 chemically denatured proteins.[44] The error bars indicate the standard deviation. The studied proteins are listed in Table S4 of the Supporting Information.

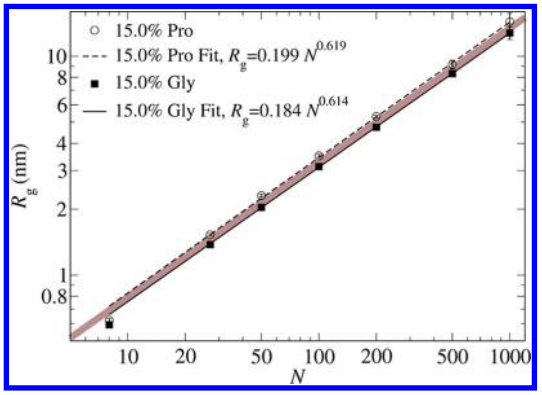


Figure 8. Simulation results for chains rich in proline and glycine residues.

corresponds to the average of 8 simulations obtained for different initial velocity distributions. The length of the chosen proteins spans 2 orders of magnitude. The simulation results show that R_g for chains longer than $N \sim 30$ beads follows a power-law scaling, but the shorter chains do not. This suggests that for short chains, the local stiffness of the chain is the dominant factor in determining the structure rather than the entropic motion. For longer chains, it can be seen that the R_g of denatured proteins is well predicted by the developed potentials within 10% error (except two outliers). The obtained simulation results can be fitted to the power law in eq 4 with $\nu = 0.589 \pm 0.009$ and $R_0 = 0.215 \pm 0.011$ nm (95% confidence bounds), which slightly overestimates the experimental values. This small deviation could be explained by the formation of hydrophobic clusters or the presence of residual secondary structures which can affect the mean persistence length of the polymer under experimental conditions.[39−41]

The generated conformations can also be used to study the hydrodynamic properties of denatured proteins. In the HYDRO computer program[42,43] the protein is represented by spherical frictional elements centered at the α-carbons of the polypeptide chain, and the Stokes radius is calculated based on the hydrodynamic theory of interacting beads. In the hydrodynamic calculation the bead radius of each frictional element is chosen to be 0.51 nm, and the Stokes radius calculations with the suggested value give reasonable results for folded proteins with rigid structures.[44] The simulations are performed on each protein sequence for 2.5×10^7 MD steps. The Stokes radius is calculated every 5000 steps, and the average value is reported. Fourteen denatured proteins (see Table S4) with lengths ranging from $N = 16$ to $N = 605$ amino acids are studied by the developed model, and the obtained Stokes radii are fitted to the power law $R_s = 0.259 \pm 0.019N^{0.515\pm0.013}$ nm (95% confidence bounds) which is in good agreement with the reported experimental data $R_s = 0.252N^{0.522}$ nm[44] (see Figure 7).

The bending and torsion potentials in our model are uncoupled, so that the correlation between these degrees of freedom is lost.[45,46] This can affect the local conformations of the simulated proteins as well as the ensemble average properties such as gyration radius. To investigate the local conformations, we have collected the pseudo-bending and torsion angles α and θ at regular intervals during the simulations of the proteins studied in Figure 6. The results are plotted in Figure S8 (a). To check the validity of the obtained density maps, a direct mapping from the Ramachandran plot of the coil regions of proteins (Figure S8 (b)) to the (α,θ) space is constructed (Figure S8 (c)). Comparison between the obtained (α,θ) density plots (i.e., comparing Figures S8 (a) and (c)) shows that the proteins can sample conformations in (α,θ) space that are not realistic due to the absence of bending-torsion coupling in the potentials. To investigate this effect on the predicted results for the gyration radius, we have used an approximate method to account for the coupling effect by introducing a "1−4" distance potential between beads i and i+3. The procedure to obtain this potential is explained in the section "Bending-torsion coupling" in the Supporting Information. We have repeated the simulations for the denatured proteins of Figure 6 with this additional potential, and it can be seen from Figure S8 (d) that including the bending-torsion coupling effect, even in this very primitive form, pushes the system to sample more realistic conformations. It can also be observed that the most populated area in Figure S8 (d) corresponds to the poly-proline II region which is experimentally confirmed to be the dominant backbone conformation in denatured proteins.[47,48] Next, the influence of the bending-torsion coupling on the average gyration radius is studied by comparing the results for R_g with and without the 1−4 distance potentials in Figure S9. This clearly shows that using the 1−4 distance potential only results in a minor change in the reported R_g values (maximum error of 3.1%, see Figure S9).

