Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops

Ionel A. Rata,*,† Yaohang Li,‡ and Eric Jakobsson*,†

Department of Molecular and Integrative Physiology, Department of Biochemistry, UIUC Programs in Biophysics, Neuroscience, and Bioengineering, National Center for Supercomputing Applications, and Beckman Institute, University of Illinois, Urbana, Illinois 61801, and Department of Computer Science, North Carolina A&T State University, Greensboro, North Carolina 27411

Received: October 14, 2009; Revised Manuscript Received: December 18, 2009

Native proteins have been optimized by evolution simultaneously for structure and sequence. Structural databases reflect this interdependency. In this paper, we present a new statistical potential for a reduced backbone representation that has both structure and sequence characteristics as variables. We use information from structural data available in the Protein Coil Library, selected on the basis of resolution and refinement factor. In these structures, the nonlocal interactions are randomly distributed and, thus, average out in statistics, so structural propensities due to local backbone-based interactions can be studied separately. We collect data in the form of local sequence-specific $\varphi-\psi$ backbone dihedral pairs. From these data, we construct dihedral probability density functions (DPDFs) that quantify any adjacent $\varphi-\psi$ pair distribution in the context of all possible combinations of local residue types. We use a probabilistic analysis to deduce how the correlations encoded in the various DPDFs as well as in residue frequencies propagate along the sequence and can be cumulated in a statistical potential capable of efficiently scoring a loop by its backbone conformation and sequence only. Our potential is able to identify with high accuracy the native structure of a loop with a given sequence among possible alternative conformations from sets of well-constructed decoys. Conversely, the potential can also be used for sequence prediction problems and is shown to score the native sequence of a given loop structure among the most fit of the possible sequence combinations. Applications for both structure prediction and sequence design are discussed.

Introduction

Protein structure is fundamental for understanding biological function at the basic molecular level; studying it is an important theoretical and experimental task. The factors that determine the structure and folding of a protein are diverse. A large variety of physical interactions are involved, and their classification and relative importance are sensitive problems.[1,2] For the purpose of this paper, we separate them into local and nonlocal interactions, and we consider "local" the intraresidue interactions and the interactions between nearest-neighbor residues along protein sequence and as "nonlocal" all the other interactions in which the residue participates. Average residue solvation and entropy effects are also included implicitly as part of local interactions.

Local interactions are significant for all levels of protein structure and dynamics. The physical causes involved are understood to a considerable extent.[3–8] The main ones are local steric interactions and entropic effects; local hydrophobic interactions and solvation issues (such as desolvating and screening of the peptide dipoles by the nearby side chains); local hydrogen bonds; polarization; and other electrostatic effects, such as dipole−dipole interactions.

Torsion angles are characteristic to interactions between atoms situated within three consecutive covalent bonds. There are four atoms in such a subsystem, and its corresponding torsion angle is the dihedral formed between the two planes of every three consecutive atoms. The middle bond is the cutting edge of the two planes, and the torsion angle measures the degree of rotation around this bond, with the fully "extended" trans conformation at 180° and the "compact" cis conformation at 0°. In the backbone representation, the main degrees of freedom allowed for structural variation are φ and ψ torsion angles around the backbone bonds that have the C_α atom of a residue at one end. As compared to them, the other backbone structural variations (of bonds, angles, and peptide bond dihedrals) consist only of small displacements around their standard values, which can be incorporated in a mean field $\varphi-\psi$ potential.

It is clear from the Ramachandran plots[9,10] that there is a strong correlation between φ and ψ dihedral angles corresponding to any residue. These plots are residue-specific, showing that the dihedral angle values are also correlated with their corresponding residue type.[11–13] Less obvious, though, are the correlations among dihedrals and types of two sequence adjacent residues. The structural influence of an amino acid configuration on the configurations of its sequence neighbors is called the nearest neighbor effect (NNE). To a first approximation, Flory[14] assumed that the backbone dihedral angles corresponding to one residue are statistically independent from the backbone dihedrals of its neighbors. This assumption basically considers NNE to be negligible (Flory's hypothesis), but an increasing amount of experimental data comes to contradict this hypothesis (see, for example, Grdadolnik et al.[15]). In our work, we show that the near neighbor correlations are sufficiently strong to account for substantial changes in the overall structure of loops and coils. NNE have been recently characterized as having effects as high as changing a residue type.[16] Accurate estimations of the local backbone interactions can compensate for missing side chain

* Corresponding authors. E-mails: (I.A.R.) rata@illinois.edu, (E.J.) jake@ncsa.illinois.edu.
† University of Illinois.
‡ North Carolina A&T State University.

