Knowledge-Based Potential for the Polypeptide Backbone

Marcos R. Betancourt*

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

Received: August 28, 2007; In Final Form: January 14, 2008

A knowledge-based potential for the polypeptide backbone as a function of the dihedral angles is developed and tested. The potential includes correlations due to the conformations and compositions of adjacent residues. Its purpose is to serve as a major component of a coarse-grained protein potential by including the most relevant local interactions while averaging out nonbonded ones. A probability density estimation algorithm and a multi-resolution probability combination procedure are developed to produce smooth probability distributions and dihedral angle potentials. The potential is described by a set of two-dimensional dihedral angle surfaces involving the various combinations of amino acid triplets and duplets. Several tests are carried out to evaluate the quality of the potential. Monte Carlo simulations, using only the dihedral angle potential and a coarse-grained excluded volume potential, show that the resulting dihedral angle distributions and correlations are consistent with those extracted from protein structures. Additional simulations of unfolded proteins are carried out to measure the NMR residual dipolar coupling (RDC). Significant correlations are obtained between the simulations and the corresponding experiments consistent with other simulations in the literature. Finally, the dihedral angle entropies are calculated for the 20 amino acids. In particular, the entropy difference between alanine and glycine agrees with the ones computed from molecular dynamics simulations (≈0.4 kcal/mol).

1. Introduction

The long time-scale simulations of proteins by molecular dynamics remain a formidable problem. Because of the large number of atoms in proteins and the large number of possible conformations, extensive simulations are required for the determination of the protein native structure and the calculation of its thermodynamic properties. Moreover, present molecular mechanic force fields are not sufficiently accurate to perform many of these calculations. Low-resolution protein models (or coarse-grained models) are a practical alternative to molecular mechanic models. They simplify the problem by reducing the number of degrees of freedom at the expense of eliminating fine structural details and fast scale motions, which are assumed to be mostly irrelevant in the long time-scale processes of large proteins. The main problem with this approach is in the determination of the potential energy function. A reduction of the model degrees of freedom is typically achieved by clustering atoms into larger coarse-grained “atoms”. Effective potentials are often derived from the properties of experimentally determined protein structures (knowledge based potentials or KBPs) or from empirical theories. Nevertheless, the quality of these potentials remains low.

A common statistical technique for obtaining the KBPs is based on the assumption that the coarse atom interactions satisfy the Boltzmann distribution and that they are statistically independent from each other (the quasi-chemical approximation). Solvent effects are implicitly included in these statistical energies (i.e., the implicit solvent model). More precisely, if **x** is a set of generalized coordinates describing the coarse atoms conformations and $P(\mathbf{x})$ is the observed frequency (i.e., probability) in which the coarse atoms are found with a particular **x** value, the statistical potential (or potential of mean force) can be approximated by

$$U(\mathbf{x}) = -k_B T_0 \ln\left[\frac{P(\mathbf{x})}{\hat{P}(\mathbf{x})}\right] + \text{constant} \tag{1}$$

where T_0 is the presumed temperature of the ensemble and $\hat{P}(\mathbf{x})$ is the (zero energy) reference state probability.

In this work, the objective is to derive an all-inclusive and precise backbone potential that captures the local (in sequence) protein interactions. Local interactions are one of the two main contributions of coarse-grained potentials (nonlocal being the other one, such as pairwise interactions among coarse atoms). A local potential can be described by the dihedral angle preferences of a specific residue. In addition, it has been demonstrated in molecular dynamic[1] and Langevin dynamic[2] simulations of polyalanine chains that these preferences also depend on the composition and conformations of neighboring residues. The present potential is based on a previous dihedral angle potential that include correlations between the dihedral angles and composition for all of the amino acid combinations of residue triplets.[3] The current model improves the weight and accuracy of the correlations between the dihedral angles from adjacent residues, minimizes the contribution form nonlocal interactions, and maximizes the information extracted from protein structures.

One assumption in deriving dihedral angle KBPs is that the nonbonded interaction effects are averaged out. This is generally a good assumption except for the case of backbone hydrogen bonds in regular secondary structures, which are related to specific dihedral angle values. A strategy that can be used for reducing these effects is the one introduced by Jha and co-workers[4] and by Bernardo et al.[5] to represent the unfolded

* Corresponding author. Phone: (317) 274-6910. Fax: (317) 274-2393. E-mail: mbetancourt@mailaps.org.

10.1021/jp076906+ CCC: $40.75
Published on Web 03/29/2008

state. It consists of using only dihedral angle information from residues involved in coils (and not in or near helices, sheets, or turns). This approach reduces the hydrogen bond biases on the dihedral angles, but it is not clear if it adds new ones by undersampling residues in sequences with a natural tendency to adopt dihedral angles typical of regular secondary structures. That is, it is not clear to what extent is a hydrogen bond inducing the angle or the angle facilitating the hydrogen bond. Here, the dihedral angle potential is derived from residues that do not form hydrogen bonds within the backbone, although simulations using the potential derived from unrestricted residues are also carried out for comparison. The potential consists of a combination of histograms of dihedral angle pairs from the same or from adjacent residues. A number of data smoothing and averaging techniques are applied to improve the probability histograms and reduce the noise. Potentials with different levels of approximations are combined in an approach similar to Sippl sparse data approximation.[6]

As a potential validation test, simulations of the protein denaturated state are carried out to estimate the nuclear magnetic resonance (NMR) residual dipolar couplings (RDC)[7] and to compare them to known experimental values of several proteins. RDCs are used to measure the average persistent orientations between magnetic dipoles in molecules. Similar tests were carried out by Jha et al.[4] and by Bernardo et al.[5] However, a difference from the present approach and these other methods is that they build the structures partially by selecting them from a library of structures. For example, in the work of Jha et al., a potential is derived in terms of a very coarse dihedral angle space (five possible ϕ, ψ combinations), which is then used to select a structure from a coil library, followed by an excluded volume potential minimization involving atomic (hard sphere) level contributions as well as soft-sphere interactions between side chains (the latter added by a third party algorithm). In the algorithm of Bernardo et al., the chain is grown from one end to another by selecting amino acid specific ϕ, ψ combinations from a loop database, which are rejected and another selected in the case of steric clashes. Side chains are also added by a different third party program. Unlike the present approach, a disadvantage of these methods is that they are not suitable for performing protein dynamics and cannot be simply integrated into a molecular simulation program (coarse-grained or not) based on potential energy functions. Another problem evident with the Bernardo et al. method is that the structures are biased by only accepting structures that avoid steric clashes. To avoid these biases, the chain growth algorithm[8] (or one of its variants) would have to be used.

Additional potential tests are carried out. The dihedral angle probability distribution is calculated from denaturated ensemble simulations and compared to the ones extracted from protein structures. In addition, the backbone dihedral angle entropies for all amino acids are computed and compared to those obtained by detailed atomic simulations. The following section describes the derivation of the dihedral angle potential from a diverse set of protein structures.

2. Methods

Dihedral Angle Potential Definition. Local interactions involve the geometric and energetic effects between the bonded atoms of adjacent residues. The backbone geometry of the current model only considers the ϕ, ψ dihedral angles as coordinates, with the exception of proline in which the ω angle is also included. For every consecutive triplet of residues with amino acid composition *abc*, a potential energy function that captures the effects of the conformation (ignoring ω for proline) and composition of neighboring residues can be generally written as

$$U_{abc}(\phi_a, \psi_a; \phi_b, \psi_b; \phi_c, \psi_c) \tag{2}$$

Note that in KBPs, U_{abc} is actually a free energy in units of k_BT_0. There are 8000 of these functions resulting from the combination of the 20 amino acids. Each represents a surface in a 9 dimensional space, and in practice, there are not enough data to generate them from known protein structures. Instead, only correlations between dihedral angle pairs are considered here. Following our previous work,[3] it is assumed that the dihedral angle potential for a triplet of consecutive residues can be approximated by

$$\begin{aligned} U_{abc}(\phi_a, \psi_a; \phi_b, \psi_b; \phi_\mathrm{c}, \psi_c) = U_{0;abc}(\phi_b, \psi_b) + \\ \frac{1}{2}[\Delta U_{1;ab}(\psi_a, \phi_b) + \Delta U_{2;ab}(\psi_a, \psi_b) + \Delta U_{3;ab}(\phi_a, \phi_b)] + \\ \frac{1}{2}[\Delta U_{1;bc}(\psi_b, \phi_c) + \Delta U_{2;bc}(\psi_b, \psi_c) + \Delta U_{3;bc}(\phi_b, \phi_c)] \end{aligned} \tag{3}$$