dx.doi.org/10.1021/ct300684j | J. Chem. Theory Comput. 2013, 9, 432−440

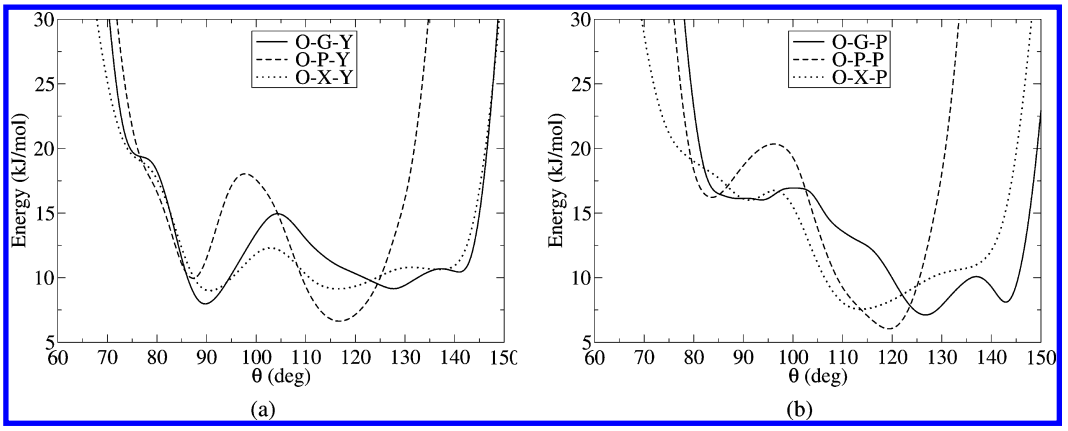


Figure 9. Level-2 bending potentials for triple-combinations with G, P, and X as central residues that (a) does not have a proline following them and (b) those which have a proline following them. O represents any type of amino acid, while Y represents any type of amino acid except P.

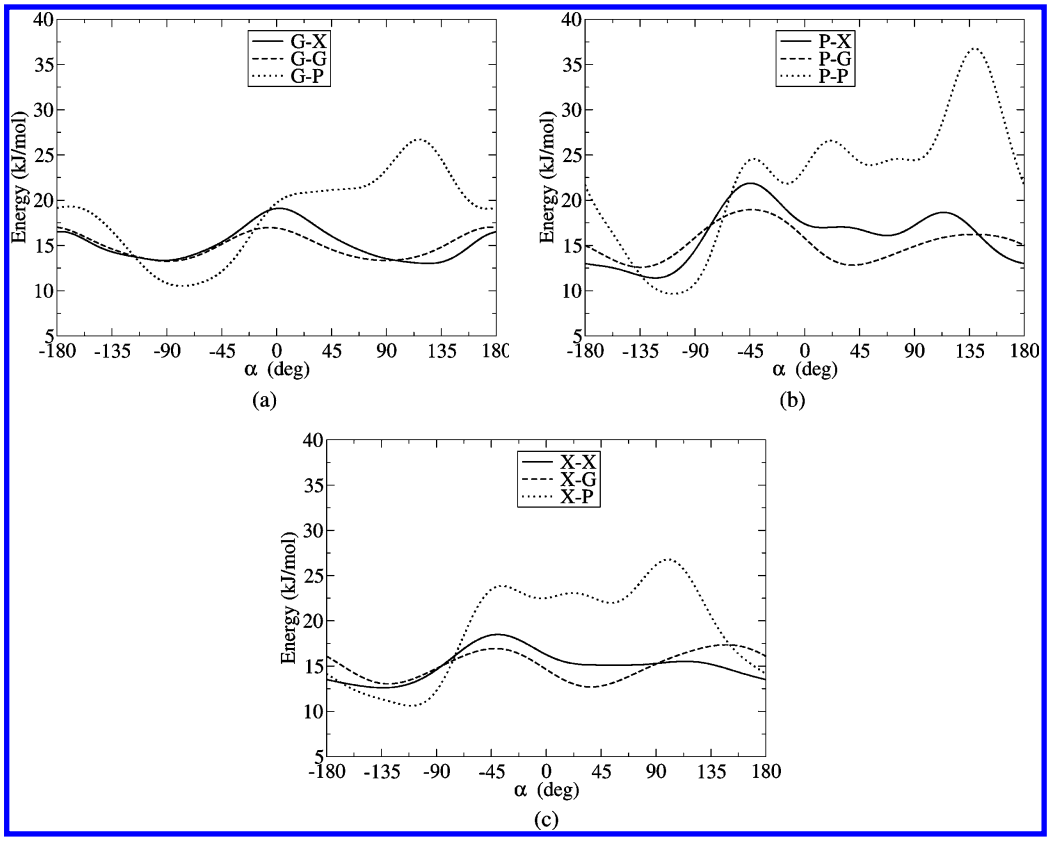


Figure 10. Torsion potentials for (a) G-X, G-G, and G-P combinations; (b) P-X, P-G, and P-P combinations; and (c) X-X, X-G, and X-P combinations.