10.1021/jp909874g © 2010 American Chemical Society
Published on Web 01/13/2010

structures, allowing for reduced backbone representations in protein simulations.[17,18]

Evaluating local interactions and their energy contributions in various configurations is a sensible task for any force field that aims to accurately reproduce the thermodynamics of protein systems. Attempts to quantify the local interactions through torsion angle statistics have been made in the literature, including probabilistic analyses based on Boltzmann distribution.[19] Local backbone interactions can be used alone for discrimination purposes[20–25] or in combination with other potentials[17,18,26–30] with special attention for the reference state.[18,29]

In this paper, we present a dihedral potential based on local interactions alone for the reduced protein representation in which the only variables are the residue sequence and the backbone torsion angles. Our potential aims to be useful for both protein structure prediction and design, and our validation methods are focused on these aspects.

Methods

Data Collection. We use a knowledge-based approach of analyzing the sequence−structure data collected from the Protein Coil Library,[31] which contains coils compiled from nonredundant protein structures available in the Protein Data Bank.[32] We chose a coil set selected from proteins with maximum 90% chain sequence identity, maximum 2 Å resolution, and maximum 2.5 refinement factor. We assume that the coil structures are representative for the local interactions so that (1) the effects of the other physical interactions are randomly distributed in the coil database and the statistics will average them out as part of a mean coil environment (the potential of mean force assumption) and (2) the structural effects of the local interactions are distributed according to Boltzmann statistics.[33] This assumption has been shown to work well in similar situations.[34–36]

To improve the validity of these assumptions, we avoid secondary structure elements because they introduce structural biases not representative for local interactions. These are mostly due to repetitive hydrogen bonding patterns. Our purpose is to study the local interactions separately from other interactions, such as nonlocal hydrogen bonds, so we need to avoid any statistical biases generated by them. In coils, nonlocal interactions are also present, but because they are randomly distributed, they do not produce systematic preferences for particular conformations.

At the inception of this work, we used statistical potentials derived from the full PDB library; however, we found that for loops and coils, the propensity of the full PDB potentials for alpha and beta structures overwhelmed other possibilities. With our approach, we got better results using just the structures in the coil library, provided we applied them only to nonalpha, nonbeta regions. Thus, our potentials as we have constructed them cannot be used alone for secondary structure prediction. However, the $\varphi-\psi$ values typically present in secondary structure elements are frequently occurring in coils, as well, showing that even residues in nonregular structures have good "a priori" propensities for secondary structurelike backbone conformations.[37] This fact suggests that a local interaction potential extracted from coil data is "transferable" to secondary structures but has to be used in combination with an additional energy contribution for hydrogen bonding (as successfully accomplished by Frishman et al.[30] for secondary structure prediction).

We do not exclude polyproline II helices from our data set, because they are not hydrogen-bond-driven and are not cooperative beyond the nearest neighbors.[38] Their high frequency in loops is mainly a consequence of backbone solvation[39] and nearest neighbor carbonyl interactions,[40] and thus, they are representative for local interactions as we defined them above. We consider only coils larger than four residues (to avoid short hairpin-like turns with regular structures comprising a hydrogen bond).

The sequence-related structural information collected from coils is organized as follows: For each residue type R, we record its corresponding φ and ψ backbone dihedral angle values from all its occurrences in the coil database. If we represent these values as data points in a $\varphi-\psi$ plane, we obtain coil-specific regular Ramachandran plots for all the 20 residue types. A similar procedure is applied to pairs of residue types when occurring as sequence neighbors: R_i-R_{i+1}. There are 400 possible combinations of residue pairs (doublets). For each of these neighboring residue doublets, we record three kinds of Ramachandran-like distribution plots for all possible adjacent dihedral pairs: for $\varphi_i-\psi_i$ angles of the first residue type (R_i), for $\varphi_{i+1}-\psi_{i+1}$ of the second residue type (R_{i+1}), and for φ_{i+1} of the second residue and ψ_i of the first. A considerably smaller number of data points are available for these doublet residue plots in which all the dihedral angle data are partitioned among 400 plots, as compared to the singlet residue plots in which data is divided into only 20 plots, but the doublet residue plots offer additional sequence−context information, taking into account the nearest neighbor residue effects. Thus, there is a trade-off between the amount of data available for the $\varphi-\psi$ plots and the specificity of their sequence context (singlet- vs doublet-residue-specific). Some methods obtain good results by collecting torsion angle data in the context of triple residue sequence neighbors by dividing the $\varphi-\psi$ plane into regions coarse enough to comprise sufficient available data points.[22]

We also record the number of occurrences of each residue type in the Coil Library. When normalized by dividing by the total number of all residues, they represent the frequency of occurrence of each residue type R. In the limit of large data sets, this frequency is representative for the probability of appearance in coils of the corresponding residue type, $P(R)$. A similar analysis is done for neighboring residue doublets. We compute the frequency of occurrence for all 20 possible residue types, and all 400 residue doublets from our database, through which we obtain the singlet and doublet residue probabilities $P(R_i)$ and, respectively, $P(R_iR_{i+1})$, where R_i and R_{i+1} designates any successive residue types in coils.