There are a few differences from eq 3 and the previous model. The inter-residue cross terms, $\Delta U_{1;ab}(\psi_a, \phi_b)$ for example, are now given by

$$\Delta U_{1;ab}(\psi_a, \phi_b) = U_{1;ab}(\psi_a, \phi_b) - [U_{1;ab}(\psi_a) + U_{1;ab}(\phi_b)] \tag{4}$$

where the last two terms are obtained by integrating $U_{1;ab}(\psi_a, \phi_b)$ over one of the dihedral angles. These cross terms account for the dihedral angle correlation energy between two consecutive residues and vanish if ψ_a and ϕ_b are uncorrelated. Note that even if these terms vanish, the central $U_{0;abc}$ potential can still capture the effect of the surrounding residue types *a* and *c* on the central dihedral angles. Another difference in eq 3 is that, unlike the central term, the cross terms are now a function of two adjacent residues only. The 1/2 factors correct for the double appearance between two consecutive residues when the energy for the entire chain is considered. An additional term $U_P(\omega)$ is added when the middle residue of the triplet is proline. The potentials at the chain ends are similarly obtained from

$$\begin{aligned} U_{*bc}(\phi_b, \psi_b; \phi_c, \psi_c) = U_{0;*bc}(\phi_b, \psi_b) + \\ \frac{1}{2}[\Delta U_{1;bc}(\psi_b, \phi_c) + \Delta U_{2;bc}(\psi_b, \psi_c) + \Delta U_{3;bc}(\phi_b, \phi_c)] \end{aligned} \tag{5}$$

$$\begin{aligned} U_{ab*}(\phi_a, \psi_a; \phi_b, \psi_b) = U_{0;ab*}(\phi_b, \psi_b) + \\ \frac{1}{2}[\Delta U_{1;ab}(\psi_a, \phi_b) + \Delta U_{2;ab}(\psi_a,\psi_b) + \Delta U_{3;ab}(\phi_a, \phi_b)] \end{aligned} \tag{6}$$

where the symbol * represents the contribution from all residues equally weighted. Ideally, a special potential should be generated for the chain ends because of the differences in geometry and chemical composition, but these effects are ignored here.

The potential functions given by eqs 3−6 are obtained using eq 1. This requires the calculation of the corresponding probability densities as described in the following section.

Probability Density and Potential Energy Calculation. In preparation for calculating the dihedral angle potential, a nonredundant protein database is built from high-quality protein structures obtained by crystallographic techniques and deposited in the Protein Data Bank (PDB). In particular, the structures are limited to 3 Å of resolution or better. To minimize information redundancy and structural biases, the triplets are selected from proteins with 20% of sequence identity or less. However, to maximize the number of triplets in the database

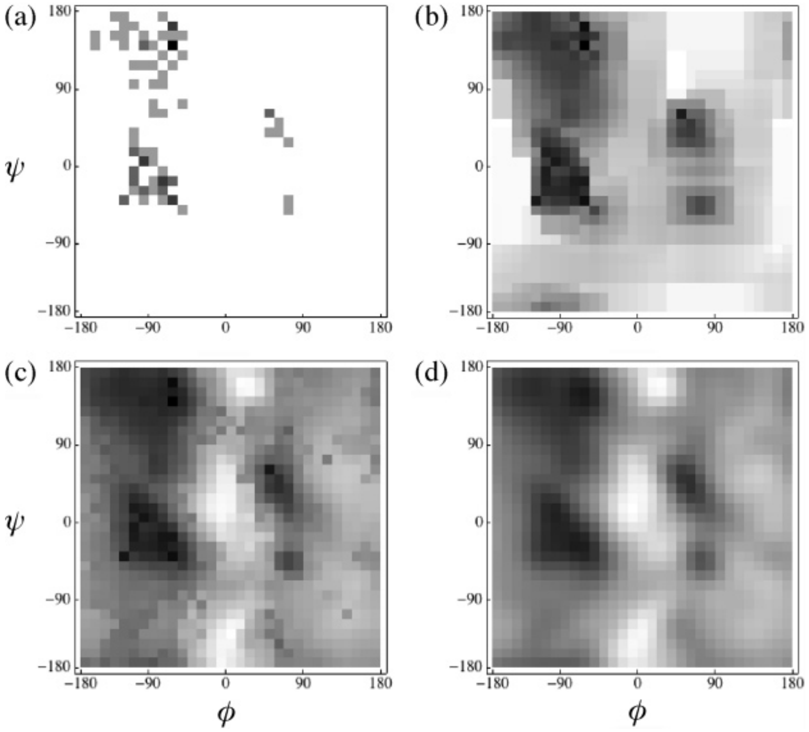

Figure 1. Example of probability density estimation steps for the Leu-Lys-Thr triplet, $P_{LKT}(\phi, \psi)$. Dark regions correspond to high probabilities, shown in a logarithmic scale. (a) Initial "raw" data points ($\ln(1 + n)$). (b) Probability after the application of the density estimation algorithm. (c) Result after combination with lower residue-specific probabilities. (d) Final probability after Fourier-transform smoothing.

the following procedure is used. First, a list of all of the proteins from the PDB with less 98% identity is compiled (call it the 98% list), which resulted in 22 626 sequences. This step eliminates identical or nearly identical sequences from the list. Next, the proteins are clustered in groups with less than or equal to 20% sequence identity. The dihedral angle events for a given triplet are obtained from all of the similar triplets in the 98% list that exclusively belong to structures with less than 20% identity. In this way, instead of arbitrarily selecting a small group of proteins with less than 20% sequence identity, all of the proteins from the 98% list are available for the analysis. Some other details of this procedure can be found in the previous work.[3]

The algorithm for the calculation of the probabilities and potential can be summarized as follows:

(1) Build triplet histograms from the observed dihedral angles.

(2) Compute the "raw" probability densities for all three-residue specific combinations as well as for two-, one-, and a no-residue specific cases.

(3) Improve the probability densities by applying a density estimation algorithm that reduces bin size effects.

(4) Compensate for sparse data by combining the three-, two-, one-, and no-residue specific probabilities.

(5) Transform the probabilities into energies and reduce noise fluctuations.

1. Build Triplet Histograms from the Observed Dihedral Angles. For each triplet, event histograms corresponding to each of the energy terms in eq 3 are computed by calculating the dihedral angles of all triplets in the database satisfying the required sequence identity conditions. The histograms are divided in two: one where the center residue is not in regular secondary structures (coiled residues) and another where no such distinction is made (unrestricted residues). Coiled residues are determined by classifying the residues into secondary structures according to the DSSP algorithm.[9] Center residues classified as (H,G,I,E,T,B) by DSSP are not used in the histograms of coiled residues. The histograms consist of square lattices, with each angle range divided into $B = 32$ equal parts, for a resolution of 11.25° × 11.25°. This choice provides enough details to adequately encode the conformation of a residue backbone. On average, the database provides 46 events per each histogram of coiled-residue triplets and 163 for the histograms of unrestricted residue triplets. Not all coiled-residue triplets are present. Thirty coiled-residue triplet histograms contain 0 events, and 2622 of them contain 20 events or less.

2. Compute the "Raw" Probability Densities for All Three-Residue Specific Combinations as well as for Two-, One-, and a No-Residue Specific Cases. In addition to the three-residue specific histograms, two-, one-, and no-residue specific histograms are calculated by averaging over residue types. These additional histograms are used to approximate and complement sparse three-residue histograms. They can be computed as follows. If $N_{abc}(\phi, \psi)$ is the number of observed events, where it is implied that the angles correspond to the center residue b whenever a triplet is specified, then the total number of events for the particular triplet is

$$C_{abc} = \sum_{\phi,\psi} N_{abc}(\phi, \psi) \qquad (7)$$

and the probability can be initially approximated as

$$P''_{abc}(\phi, \psi) = \frac{N_{abc}(\phi, \psi)}{C_{abc}} \qquad (8)$$

where the double prime indicate the first approximation to be made. As an example, the distribution for the Lue-Lys-Thr (coiled-residue) triplet is shown in Figure 1a. The two-, one-, and no-residue specific probabilities are calculated by adding all histograms of the particular residue type (or types) to be averaged (except when b or c are proline). However, one must remove biases caused by the uneven representation of the residues being averaged, while reducing the inaccuracies due to residues with low representations. The two-, one-, and no-residue specific probabilities are computed using

$$P''_{ab*}(\phi, \psi) = \left[\sum_{c} \frac{C_{abc}}{K + C_{abc}} P''_{abc}(\phi, \psi)\right] / Z''_{ab*} \tag{9}$$

$$P''_{*bc}(\phi, \psi) = \left[\sum_{a} \frac{C_{abc}}{K + C_{abc}} P''_{abc}(\phi, \psi)\right] / Z''_{*bc} \tag{10}$$

$$P''_{*b*}(\phi, \psi) = \left[\sum_{a,c} \frac{C_{abc}}{K + C_{abc}} P''_{abc}(\phi, \psi)\right] / Z''_{*b*} \tag{11}$$

$$P''_{***}(\phi, \psi) = \left[\sum_{a,b,c} \frac{C_{abc}}{K + C_{abc}} P''_{abc}(\phi, \psi)\right] / Z''_{***} \tag{12}$$

where the Z'''s are normalization factors. Adding the probabilities instead of the histograms reduces biases while the ratio $C_{abc}/(K + C_{abc})$ is introduced to decrease the weight of poorly sampled triplets. K is a constant (chosen as $K = 30$) that approximately represents the value of C_{abc} around and above which the probabilities are more robust. Probabilities that are less residue specific are used to complement more specific ones. In addition, the two-residue specific probabilities, $P''_{ab*}(\phi,\psi)$ and $P''_{*bc}(\phi,\psi)$, are also used to describe the chain ends, where a third residue does not exist. A separate probability histogram is computed for the ω angle of proline, independent of other residues. For simplicity and to avoid large sparse data errors, the correlations between the ω and the ϕ angles in proline are ignored, even though these may be significant.