This implies that the correlation between bending and torsion potentials does not have a considerable effect on the ensemble average R_g of the denatured proteins studied in this work, despite the fact that additional regions are sampled in (α,θ) space.

4.2. Sequence and Composition Effect. The polymer chains studied in section 3 clearly show the effect of amino acid composition of the chain on R_g, highlighting the role of Pro in enlarging and that of Gly in reducing the conformational radius. A survey on the protein sequences studied in the previous section (Figure 6) and the proteins in the DASSD library shows that the amount of Pro and Gly residues never exceeds 15%. In order to study the effect of sequence and composition on the R_g of unfolded proteins, a series of simulations has been performed on protein chains with different lengths, containing either 15% of Gly or 15% of Pro residues randomly distributed along the chain (the rest of the beads are of type X and the same simulation parameters as in section 3 and subsection 4.1 are used). As expected, the sequences rich in Pro residues adopt a higher R_g compared to the chains rich in Gly, producing an upper and lower bound for the R_g of denatured proteins (see Figure 8). Indeed, the simulation results for the specific denatured proteins of Figure 6 fall in between the bounds of Figure 8.

5. CONCLUSION

We have proposed an implicit solvent one-bead-per-amino-acid coarse-grained model to study the unfolded state of proteins.

dx.doi.org/10.1021/ct300684j | J. Chem. Theory Comput. 2013, 9, 432−440

Bending and torsion potentials for the bonded interactions are extracted from Ramachandran data of the coil regions of proteins in the protein data bank and hence the obtained potentials are not biased to any specific secondary structure. The accuracy of the potential functions is a comprise that maintains residue and sequence specificity and at the same time allows for data reduction. For the bending potentials, the neighbor residue effect does not play an important role except in the case of preproline residues. On the other hand, accounting for the neighboring residue effect in the extracted Ramachandran density plots improved the torsion potentials to a great extent. The validity of the model was demonstrated by comparing the ensemble average R_g of a series of denatured proteins with experimental values. It was shown that the developed potential functions for bonded interactions can successfully predict the scaling relations of denatured proteins. Future work will be focused on extending the current model with electrostatic and hydrophobic interactions in order to study natively unfolded proteins under physiological conditions.

6. APPENDIX

Detailed Potentials

The bending and torsion potentials resulting from the methodology in section 3 are presented in Figures 9 and 10.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