Constructing the Dihedral Probability Distribution Functions. From the Ramachandran-like plots of adjacent $\varphi-\psi$ pairs from singlet and doublet residues, we extract sequence-specific quantitative data about dihedral angle distribution. The $\varphi-\psi$ data points in these plots tend to be distributed according to their conformational preferences. In other words, the local density of data points around a $\varphi-\psi$ pair from the distribution plot is proportional to the probability of occurrence of that $\varphi-\psi$ conformation in the sequence context of the plot. In addition, the local density of data points is proportional to the total number of available data points, N. We use these statistical assumptions to convert the Ramachandran-like distribution plots to dihedral probability density functions (DPDFs). A simple method to do this conversion is to divide the $\varphi-\psi$ plane into square bins and to compute the density of data points for each bin (the two-dimensional histogram technique), but because the distribution of data points is irregular (clustering around a few regions) and sparse (especially in the case of rare residue doublets), the histogram method is ineffective for this problem without further optimizations. An optimized histogram technique for obtaining

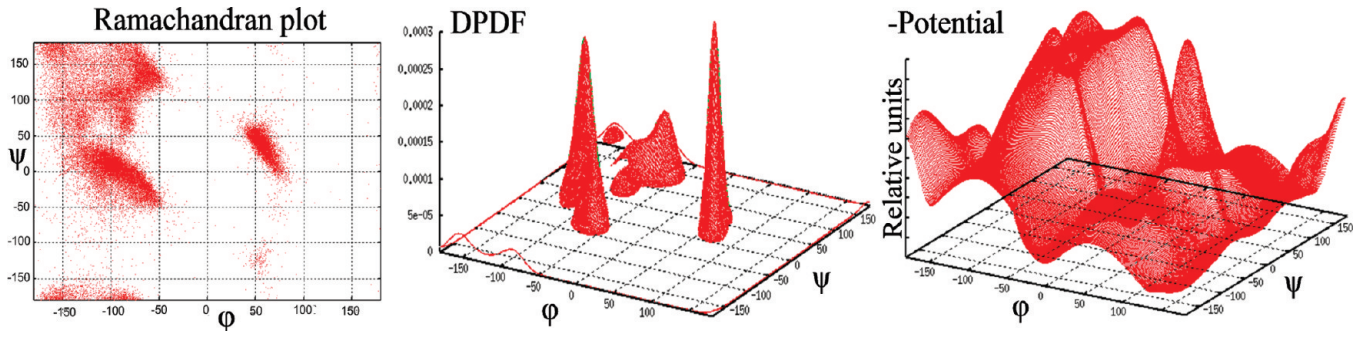

Figure 1. Ramachandran plot, DPDF surface, and minus potential surface for asparagine.

smooth distribution functions has been successfully used for the same problem by Betancourt.[41]

A commonly used method is to divide the $\varphi-\psi$ plane into several predetermined regions of various shapes related to the way the data points are distributed and to compute densities and probabilities for those discrete regions. These are coarse-grained methods that do not allow for computing probability density functions, but offer the possibility of obtaining statistics for up to four consecutive dihedrals corresponding to any two residue neighbors.[23,24,29] On the other hand, constructing smooth DPDFs with available derivatives offers many advantages[20] without constituting a trade-off between the quality of the plots and the data they are constructed from.[21,42] This method involves data point density estimator and smoothing techniques.[12,21,41–44]

Our procedure to construct a DPDF surface and its rationale are described below. A Ramachandran-like plot represents a discrete distribution because of the limited number of data points it contains. For a couple of reasons, when the number of data points is limited, there is a degree of randomness in their distribution beyond such factors as imperfect precision in structure determination. One is that for a small sample size, statistics may not accurately represent the distributions that would be obtained from a large set. A second reason is that the above-mentioned biases due to tertiary environment cancel out only for large sample sizes. The sparser the available data are, the higher the degree of arbitrary biases. That is the reason we cannot know with certainty the influence that each data point should have in constructing the distribution function and the DPDF.