The cross-term probabilities $P''_{ab}(\psi,\phi)$, $P''_{a*}(\psi,\phi)$, $P''_{*b}(\psi,\phi)$, and $P''_{**}(\psi,\phi)$ are calculated in a similar fashion, starting from the number of observed events $N_{ab}(\psi, \phi)$. In this case, whenever a residue duplet is considered, it is implied that the first angle corresponds to the left residue and the second angle to the right one. The calculation of the other two cross terms ($P''_{ab}(\phi,\phi)$ and $P''_{ab}(\psi,\psi)$) is identical, and only the derivation of the $P''_{ab}(\psi,\phi)$ term will be described in what follows.

3. Improve the Probability Densities by Applying a Density Estimation Algorithm That Reduces Bin Size Effects. The histogram resolution chosen affects the values of the probability density and the extracted energy. For sparse data, small bin sizes can lead to a rough representation of the probability density while large bin sizes lead to a loss of information. Therefore, an improved version of the probability density estimation algorithm described in the previous work is implemented here.[3] This algorithm selects the most appropriate bin size (up to the minimum, which in this case corresponds to 11.25° for each dihedral angle) by recursively subdividing the two-dimensional lattice in equal area quadrants, until the density in each quadrant becomes uniform according to a statistical criteria (χ^2) of uniformity. This procedure can lead to some artificially flat probability regions, depending on the point of origin where the space is initially subdivided. Because the lattice has toroidal symmetry, the algorithm can be improved by calculating the density many times but in each case shifting the point of origin to every point (or every other point for computational efficiency) of the lattice before the recursive subdivision procedure is applied. For example, in the first density calculation, the lattice gets recursively divided centered around the midpoint {16, 16}, and in the next density calculation, it is divided centered around the point {16, 18}, and so on. All of the density histograms obtained in this manner are then averaged and a probability is assigned to each of the 32 × 32 bins. This algorithm results in smoother probabilities, $P'_{abc}(\phi,\psi)$ and $P'_{ab}(\psi,\phi)$, which are less dependent on bin size. Note that the probabilities are first multiplied by their corresponding C_{abc} or C_{ab} factors for the statistical criteria to be applicable. This procedure is applied to the three-, two-, one-, and no-residue specific probabilities, even though the largest impact occurs for the three-residue one. An example is shown in Figure 1b.

4. Compensate for Sparse Data by Combining the Three-, Two-, One-, and No-Residue Specific Probabilities. Because of the limited amount of data for each triplet, many of the dihedral angle bins contain no information. The probabilities given by eqs 9−12 (after applying the density estimation algorithm) can be used to approximate the empty bins. The method works as follows. Let $p'_1(\phi,\psi)$ be a probability (either for a residue triplet or duplet) that is to be complemented by a probability with lower residue specificity $p'_0(\phi,\psi)$. The new probability is obtained using

$$p_1(\phi, \psi) = \frac{p'_0(\phi, \psi)/w(\phi, \psi) + C_1 p'_1(\phi, \psi)}{Z_1} \tag{13}$$

where C_1 is the number of events involved in $p'_1(\phi,\psi)$, Z_1 is a normalization, and $w(\phi, \psi)$ is a weight factor to be described below. This is similar to the method proposed by Sippl,[6] except that the factor $w(\phi, \psi)$ here depends on the dihedral angles. If the operation in eq 13 is represented as

$$p_1(\phi, \psi) = p'_0(\phi, \psi) \oplus p'_1(\phi, \psi) \tag{14}$$

then the combination of the probabilities with several levels of residue specificity can be obtained from

$$P_{abc} = [[q \oplus P'_{***}] \oplus P'_{*b*}] \oplus P'_{abc} \tag{15}$$

$$P_{ab} = [q \oplus P'_{**}] \oplus P'_{ab} \tag{16}$$

where the factor $q = 1/B^2$ is the uniform probability and where the angles have been omitted for clarity. Notice that eqs 15 and 16 could have included terms of the form $(P'_{ab*} + P'_{*bc})/2$ and $(P'_{a*} + P'_{*b})/2$, respectively, but it was found that such a combination does not seem to significantly add more information. For the chain ends, the probabilities are obtained in a similar fashion, for example,

$$P_{ab*} = [[q \oplus P'_{***}] \oplus P'_{*b*}] \oplus P'_{ab*} \tag{17}$$

The proline residue requires special attention because of the stronger restrictions of the proline ϕ angle. In general, whenever the second and third residues in a triplet or the second residue in a duplet is proline, one cannot approximate it by the average of all residues because the ϕ angle distribution of the non-proline residues is significantly different. The ϕ angle of the first residue in a triplet or a duplet does not affects the probability (except for the ϕ_a, ϕ_b duplet cross terms). The following four special cases are considered as a replacement of eqs 15 and 16, whenever proline (p) is involved in the specified locations:

$$P_{apc} = [q \oplus P'_{*p*}] \oplus P'_{apc} \tag{18}$$

$$P_{abp} = [[q \oplus P'_{**p}] \oplus P'_{*bp}] \oplus P'_{abp} \tag{19}$$

$$P_{app} = [q \oplus P'_{*pp}] \oplus P'_{app} \tag{20}$$

$$P_{ap} = [q \oplus P'_{*p}] \oplus P'_{ap} \tag{21}$$

Note that P'_{**p} is a probability calculated specially for this situation.

Equation 13 can be obtained in an alternative way, which leads to the value of $w(\phi, \psi)$. Let

$$\sigma[N] = \frac{1}{\sqrt{1 + N}} \quad (22)$$

be an approximation to the relative error in a bin, which varies from zero to one depending on the number of events N in the bin. Given $N_1(\phi, \psi) = C_1 p'_1(\phi,\psi)$, a weight of $1 - \sigma^2[N_1(\phi, \psi)]$ is assigned to each corresponding bin. The objective is to merge $p'_1(\phi,\psi)$ and $p'_0(\phi,\psi)$ such that $p'_1(\phi,\psi)$ dominates when the number of events in the (ϕ, ψ) bin is significant, and such that $p'_0(\phi,\psi)$ dominates otherwise. This combination is achieved with

$$p_1(\phi, \psi) = \frac{\sum_{\phi,\psi} \sigma^2[N_1(\phi, \psi)]p'_1(\phi, \psi)}{\sum_{\phi,\psi} \sigma^2[N_1(\phi, \psi)]p'_0(\phi, \psi)} \sigma^2[N_1(\phi, \psi)]p'_0(\phi, \psi) + \{1 - \sigma^2[N_1(\phi, \psi)]\}p'_1(\phi, \psi) \quad (23)$$

The factors $\sigma^2[N_1(\phi, \psi)]$ and $\{1 - \sigma^2[N_1(\phi, \psi)]\}$ linearly interpolate between the two probabilities. The quantity $\sum_{\phi,\psi}\sigma^2[N_1(\phi,\psi)]p'_1(\phi,\psi)$ is the total probability of $p'_1(\phi,\psi)$ that has been reduced by the weight involving the errors. By multiplying the $p'_0(\phi,\psi)$ term by this factor, the $p'_0(\phi,\psi)$ term compensates for the $p'_1(\phi,\psi)$ reduction. Identifying eq 23 with eq 13, one can see that

$$w(\phi, \psi) = \left[\frac{\sum_{\phi,\psi} \sigma^2[N_1(\phi, \psi)]p'_0(\phi, \psi)}{\sum_{\phi,\psi} \sigma^2[N_1(\phi, \psi)]p'_1(\phi, \psi)}\right] p'_1(\phi, \psi) \quad (24)$$

An example of a combined probability is shown in Figure 1c.

5. Transform the Probabilities into Energies and Reduce Noise Fluctuations. With the probabilities obtained as described above, the energy terms in eq 3 are explicitly computed as

$$U_{0;abc}(\phi_b, \psi_b) = -\ln\left[\frac{P_{abc}(\phi_b, \psi_b)}{q}\right] \quad (25)$$

$$\Delta U_{1;ab}(\psi_a, \phi_b) = -\ln\left[\frac{P_{ab}(\psi_a, \phi_b) + 1/C_{ab}}{\sum_{\phi_b} P_{ab}(\psi_a, \phi_b) \sum_{\psi_a} P_{ab}(\psi_a, \phi_b) + 1/C_{ab}}\right] \quad (26)$$

in units of k_BT_0. The reciprocal of $C_{ab} = \sum_{\psi,\phi} N_{ab}(\psi, \phi)$ has been added to the cross potential term to suppress large energy fluctuations in regions of low probability and large errors. Similar equations are used for $\Delta U_{2;ab}(\psi_a, \psi_b)$ and $\Delta U_{3;ab}(\phi_a, \phi_b)$. Here, q defines the zero energy reference state.