A list of the 3-residue-fragments and the associated number of Ramachandran data points (Table S1), bending potentials of level-1 and level-3 (Figures S1 to S4), the fitting equation for the torsion potentials (eq S1), the fitting coefficients for the torsion potentials of Level-1 (Table S2), a convergence study (Figure S5), the time step size selection (Figure S6), the list of simulated proteins (Tables S3 and S4), the method to obtain the 1−4 distance potential (section 'Bending-Torsion Coupling'), details regarding the (α,θ) density maps (Figure S8), and the gyration radius obtained with and without using the 1−4 distance potential (Figure S9). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: P.R.Onck@rug.nl.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Fink, A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35−41.
(2) Tompa, P. *Structure and function of intrinsically disordered proteins*; Chapman & Hall/CRC: 2009; pp 163−188.
(3) Dunker, A.; Brown, C.; Lawson, J.; Iakoucheva, L.; Obradović, Z. *Biochemistry* **2002**, *41*, 6573−6582.
(4) Radivojac, P.; Iakoucheva, L.; Oldfield, C.; Obradovic, Z.; Uversky, V.; Dunker, A. *Biophys. J.* **2007**, *92*, 1439−1456.
(5) Tompa, P. *Bioessays* **2003**, *25*, 847−855.
(6) Tozzini, V. *Acc. Chem. Res.* **2010**, *43*, 220−230.
(7) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.
(8) Tozzini, V. *Q. Rev. Biophys.* **2010**, *43*, 333−371.
(9) Feig, M. *Modeling Solvent Environments: Applications to Simulations of Biomolecules*; Wiley-VCH: 2010.
(10) Sorenson, J.; Head-Gordon, T. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 368−379.
(11) Korkut, A.; Hendrickson, W. *Proc. Natl. Acad. Sci.* **2009**, *106*, 15667.
(12) Tozzini, V.; McCammon, J. *Chem. Phys. Lett.* **2005**, *413*, 123−128.
(13) Basdevant, N.; Borgis, D.; Ha-Duong, T. *J. Phys. Chem. B* **2007**, *111*, 9390−9399.
(14) Smith, P.; Pettitt, B. *J. Phys. Chem.* **1994**, *98*, 9700−9711.
(15) Tirion, M. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(16) Ercolessi, F.; Adams, J. *Europhys. Lett.* **1994**, *26*, 583.
(17) Monticelli, L.; Kandasamy, S.; Periole, X.; Larson, R.; Tieleman, D.; Marrink, S. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(18) Yap, E.; Fawzi, N.; Head-Gordon, T. *Proteins: Struct., Funct., Bioinf.* **2007**, *70*, 626−638.
(19) Tozzini, V.; Rocchia, W.; McCammon, J. *J. Chem. Theory Comput.* **2006**, *2*, 667−673.
(20) Tozzini, V.; Trylska, J.; Chang, C.; McCammon, J. *J. Struct. Biol.* **2007**, *157*, 606−615.
(21) Bereau, T.; Deserno, M. *J. Chem. Phys.* **2009**, *130*, 235106.
(22) Finkelstein, A.; Ptitiÿăsiÿąyn, O. *Protein physics: a course of lectures*; Academic Press: 2002; pp 15−22.
(23) Creighton, T. E. *Proteins: structures and molecular properties*; WH Freeman: 1993.
(24) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59−107.
(25) Flory, P. *Macromolecules* **1974**, *7*, 381−392.
(26) Avbelj, F.; Grdadolnik, S.; Grdadolnik, J.; Baldwin, R. L. *Proc. Natl. Acad. Sci.* **2006**, *103*, 1272.
(27) Dayalan, S.; Gooneratne, N.; Bevinakoppa, S.; Schroder, H. *Bioinformation* **2006**, *1*, 78.
(28) Ho, B.; Brasseur, R. *BMC Struct. Biol.* **2005**, *5*, 14.
(29) Shimazaki, H.; Shinomoto, S. *Neural Comput.* **2007**, *19*, 1503−1527.
(30) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(31) Ashbaugh, H.; Hatch, H. *J. Am. Chem. Soc.* **2008**, *130*, 9536−9542.
(32) Soranno, A.; Longhi, R.; Bellini, T.; Buscaglia, M. *Biophys. J.* **2009**, *96*, 1515−1528.
(33) Rauscher, S.; Baud, S.; Miao, M.; Keeley, F.; Pomès, R. *Structure* **2006**, *14*, 1667−1676.
(34) Rose, G. *Advances in protein chemistry*; Academic Press: 2002; Vol. *62*, pp 1−22.
(35) Zangi, R.; Zhou, R.; Berne, B. *J. Chem. Theory Comput.* **2009**, *131*, 1535−1541.
(36) Lim, W.; Rösgen, J.; Englander, S. *Proc. Natl. Acad. Sci.* **2009**, *106*, 2595.
(37) Das, A.; Mukhopadhyay, C. *J. Phys. Chem. B* **2008**, *112*, 7903−7908.
(38) Creamer, T. *Unfolded proteins: from denatured to intrinsically disordered*; Nova Publishers: 2008; pp 1−21.
(39) Kohn, J.; Millett, I.; Jacob, J.; Zagrovic, B.; Dillon, T.; Cingel, N.; Dothager, R.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. *Proc. Natl. Acad. Sci.* **2004**, *101*, 12491.
(40) Baldwin, R.; Zimm, B. *Proc. Natl. Acad. Sci.* **2000**, *97*, 12391.
(41) Fitzkee, N.; Rose, G. *Proc. Natl. Acad. Sci.* **2004**, *101*, 12497.
(42) Garcia de la Torre, J.; Navarro, S.; Lopez Martinez, M.; Diaz, F.; Lopez Cascales, J. *Biophys. J.* **1994**, *67*, 530−531.
(43) Carrasco, B.; Garcia de la Torre, J. *Biophys. J.* **1999**, *76*, 3044−3057.
(44) Zhou, H. *J. Phys. Chem. B* **2002**, *106*, 5769−5775.
(45) Alemani, D.; Collu, F.; Cascella, M.; Dal Peraro, M. *J. Chem. Theory Comput.* **2009**, *6*, 315−324.
(46) Trovato, F.; Tozzini, V. *AIP Conf. Proc.* 2012, *1456*, 187.
(47) Shi, Z.; Woody, R.; Kallenbach, N. *Adv. Protein Chem.* **2002**, *62*, 163−240.
(48) Whittington, S.; Chellgren, B.; Hermann, V.; Creamer, T. *Biochemistry* **2005**, *44*, 6269−6275.