To deal with the above issues and convert a discrete set of data points to a continuous DPDF, we consider the local contribution of each $\varphi-\psi$ data point as a 2D normalized Gaussian surface (or mask) centralized in that data point. For normalization purpose, the resulting DPDF surface will be computed as the sum of the Gaussian surfaces for all data points divided by the number of data points. With this approach, each data point influence on the DPDF surface is focused on a restricted region around it. The closer we are positioned with respect to the data point, the larger its influence is. The half-width sigma of the Gaussian is a measure of how localized this contribution is. In our approach, the half-widths are allowed to have different values for different data points. In regions where the data points are sparsely distributed, their half-widths should be high enough so that they mix their contribution over a larger region and don′t introduce large irregularities in the resulting DPDF due to arbitrary distribution. On the other hand, in highly populated regions, the half width should be low enough to allow for a more localized and refined calculation of the DPDF, which is possible in this case. Therefore, each data point from the plot should be represented by a normalized Gaussian with a half-width depending on the local distribution density, that is, on the DPDF itself. In practice, we deploy an iterative procedure of adjusting the sigma parameters (that determine the half widths) and the DPDF surface successively in a self-consistent manner until convergence is attained.

The algorithm flow is illustrated below:

1. Start with the same values of the half-widths, σ_i, for each data point because at this point, we do not have any information about the distribution function. We should chose a large enough starting value (say, $\sigma_i = 20$ degrees).

2. Construct a DPDF surface by adding up for each data point of index, i, a normalized 2D Gaussian surface of half-width σ_i, centralized in (φ_i, ψ_i). Divide by the total number of available data points, N, for normalization.

3. Compute new values of σ_i for each data point, i, using the current DPDF value calculated at (φ_i, ψ_i). The local density function and, therefore, the "un-normalized" DPDF ($N \times$ DPDF) should be inversely proportional to σ_i^2, so σ_i is inversely proportional to the square root of $N \times$ DPDF(φ_i, ψ_i).

4. If for any data point σ_i differs significantly (more than a predefined margin) from its previous iteration value, go to step 2, where a new iteration is started.

5. Else stop. All σ_i's and the DPDF surface have reached convergence.

This algorithm makes use of only one arbitrary parameter: the proportionality constant from step 3, which we adjusted from our statistical tests presented below. According to assumption 2 from the previous section, the dihedral conformations occurring in the Protein Coil Library can be assumed to be distributed according to Boltzmann statistics.[33] This means that any conformation distribution probability, such as a DPDF, is related to a mean-field potential energy according to Boltzmann formula: DPDF $\propto \exp(-\text{potential}/kT)$. This formula can be used to derive a potential function for the interactions characterized by any DPDF:

$$\text{potential} = -kT\ln(\text{DPDF}) + \text{const.} \qquad (1)$$

For illustration, we show in Figure 1, in order, the Ramachandran plot for asparagine, the DPDF surface obtained with our statistical procedure, and the potential surface with opposite sign calculated with formula 1. This potential is expressed in relative units because of the lack of an absolute reference state. In our work, we circumvent the reference state problem by performing a strictly probabilistic calculus in scoring and comparing various configurations.

We tested the accuracy and consistency of our method for constructing DPDFs and the extent to which it can be reliably used for cases of sparse available data. The test case of asparagine DPDF is presented below. From about 30 000 data points gathered in this case (for which the Ramachandran plot