A final smoothing of the energies is made to reduce high wavenumber (mainly noise) fluctuations. In applications, such as dynamic simulations where the potential gradient needs to be computed, it is convenient to have functions that are as smooth as possible. In addition, averaging and smoothing procedures in coarse-grained models tend to improve the potential energy functions. To reduce the noise fluctuation, a low-pass Fourier transform filter (removing fluctuations with wavelength of 8 or more in lattice units out of a maximum of 16) was applied to the potential. An example of the resulting energy for Leu-Lys-Thr is shown in Figure 1d. Notice that this potential also models the energy barriers between minima. A three-dimensional plot of the entire energy surface for the same triplet is shown in Figure 2. In this example, the energy range extends from near $-3k_BT_0$ to about $8k_BT_0$.

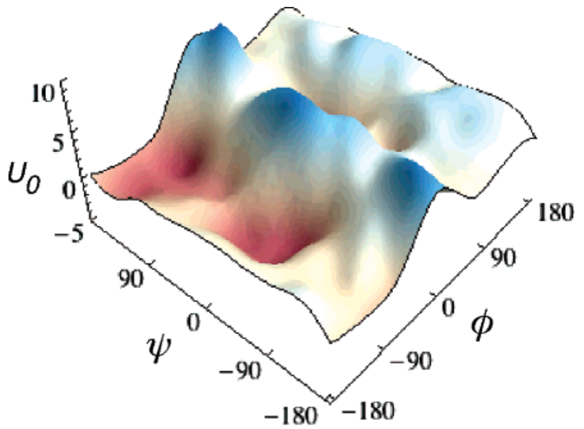

Figure 2. Three-dimensional view of the energies obtained for the Leu-Lys-Thr triplet, $U_{0;LKT}(\phi, \psi)$. The plot covers an energy range of approximately 10 k_BT_0. Note that the plotting program interpolates between bins creating an even smoother appearance.

3. Results and Discussion

Three different tests were carried out to evaluate the dihedral angle potential. First, consistency tests were carried out to verify the correspondence between the observed protein dihedral angle distributions, the probability derived from the potential, and the probabilities resulting from simulations. Second, experimental measurements of RDCs for denatured proteins were compared to those obtained by Monte Carlo simulations with the potential. Third, the entropy loss upon substitution of alanine by glycine obtained from the potential is compared to those obtained from molecular dynamic simulations. This entropy change is also computed for the other amino acids.

Monte Carlo simulations were carried out using a coarse-grained model that combines the dihedral angle potential with nonlocal excluded volume interactions. The model keeps track of the atomic positions to compute the backbone dihedral angles and the position of the coarse-grained atoms, formed by uniting selected groups of atoms. This hybrid (atomic plus coarse-atomic) model is similar to the one used in a previous work, differing mainly in the selection of the atoms composing the coarse-grained atoms.[3] One, two, or three coarse-atoms describe the various residues and are used to compute the nonbonded interactions (i.e., excluded volume effects). The excluded volume interactions are modeled by a soft repulsive potential of the form $(1.5\text{Å}/r)^{12}$ between all coarse-grained atom pairs from residues separated by more than 2 peptide bonds, where r is their pair wise distance. Some other details of this Monte Carlo algorithm are described elsewhere.[10]

Consistency Tests. The probability obtained in step two of the potential energy calculation can be considered the observed probability density for the dihedral angle propensities. The process of combining the observed probability densities (step three) and smoothing the resulting potential (step four) modify the probabilities into the potential energy model. To verify that these steps produce potential energy functions consistent with the observed probability densities, a root-mean-square error

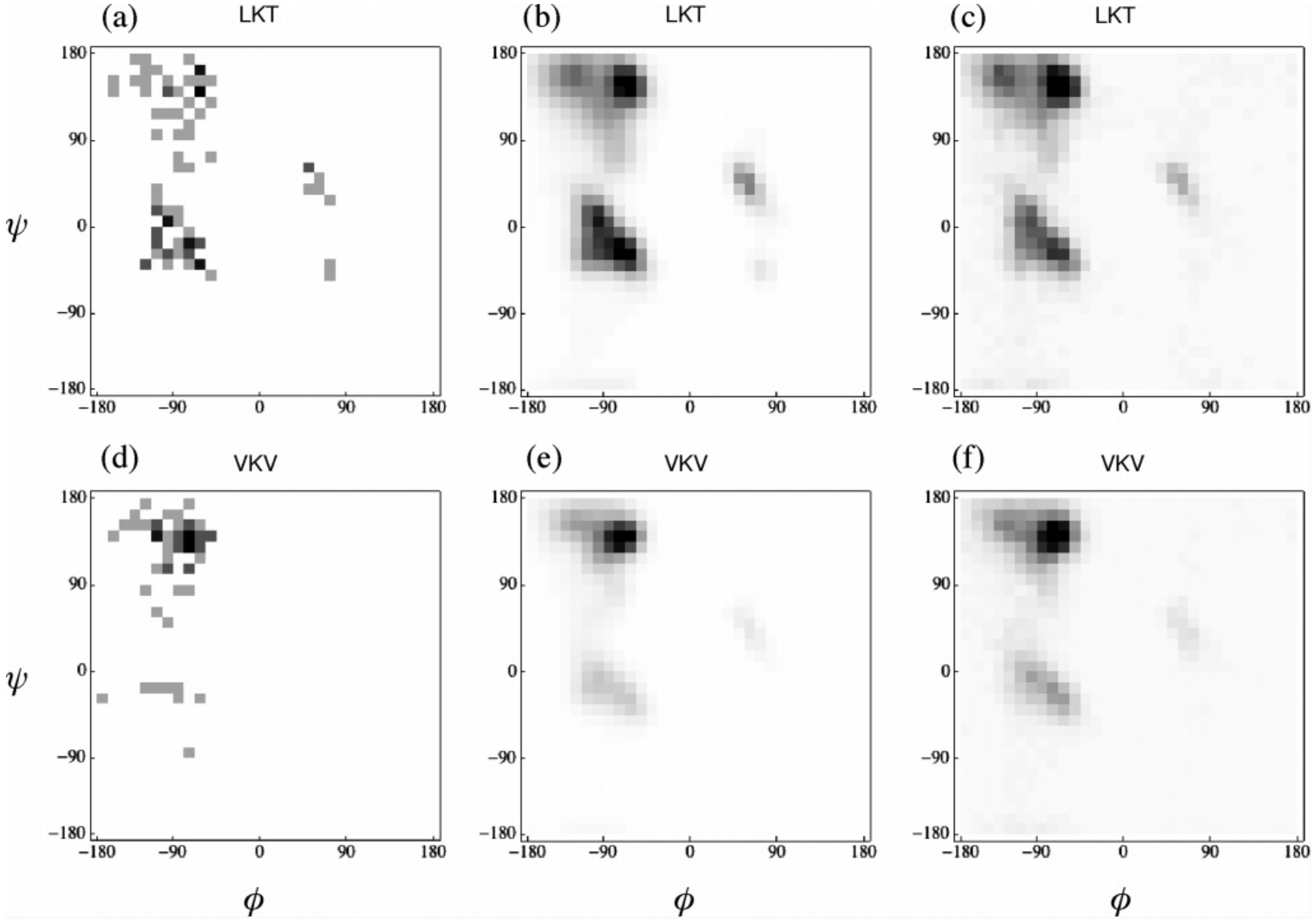

Figure 3. Comparison between initial data points (raw probabilities), final estimated probabilities from the potential, and probabilities obtained from simulations for the Leu-Lys-Thr triplet and for the variant Val-Lys-Val. (a) Initial data points for Leu-Lys-Thr. (b) Probability obtained by exponentiating the negative of the final potential for Leu-Lys-Thr. (c) Probability extracted from simulating an 11-residue fragment of apoMb (RFKHLKTEAEM). (d−f) Same as a−c but using Val-Lys-Val. Note that the dark regions correspond to high probabilities in a linear scale.

(rms) can be computed between the probabilities in step two and the exponential of the negative potentials ($U_{0;abs}$) obtained in step four (after normalizing). The rms error of the probability per bin, averaged over all 8000 probability functions, was found to be only $(8.2 \pm 7) \times 10^{-5}$. A direct comparison between the "raw" data and the probability distribution derived from the potential energy is shown in Figure 3 for two coiled-residue triplets, LKT and VKV. The correspondence of the populated areas between a and b and between d and e is clear. The effects of the probability combinations (step three) are not visible in this linear scale. Unfolded state simulations are also conducted for a selected few proteins (the four used in the next section) to reconstruct the probability distributions of the ϕ, ψ angles in the $U_{0;abs}$ term. Two examples are shown in Figure 3c,f, showing a close correspondence to the potentials in b and e, respectively.