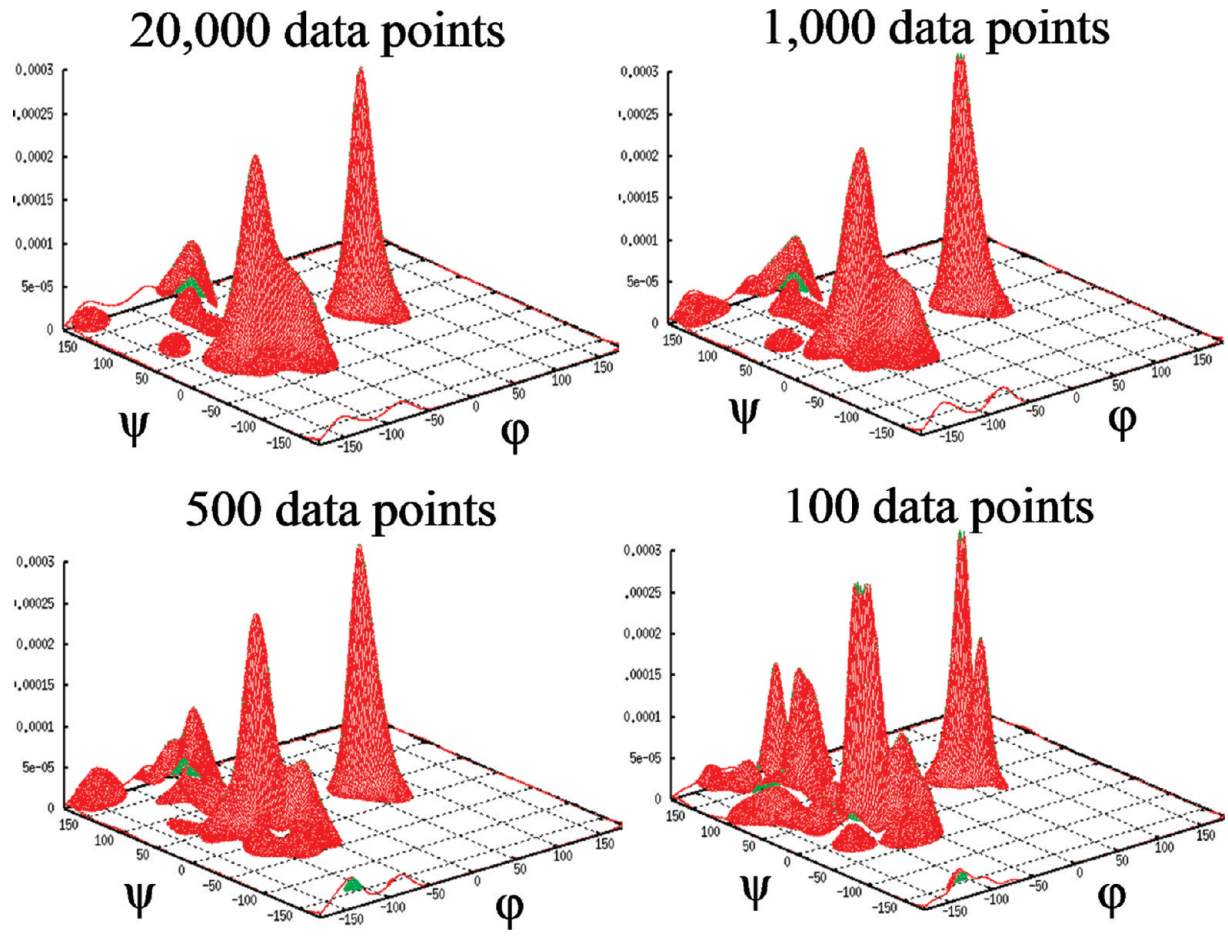

Figure 2. DPDFs for asparagine generated for disjoint data sets with various numbers of data points.

is shown in Figure 1), we arbitrarily separate disjoint sets of, respectively, 20 000, 1000, 500, and 100 data points. DPDFs have been constructed for each set, and their plots are shown in Figure 2. Two important observations can be made by inspecting Figure 2. First, there are no significant differences between the plots corresponding to disjoint data sets if they contain enough data points (see the first two plots from Figure 2). This shows that a DPDF is consistently reproducing the correlations represented by the plot and does not arbitrarily depend on the data set used. Second, Figure 2 shows that our statistical method is consistent for a large range of available data points. The method is reliable even for data sets with a small number of data points (its precision is gradually lost down to around 100 data points). The fact that the method is able to extract a considerable amount of information, even in the cases of sparse data (which occur for some rare residue doublets), is an important advantage of the method.

Backbone Scoring Function for Coils in Sequence and Structure Dimensions. In this section, we calculate an overall configuration probability formula for evaluating coil backbone conformations on the basis of local sequence-structure interactions and the way they propagate along the backbone. With a probabilistic analysis, we can infer the joint probability for a longer chain configuration from local probabilities that we computed from the Protein Coil Library (namely, the DPDFs and the residue frequencies). We consider both backbone dihedrals and residue types as variables in our attempt to take into account the interdependency between the structure and sequence characteristics as they emerge from their distribution in the coil database. For example, the Ramachandran plot for the otherwise symmetric glycine is asymmetric in (φ, ψ), strongly favoring positive φ dihedrals, even though glycine has a symmetric conformation with respect to (φ, ψ); that is, glycine has a strong preference for the left-handed α-helical region in Ramachandran space that is very sparsely populated in other residues because of their chirality. In other words, the relative frequence or occurrence probability of a residue strongly correlates with its backbone conformation. The residues' conditional occurrence probabilities have to be carefully considered for a strict probabilistic calculus.

We consider a coil backbone, as described by the following set of "configuration variables": (φ_1, ψ_1, R_1, ..., φ_i, ψ_i, R_i, ..., φ_n, ψ_n, R_n), where φ_i and ψ_i are the backbone dihedrals of the ith residue in the coil sequence, and R_i denotes its residue type. Our nearest-neighbor-only analysis is implicitly based on several assumptions about local backbone sequence−structure correlations. The assumptions are given below, followed by an assessment of their validity. Note that these assumptions are asserted for nonalpha nonbeta structures only.

Assumption 1: Any backbone dihedral (φ_i or ψ_i) is directly correlated only with its corresponding adjacent dihedrals along the backbone (ψ_{i-1}, ψ_i, or, respectively, φ_i, φ_{i+1}).

Assumption 2: Any backbone dihedral (φ_i or ψ_i) is correlated with its own and with its neighboring residue types (R_{i-1}, R_i, R_{i+1}) and is independent of residues that are farther apart in the sequence. Conversely, any residue type (R_i) is correlated with its own and its neighboring residues torsion angles (φ_{i-1}, ψ_{i-1}, φ_i, ψ_i, φ_{i+1}, ψ_{i+1}) and is not significantly correlated with any other dihedrals.

Assumption 3: Any residue type (R_i) is correlated with the residue types of its nearest neighbors on each side of the

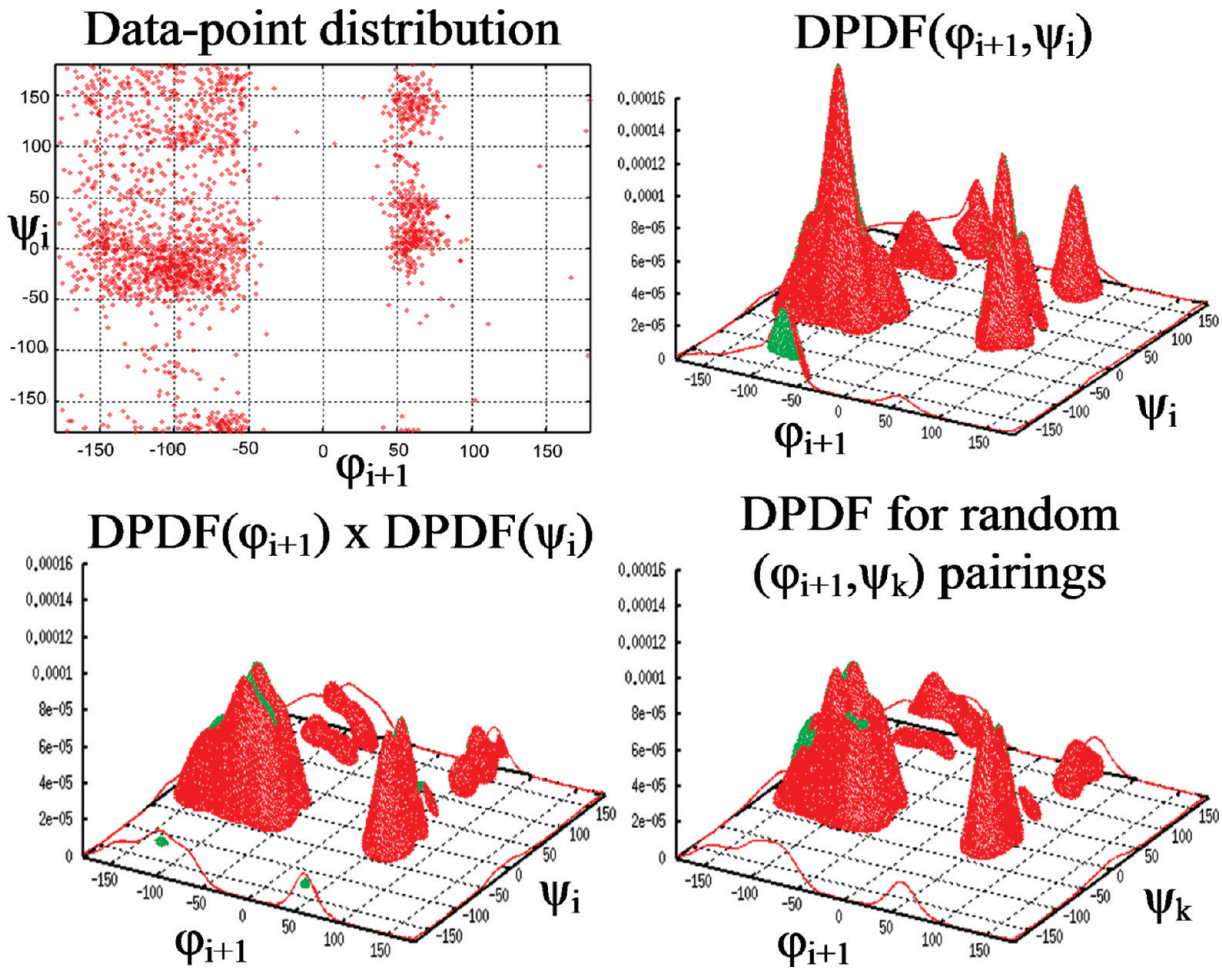

Figure 3. Correlation test for the dihedral angles $\varphi_{i+1}-\psi_i$ corresponding to adjacent Asn-Asp residues.

sequence (R_{i-1} and R_{i+1}) and is independent of farther residue types along the sequence.