Figure 3 also shows evidence of the effects of the neighboring residue composition on the dihedral angle of the center residue. In coiled conformations, when lysine is surrounded by leucine and threonine, the population of right handed helices and extended conformations are approximately even, Figures 3a−c. However, when the neighbors of lysine are replaced by valines, which are more propense to be in extended conformations, then lysine shows a higher tendency to be found also in an extended conformation, Figures 3c−f.

The potential cross terms (eq 4) show some of the correlation effects between the neighboring residue conformations and the center residue conformations. Several examples are shown in Figure 4. Figure 4a shows that, when leucine is followed by a lysine, there is a higher tendency for the leucine ψ angle to be in a right-handed helical conformation and the lysine ψ angle in a left-handed helical conformation, as shown by the energy minima of $-0.6k_BT_0$ around $(\psi_L, \phi_K) \approx (-20°, 60°)$. Figure 4b shows the ψ, ψ angle correlation potential of a glutamic acid followed by an alanine. There is a preference for both residues to be in a right-handed helical conformation in this case. Figure 4c shows the correlation between the ψ, ψ angles of two consecutive lysines. In this situation, the first lysine shows a preference to be in a left-handed helix when the second one is in an extended conformation typical of β strands but not of polyproline II structures. In general, the potential correlation terms do not show minima lower than $-1.1k_BT_0$, with average minima values (among all duplets) of near $-0.33k_BT_0$. Many of them show fluctuations on the order of $0.1k_BT_0$, which is probably within the noise level. Nevertheless, all three cross terms seem to contribute (for some residues more than for others) to the conformation correlations between adjacent residues. However, the importance of these correlations can be expected to diminish rapidly with rising temperature.

Residual Dipolar Coupling Tests. The RDC obtained by NMR can be used to characterize a persistent (or residual) protein structure in the denaturated state. In the experiments of interest, RDCs are measured by weakly aligning a protein in a slightly anisotropically compressed gel, such as strained-aligned polyacrylamide gels.[11,12] The RDC values depend on the direction of bonds such as the backbone N−H or the C−H bond, which are determined relative to the applied magnetic field,

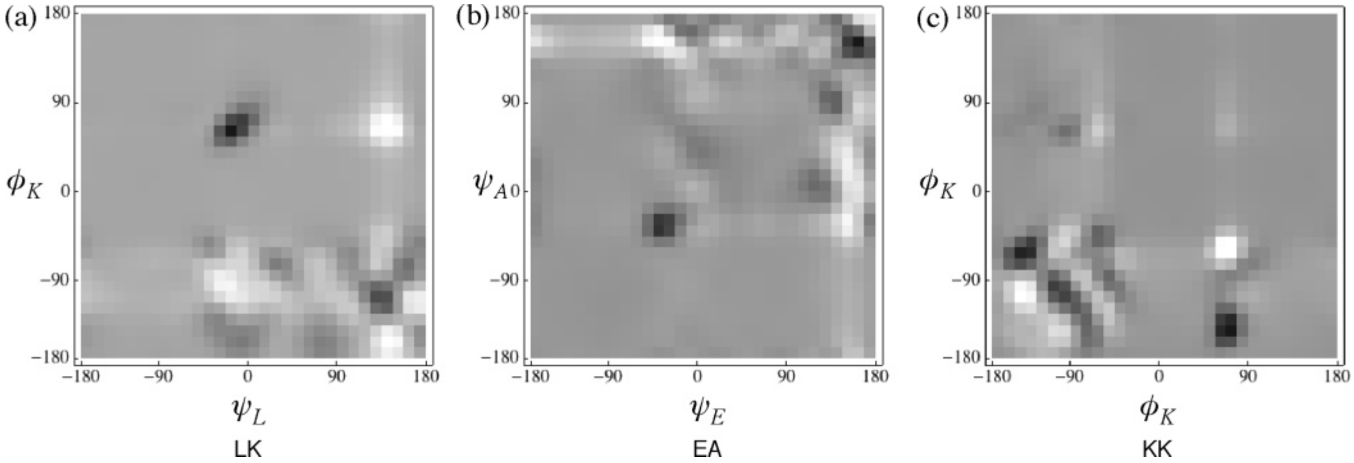

Figure 4. Energy cross terms examples. (a) $U_{1;LK}(\phi_L, \psi_K)$. (b) $U_{2;EA}(\psi_E, \phi_A)$. (c) $U_{3;KK}(\psi_K, \phi_K)$.

which in turn is oriented in the direction of the axis of symmetry of the gel. The protein is inserted into the gel and partially aligned by either stretching or compressing the gel. Because some of the principal axes of unfolded proteins are larger than others, the largest axis is weakly aligned in the direction of the gel stretching, or the shortest axis is weakly aligned in the direction of the gel compression, depending on which straining method is used. For a protein ensemble, this results in a non-vanishing average orientation of either the N−H or the C−H bond. For gel stretching, the average residual dipolar coupling can be computed using

$$R_{\mathrm{NH}} = R_{\max}\langle(3\cos^2\theta - 1)/2\rangle \tag{27}$$

where $R_{\max}$ is a constant, θ is the angle between the N−H bond and the gel stretching direction, and the brackets represent an ensemble average. Note that for the gel compression case, the R_{NH} is obtained from eq 27 after multiplication by a factor of $-1/2$. In the present simulations only the relative RDC can be obtained, that is, $R_{NH}/R_{\max}$ as a function of residue position along the chain.

In computing the RDC, more than 80 000 independent structures were generated for each of the sequences under consideration. A large number of structures were necessary because of the weak signal. The structures were equilibrated at a fixed simulation temperature ($t = 1$) long enough so that the energy converged and the energy auto correlation function became zero. Note that, because the backbone potential is in units of k_BT_0, the actual temperature of the system is $T = tT_0$. A procedure for calculating the confinement effects is given by the PALES[13] algorithm, which works by taking a structure and gradually rotating it in the presence of a wall, eliminating from the ensemble those orientations that collide with the wall. A faster approximation to this method was derived by Almond and Axelsen.[14] Here, instead, a more straightforward approach was used. To simulate confinement effects, the RDC calculations were carried out in a simulated gel consisting of four walls (with square cross section) and two open ends. A two walls case was also investigated, but the results were similar so a four-wall cavity was used to simulate both gel stretching and compressing experiments. The walls consisted of a soft potential given by a linearly increasing repulsive energy outside the wall boundaries, with a slope of 0.5 k_BT_0/Å. The structures were first equilibrated in the absence of walls; then, their longest principal axis was aligned in the direction of the cavity (toward the open ends) and finally equilibrated in the presence of the walls. Wall separations were selected to be smaller than the random flight radius of gyration but larger than their compact radius of gyration.

Potential Components Effects. The derivation and functional form of the potential energy can be partially justified by computing the RDC for several proteins in the unfolded state. One case for which experimental values are available is apomyoglobin (apoMb).[11] The protein was chemically denaturated at room temperature by adding 8 M urea and lowering the pH to 2.3. This has the effect of disrupting the nonbonded interactions without having a major influence on the local structural propensities. In the apoMb simulations, the wall separation value was set to 40 Å, which is smaller than the unfolded radius of gyration (≈42 Å) but larger than the folded radius of gyration (15 Å). The simulations were performed for the sequence given by the file with PDB code 1bvc, which contains 153 residues. The relative RDC was computed for all of the residues (except the first) and aligned to the experimental values by fitting the multiplicative scale of the computed values. To compare the experimental values to the theoretical values, the (Pearson) correlation coefficient $r = (\langle xy\rangle - \langle x\rangle\langle y\rangle)/(\sigma_x\sigma_y)$ and the "overlap" correlation coefficient $\rho = \langle xy\rangle/\sqrt{\langle x^2\rangle\langle y^2\rangle}$ were computed, where x and y are the two variables being compared and σ_x and σ_y are their standard deviations. These numbers also allows us to compare the results to the ones obtained by Jha et al.[4]

Figure 5 shows the experimental RDC values for apoMb and the simulated ones obtained from the optimal potential and from several simplified versions. Figure 5a shows the optimal potential RDC with correlation coefficients of $r = 0.77$ and $\rho = 0.95$. This is a good agreement between the theory and the experiments and is slightly better than the ones obtained by Jha et al.[4] The effects of neighboring residue composition and conformation can be explored by replacing the triplet potential $U_{0;abs}$ replaced by the single residue one $U_{0;*b*}$, and neglecting the cross terms $\Delta U_{i;ab}$. With this modification, the resulting RDC is shown in Figure 5b, resulting in correlation coefficients of $r = 0.63$ and $\rho = 0.93$. In this case, a large overlap is still present but the Pearson correlation has been significantly diminished. Some of the main features are still present, such as the increased tendency toward helical structures near residues 25 and 125. Next, the effects of a residue type independent model can be seen by letting each residue depend on the potential $U_{0;***}$ only. Figure 5c shows the RDC in this case, which results in correlation coefficients of $r = 0.58$ and $\rho = 0.93$. The Pearson correlation is reduced further showing the need for composition