Assumption 1 is equivalent to assuming that more distant correlations emerge from the propagation along the chain of the nearest neighbor correlations. Betancourt,[41] for example, considers direct correlations between three consecutive backbone dihedrals using a first-order approximation. For our purpose of constructing a potential characterized only by local interactions, we use only adjacent dihedral pair correlations. The correlation between φ_i and ψ_i is evident from any regular Ramachandran plot (see Figure 1, for example). The data points are clusters in a few particular "allowed regions" of the Ramachandran map. In addition, in three such regions (right- and left-handed α-helix and polyproline II), the distributions are preferentially oriented along negative slope lines in the $\varphi-\psi$ plane, implying a strong "anti-correlation" between the two torsion angles in these regions (Figure 1).

To illustrate the correlation between φ_{i+1} and ψ_i (that is neglected in Flory's hypothesis), we use the DPDF representation for a pair of neighboring residues. We have to show that there is a significant difference between DPDF(φ_{i+1}, ψ_i) and the product DPDF(φ_{i+1}) × DPDF(ψ_i) of the one-dimensional probability density functions. The 1D DPDF for a dihedral is calculated from the 2D DPDF by integration over the other dihedral. Another way to construct the distribution function of the uncorrelated dihedrals is to use the data available for φ_{i+1} and for ψ_i but to randomly mix up their pairings. We can maintain the initial order of data points for the φ_{i+1} values and randomly scramble the order of the ψ_i data set. After scrambling, the new (φ_{i+1}, ψ_i) pairs will no longer correspond to residue neighbors, and there should be no correlation between φ_{i+1} and ψ_i so that DPDF(φ_{i+1}, ψ_i) = DPDF(φ_{i+1}) × DPDF(ψ_i). The resulting functions are plotted in Figure 3 for the case of Asp-Asn residue pair. It can be easily noticed that the first DPDF plot is different from those for the uncorrelated dihedrals. This means that there is a significant correlation between the adjacent dihedrals belonging to successive residues in the sequence. Figure 3 also shows that the last two plots for the uncorrelated dihedrals are very similar, as they should be because they represent the same function (of the uncorrelated distribution), only obtained in different ways. This confirms once again the consistency of the DPDF generation approach and shows that the difference from the first DPDF plot is not just an artifact of the method.

Regarding assumption 2, the fact that the backbone dihedrals of a residue are correlated with the residue type is obvious from the differences in the Ramachandran plots for different residues. The residue type influences the distribution of the backbone dihedrals and therefore is correlated with them. To illustrate the correlation between a residue's dihedrals and a neighboring residue type, we plotted in Figure 4 several DPDFs for the $\varphi-\psi$ angles of isoleucine in the context of different succeeding sequence neighbors (doublet residue DPDFs). The differences in these plots prove the influence that the adjacent residue type has on the $\varphi-\psi$ angles of its preceding residue.

For assumption 3, to prove that R_i and R_{i+1} are correlated, we just need to show that there are instances in which the frequency (or probability) of occurrence of a neighboring residue pair $P(R_iR_{i+1})$ is significantly different from the product of the two residues' individual frequencies $P(R_i) \times P(R_{i+1})$. One