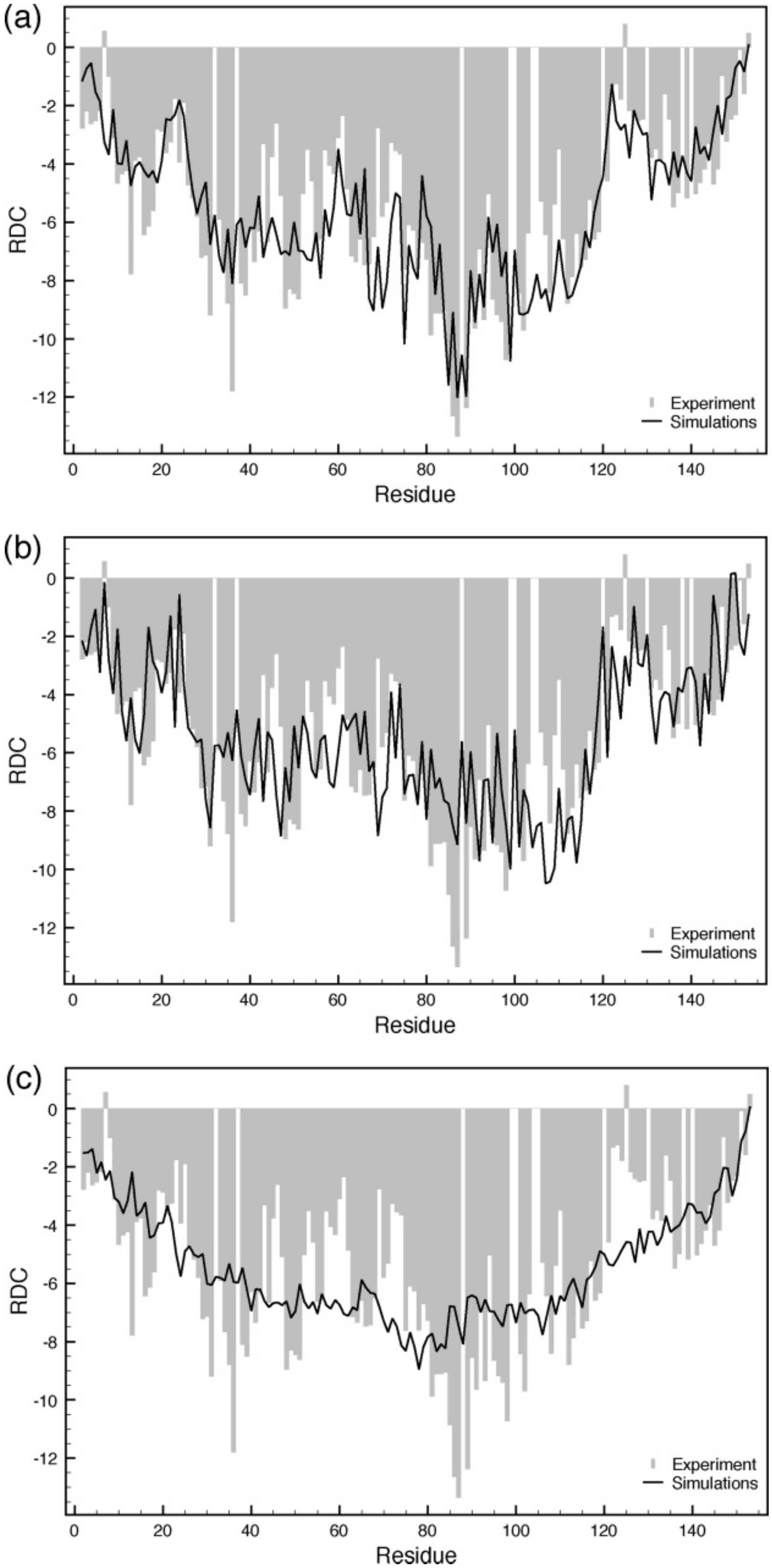

Figure 5. RDC for apoMb and the effects caused by the correlations. (a) RDC obtained from the full potential (eq 3). (b) RDC obtained using a potential with no composition or conformation correlations from nearest neighbor residues ($U_{0;*b*}$ only). (c) Potential with no sequence dependence for the backbone potential (only the average $U_{0;***}$ was used). The gray bars are the denaturated experimental results. A "soft" excluded volume effects between residues separated by three peptide bonds or more is present in all simulations. The coiled residue potential was used in all cases.

dependence. It can be seen that this curve lacks the main fluctuations appearing in Figure 5a,b. Notice that the small fluctuations are due to sampling error. However, the overlap correlation is still high, showing that the preference of a residue to form extended conformation in the unfolded state is the dominant factor in the overall shape of the RDC.

The correlation effects can be eliminated one at a time instead of all together (as was done in Figure 5b). Starting from the

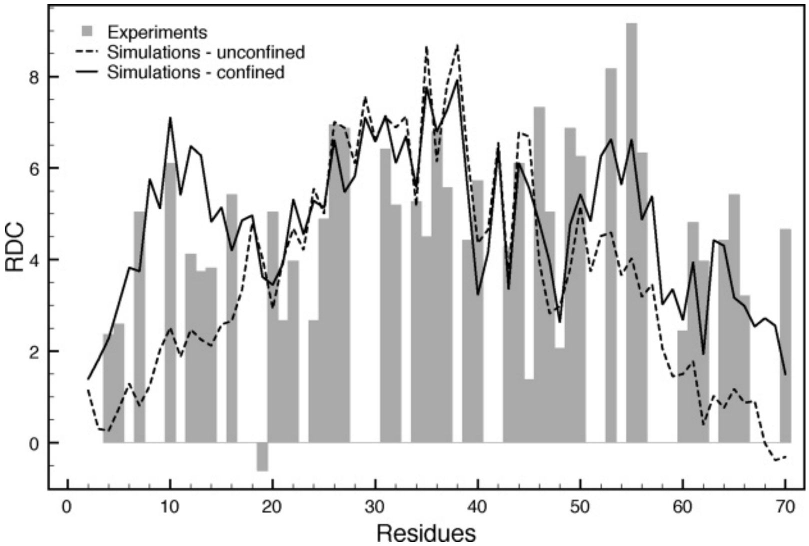

Figure 6. Confinement effects on the RDC of eglin C. The correlation coefficients are $r = 0.32$, $\rho = 0.86$ for the unconfined case and $r = 0.41$, $\rho = 0.93$ for the confined case. The gray bars are the denaturated experimental results. The coiled residue potential was used.

optimal potential (eq 3), if the triplet potential $U_{0;abs}$ is replaced by the single residue one $U_{0;*b*}$ (keeping the cross terms), the correlations become $r = 0.67$ and $\rho = 0.93$. If $U_{0;abs}$ is kept but the cross terms are dropped, then the correlations become $r = 0.71$ and $\rho = 0.94$. In both cases, the correspondence of the model to the experiments diminishes, but the reduction is smaller when the cross terms are eliminated. This suggests that the conformation correlation effects of neighboring residues are of a higher order than the residue composition correlations, as expected from the observations of the potential cross terms.

Confinement Effects. According to Mohana et al.,[11] the conformational ensembles of apoMb in the various unfolded states are not perturbed significantly by the presence of the gel matrix, as suggested by small differences in the heteronuclear single quantum coherence spectra with and without the gel. Nevertheless, the present simulations reveal small but noticeable effects, in particular near the chain ends. To study the RDC with and without walls, the RDC is approximated by measuring the angle in relation to the protein principal axes, as suggested by Jha et al.[4] Regardless of how good this RDC approximation is, it is a clear indication of the confinement effects. The protein eglin C was used for this comparison, with the sequence given by the PDB file 1cse (I). Ohnishi et al.[12] experimentally determined the RDC for denaturated eglin C in 8 M urea and a pH of 3.0. The comparison between the simulations, with and without confinement, and the experiment is shown in Figure 6. Without confinement, the correlation values are ($r = 0.28$, $\rho = 0.85$) and with confinement they improve to ($r = 0.48$, $\rho = 0.94$). It is evident from the plot that the improvement comes from the chain ends (around 20 residues in each end).

RDCs for Various Proteins. In addition to apoMb and eglin C, two other proteins for which RDC data in the denaturated state is available are ubiquitin (Ub)[15] and the Snase fragment 131 (D131D).[7] These proteins were also chemically denaturated at room temperature by adding 8 M urea and lowering the pH (2.5 for Ub and 5.2 for D131D). Simulations were performed for the sequences given by the PDB files 1snq for D131D and 1ubq for Ub. Table 1 shows the correlations between the experimental and the simulated RDC values for all four proteins. These values are consistent and slightly better than those obtained by Jha et al.[4] Table 1 also shows the correlations resulting from a potential that uses the unrestricted residues irrespective of their secondary structure type. The simulations were also carried out at "room" temperature ($t = 1$). For all cases, with the exception of eglin C, both correlations are smaller when the unrestricted residues are used instead of the coiled ones. On the basis of these few results, one can conclude that the potential derived from the coiled residues is more representative of the unbiased dihedral angle potential, although the unrestricted residue potential can be comparable in some cases.

TABLE 1: RDC Correlation Coefficients between Simulations and Experiments

			coiled		unrestricted	
protein	residues	gap[a]	r^b	ρ^c	r	ρ
apoMb	153	40 Å	0.77	0.95	0.49	0.84
eglin C	70	15 Å	0.48	0.94	0.49	0.67
Ub	76	20 Å	0.69	0.95	0.45	0.63
D131D	131	25 Å	0.30	0.87	0.20	0.65

[a] Confinement separation gap. [b] Pearson correlation coefficient. [c] Overlap correlation coefficient.

Native RDC for Eglin C. The Pearson correlation coefficient for eglin C is not very high ($r = 0.48$), which raises questions about the quality of the potential or the correspondence of simulation conditions to the RDCs experiments. Furthermore, simulations using the unrestricted residue potential result in a nearly equal value ($r = 0.49$). Ohnishi et al.[12] found a correlation of $r = 0.51$ between the native and the denaturated RDC. However, a calculation of the overlap correlation coefficient ρ shows that between the native and the denaturated RDC, $\rho = 0.20$, while between the coiled residue potential simulations and the denaturated state RDC, $\rho = 0.94$. This is a more substantial difference, suggesting otherwise that the simulations are more representative of the unfolded state. The experimentally determined RDC for the native and denaturated states, as well as the simulated RDC using the unrestricted residue potential are depicted in Figure 7. It is evident that both the native structure and the structures determined with the unrestricted residue potential have a higher content of helical structures than either in the denaturated state. This is due to the fact that negative (positive) RDCs indicate a higher propensity toward helical (extended) structures. The similarities in the Pearson correlation coefficient but not on the overlap correlation coefficient suggest that there may be some small fraction of H bonds present,

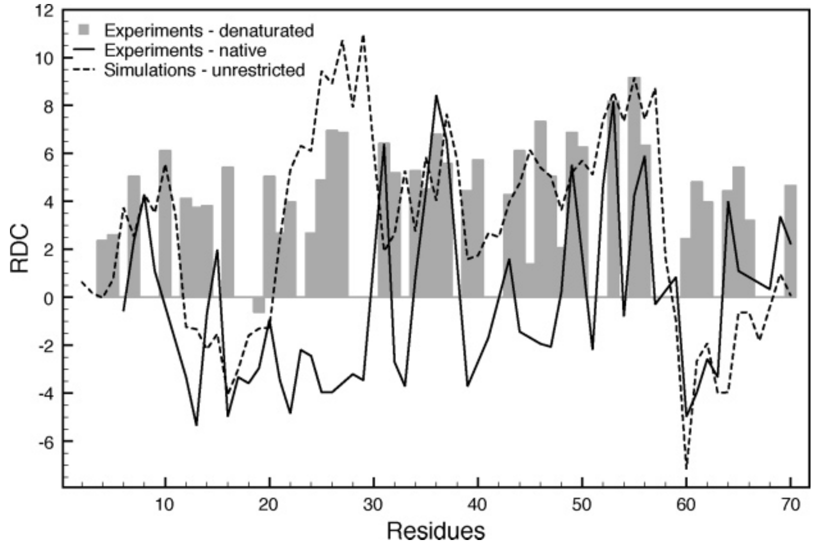

Figure 7. Comparison between the native RDC (solid line), the denaturated RDC simulated using the unrestricted residue potential (dashed line), and the experimental denaturated RDC (gray bars) for eglin C. The correlation coefficients between the denaturated and the native state are $r = 0.51$, $\rho = 0.20$ and between the denaturated state and the unrestricted residue potential simulations are $r = 0.49$, $\rho = 0.67$. In contrast, the coiled residue potential simulations yield correlations of $r = 0.48$, $\rho = 0.94$.

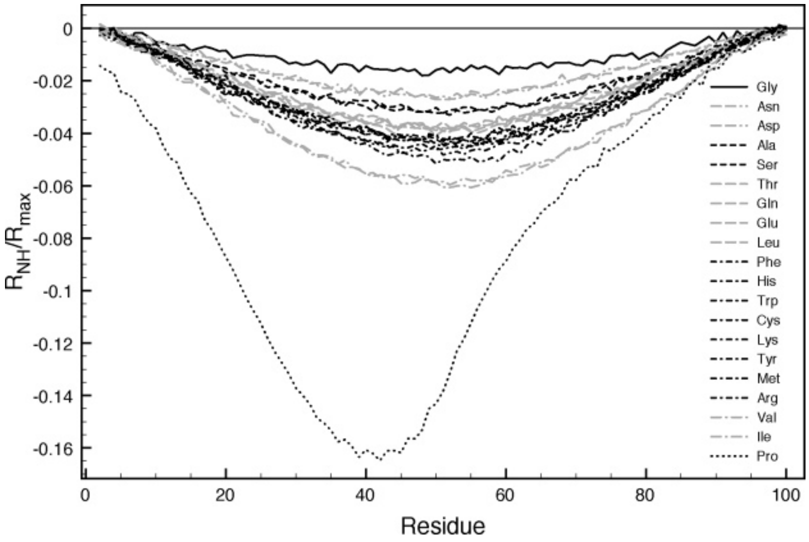

Figure 8. Relative RDC from the simulations of 20 homo-polypeptide chains of 100 residues in length for each one of the 20 amino acids. For clarity, the results are visually clustered using equal line types. The listed residues are approximately arranged in the same order as the curves minima. The coiled residue potential was used.

perhaps due to insufficiently low pH values in which the experimental RDC was determined. This could also be the case for D131D.

Homopolymers. In addition to the proteins described, the RDCs for 20 homopolymers, each composed of one of the 20 amino acid types, was simulated. While the experimental RDC for homopolymers might not be determinable by NMR, the simulation reveals the relative behavior of each of the 20 amino acids in an RDC plot. For simplicity, the RDC for these simulations was calculated by measuring the N−H angles in relation to the principal axis of the protein without a confining media. Figure 8 shows the simulated RDCs for the 20 homopolymers of 100 residues in length. The V shape is typical of random-flight homopolymers[4,16] but the variations of the minimum value with residue type indicate the different structural propensities of the different amino acids. Note that, in a stretched gel, helices with their axis in the direction of the protein principal axis will have their N−H vectors nearly parallel to the principal axis and a relative RDC close to 1, while extended structures (β strands and polyproline II) aligned in the direction of the principal axis will have their N−H vectors perpendicular to the principal axis with a relative RDC close to −1/2.[11] A salient feature of this graph is the curve of polyproline. It shows the strongest tendency to be in extended conformations (as expected). In addition, the graph shows a significant asymmetry around the center of the chain, with a stronger tendency for the N-terminus side of the chain to be in an extended state. An indication of strong asymmetric effects in proline is consistent with asymmetries in the potential cross terms (ΔU), observed in the present model and in the previous one.[3] For the other residues, the minimum in the relative RDC curves varies from −0.016 for glycine to −0.06 for isoleucine, and the curves are approximately symmetric around the middle of the chain. Even for alanine, which has a preference for helical secondary structures in native proteins, the relative RDC values are negative. This is in general agreement with the view that residues in the denaturated state adopt extended conformations.[17]

Entropy Change Upon Mutations. Several attempts to compute the entropy of amino acids in the denaturated state by molecular simulations have been previously made. One of these obtained the entropy change of all 20 amino acids upon substitution by alanine, using Langevin dynamics and employing several popular force fields.[2] The simulations were, in part, performed for single amino acids flanked by two alanines. In particular, the resulting entropy change between alanine and glycine was less than 0.5 kcal/mol for all force fields used. In a more recent and accurate calculation, molecular dynamics was used to compare the entropy change of replacing an alanine in short polypeptides by a glycine.[18] The polypeptides consisted of AXA or GGXGG, where X is either alanine or glycine. These simulations resulted in an entropy change of approximately 0.4 kcal/mol.

A direct comparison can be made between these results and the present model. The absolute entropy is computed from the equation

$$S = -\sum_{ij} P_{ij} \ln P_{ij} \tag{28}$$

where the i, j runs over all bins in the ϕ, ψ space, and P_{ij} is the probability of finding the conformation in bin (i, j). S is expressed in units of k_BT_0, which in reality is the entropic contribution to the free energy T_0S.

The entropy S can be obtained in two different ways: by computing P_{ij}'s directly from the KBP or from simulations. When computing P_{ij} from the KBP, the term U_0 in eq 3 is used for each tri-peptide in the sequence representing the average over the neighboring conformations. The probability is computed by exponentiating $-U_0$ and normalizing it over the ϕ, ψ space. Alternatively, the homo-polypeptide simulations from the previous section can be used to compute P_{ij}. This is not quite the same situation as in the atomic simulations[18] because, here, the residues in question are surrounded by residues of the same type and because the chains are much longer than 3 or 5 residues (as was the case in the atomic simulations), increasing the excluded volume effects. Two cases were considered when computing the entropy directly from the potential: a tri-peptide AXA (X surrounded by two alanines) or a tri-peptide XXX (X surrounded by two residues of the same kind), where X is any one of the 20 amino acids. The latter case was included for comparison with the entropy obtained from the simulations. All entropies are reported in units of kilocalories per mole, using $T_0 = 293$ K (room temperature) when converting from k_BT_0.

The largest absolute entropy value occurs for Gly, whose value in the case of GGG is 3.39. In contrast, the absolute entropy for a uniform (random) distribution is $T_0S_0 \approx 4.03$ kcal/mol (i.e., ln 32^2), which is not that much larger than the entropy for Gly. Table 2 shows the relative entropy ($T_0\Delta S$) of a residue upon substitution by Gly, that is, $\Delta S \equiv S_G - S_X$, which is less sensitive to bin size. The relative entropies resulting from the simulations (using the coiled residue potential) vary from 0.00 kcal/mol (for glycine) up to 1.71 kcal/mol (for proline). These entropies are strongly correlated to the ones obtained from the potential, again showing a consistency between the potential and the simulated ensembles. The entropies obtained from the simulations are slightly larger than the XXX case because of additional cross potential terms. When the unrestricted residue potential is used instead of the coiled residue potential, the absolute entropy values for the triplets surrounded by alanines (the AXA case) are lower by an average of 0.52 kcal/mol. In particular, the entropy is lower by 0.75 kcal/mol for alanine and by 0.23 kcal/mol for glycine. This confirms the expected lower entropy due to the order in regular secondary structure.

Comparisons with the relative entropy values obtained by Zaman et al.[2] for all 20 amino acids do not result in very high correlations. This is in part due to the significant differences among the values generated by the various force fields. The calculated entropy differences obtained from four different force fields (AMBER 94, C-S-94, OPLS-AA-01, and OPLS-UA) can be best correlated to the entropies of the present model using the AXA values (in relation to the AAA entropies). The best overall correlation is obtained for the OPLS-UA model (with Pearson correlation coefficient = 0.80). However, this correlation is dominated by the differences between the proline and the glycine entropies, which are well-separated from the rest. Jha et al. also obtained relative entropies from KBPs.[19] Their entropies and the ones obtained here using the U_0 result in a correlation coefficient of $r = 0.87$ (either for AXA or *X*). More recently, the relative entropy value for alanine was calculated using molecular dynamics by Scott et al.,[18] resulting in a value of ~0.4 kcal/mol for a similar bin size to the one used here at a temperature of 298 K. For the AAA case, the relative entropy value calculated here (i.e., T_0S (AGA) − T_0S (AAA)) is 0.42 kcal/mol, which is very similar.

TABLE 2: Entropy Difference $T_0\Delta S$ upon Substitution of a Residue by Glycine

X	MC[a] ...XXX...	U_0 XXX	U_0[b] AXA	U_0 *X*[c]
Ala	0.45	0.42	0.38	0.40
Arg	0.41	0.35	0.26	0.25
Asn	0.22	0.20	0.17	0.13
Asp	0.27	0.26	0.27	0.20
Cys	0.36	0.33	0.28	0.27
Gln	0.35	0.30	0.33	0.25
Glu	0.42	0.39	0.36	0.30
Gly	0.00	0.00	0.00	0.00
His	0.35	0.32	0.31	0.20
Ile	0.64	0.58	0.59	0.50
Leu	0.47	0.43	0.42	0.40
Lys	0.39	0.36	0.26	0.25
Met	0.45	0.39	0.36	0.33
Phe	0.38	0.36	0.32	0.31
Pro	1.71	1.74	1.27	1.10
Ser	0.40	0.38	0.35	0.35
Thr	0.43	0.41	0.44	0.39
Trp	0.42	0.39	0.32	0.33
Tyr	0.39	0.35	0.33	0.29
Val	0.63	0.61	0.50	0.48

[a] Monte Carlo simulations of 100-residue homo-polypeptides using the coiled residue potential. [b] Entropies computed from U_0. All entropies expressed in units of kcal/mol at 293 K. [c] The symbol * = all residues equally weighted.

4. Conclusions

A procedure for the extraction of high-definition energy functions from known protein structures for the polypetide backbone was described and tested. The potential was develop with the main objective to be used as one of the main components of a protein coarse-grained model for the study of protein folding and dynamic simulations. Unlike other models designed for constructing denaturated proteins, which rely on complementary local potentials or a library of local conformations, the current model is self-contained in the sense that it is designed to capture most of the local interaction effects. A key component of the potential was the inclusion of correlations between a residue's conformation and the conformation and composition of its flanking residues. Given the limited data available, these correlations were modeled by pairwise angle potentials depending on the specific composition of up to three consecutive residues (triplets). An additional problem caused by sparse data was the selection of the grid size in probability density calculations. The solution was to use a density estimation algorithm with an adjustable grid size, based on the satisfaction of a uniformity criteria that yielded more accurate and smoother probability densities. Higher correlations are more difficult to detect because of the lack of data, so they were approximated by merging different levels of residue specificity. In addition, this combination allowed for the estimation of the range of energies between high and low probability dihedral angles. Additional smoothing was performed on the potential functions by using Fourier transform filters. This final step generated a smoother surface suitable for the calculation of the potential gradient that could be used in force field calculations.

Simulations using the potential were able to reproduce dihedral angle distributions consistent with the original observed distributions. The dihedral angle propensities (ϕ, ψ) of a residue of a given type exhibit, in many cases, a significant dependency on the type of residues of its neighbor, irrespective of whether or not the residue is part of a coiled secondary structure. Nevertheless, these correlations could not be obtained for many residue triplets because of the lack of a significant number of events in the database at this time. The correlations between the dihedral angles of neighboring residues was approximated by a residue duplet and a dihedral angle from each residue: (ψ, ϕ), (ψ, ψ), or (ϕ, ϕ). These correlations are much weaker (less than $1.0k_BT_0$) than the composition correlations, and in many cases they were indistinguishable from the noise.

As part of the potential validation, simulations of the unfolded ensemble of proteins in confined environments were conducted. NMR residual dipolar couplings were estimated for several proteins (apoMb, eglin C, Ub, and D131D). There was a significant correlation between the simulated and the measured RDC values for apoMb and Ub and less for eglin C and D131D, consistent with other calculations based on coil library data.[4,5] This suggests that the information provided by coil libraries may be insufficient to reproduce the RDC plots at much higher correlation values, and that other experimental conditions including a more detail consideration of the properties of the solvent (urea, pH, etc.), instead of an implicit solvent, may be necessary. As evidence, Mukrasch et al.[20] showed that by replacing a coiled-based potential by one obtained from simulations using a full atomic force field on a selected few residues of protein Tau, many of the differences between the computed and the experimental RDC can be accounted for. For all proteins, there was an increase in correlation for simulations performed in a confined environment. A comparison of the RDC plots show that confinement affects mainly the chain ends. This strongly suggests that the polyacrylamide gels used in the experiments have an non-negligible effect on the structure of the denaturated proteins, as can also be deduced from the theory of Zweckstetter and Bax.[13] Additional unfolded ensemble simulations were carried out for 20 homo-polypeptides, representing each of the 20 amino acids. In particular, the (hypothetical) RDC for polyproline was found to be asymmetrical around the middle of the chain, with a tendency to be more extended toward the N termini. Finally, it was found that entropy calculations correlate relatively well with molecular dynamic simulations, in particular, with the entropy of substitution of an alanine by a glycine, which yield a value around 0.4 kcal/mol.

Acknowledgment. This work was supported by the Department of Physics at the IUPUI School of Science.

References and Notes

(1) Pappu, R. V.; Srinivasan, R.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12565.

(2) Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R. *J. Mol. Biol.* **2003**, *331*, 693.

(3) Betancourt, M. R.; Skolnick, J. *J. Mol. Biol.* **2004**, *342*, 635.

(4) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099.

(5) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002.

(6) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859.

(7) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487.

(8) Rosenbluth, M. N.; Rosenbluth, A. V. *J. Chem. Phys.* **1955**, *23*, 356.

(9) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.

(10) Betancourt, M. R. *J. Chem. Phys.* **2005**, *123*, 174905.

(11) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131.

(12) Ohnishi, S.; Lee, A. L.; Edgell, M. H.; Shortle, D. *Biochemistry* **2004**, *43*, 4064.

(13) Zweckstetter, M.; Bax. A. *J. Am. Chem. Soc.* **2000**, *122*, 3791.

(14) Almond, A.; Axelsen, J. B. *J. Am. Chem. Soc.* **2002**, *124*, 9986.

(15) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799.

(16) Louhivuori, M.; Paakkonen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647.

(17) Liu, Z.; Chen, K.; Ng, A.; Shi, Z.; Woody, R. W.; Kallenbach, N. R. *J. Am. Chem. Soc.* **2004**, *126*, 15141.

(18) Scott, K. A.; Alonso, D. O. V.; Sato, S.; Fersht, A. R.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2661.

(19) Jha, A. K.; Colubri, A.; Zaman, M. H.; Koide, S.; Sosnick, T. R.; Freed, K. F. *Biochemistry* **2005**, *44*, 9691.

(20) Mukrasch, M. D.; Markwick, P.; Biernat, J.; Bergen, M. v.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235.