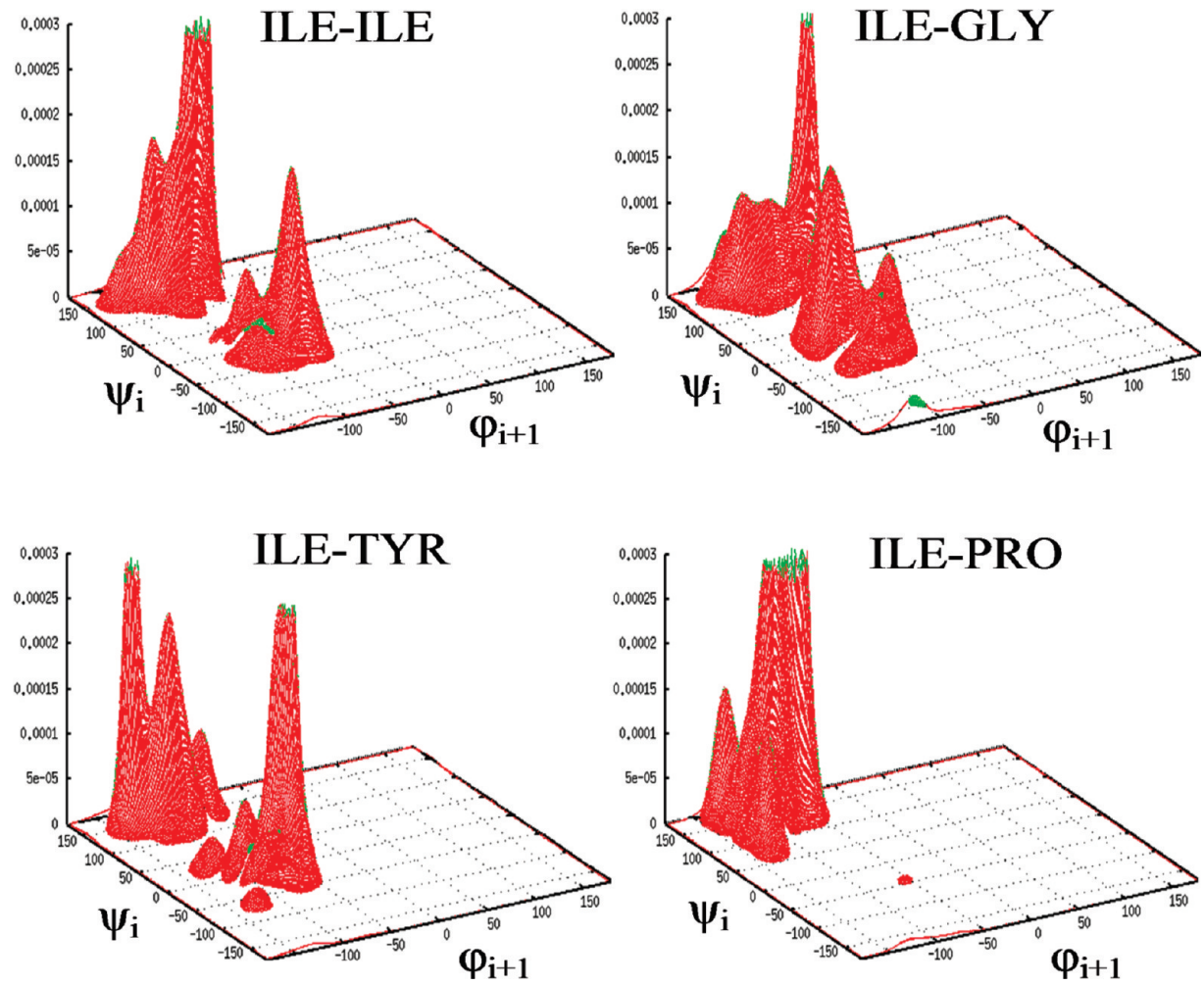

Figure 4. Doublet DPDFs for the $\varphi-\psi$ angles of Ile in the context of four different succeeding residues.

example is the case in which R_{i+1} is proline. It is known that the rigid conformation of proline with a rather fixed φ dihedral imposes restrictions to the dihedral conformation of the preceding residue in the sequence, R_i. Many R_i residue types cannot easily accommodate these restrictions, and for those residues, we have $P(R_i) \times P(R_{i+1}) > P(R_iR_{i+1})$. For example, when R_i = Leu, $P(R_i) \times P(R_{i+1})/P(R_iR_{i+1}) \approx 1.5$. When the two residues are not nearest neighbors, showing that there is no considerable correlation between R_{i-1} and R_{i+1} in the context of any intervenient R_i, φ_i, and ψ_i can be done again with the help of DPDF plots by proving the following formula (which will also be useful for our future discussion), as shown in the Appendix:

$$\mathrm{DPDF}(\varphi_i\psi_i|R_{i-i}R_iR_{i+i}) = \frac{\mathrm{DPDF}(\varphi_i\psi_i|R_{i-i}R_i)\ \mathrm{DPDF}(\varphi_i\psi_i|R_iR_{i+i})}{\mathrm{DPDF}(\varphi_i\psi_i|R_i)} \quad (2)$$

To assess the validity of formula 2, we can plot both the left-hand side (the triple residue DPDF from actual data points) and right-hand side (the "constructed" triplet DPDF) and compare the two plots. In Figure 5, we chose R_{i-1} = Asp and R_{i+1} = Lys, residues of opposing charges that are more probable to interact with each other, and plot the two sides of eq 2 for R_i = Gly and R_i = Leu cases in which there are the most triplet occurrences in the Coil Library, which can be used for more reliable statistics. We can notice the similarity of the actual and constructed triplet DPDF plots. Again, this is a direct consequence of the fact that the correlation between R_{i-1} and R_{i+1} residue types is weak in coils (see the Appendix). For whole proteins (including secondary structure), using actual data sets for triplet residue statistics has been reported to be advantageous.[22]

The above assumptions mainly state that the sequence−structure correlations manifest at a local backbone level and are lost after a reasonable sequence separation. Next, we consider how to "stitch" the local correlations together to evaluate a joint probability for the structure−sequence evaluation of an entire coil. We use the following formula for linked conditional probabilities:

$$P(...R_{i-1}\varphi_i\psi_iR_i\varphi_{i+1}\psi_{i+1}R_{i+1}...) = ...P(R_{i-1}|\varphi_i\psi_i...)\ P(\varphi_i|\psi_iR_i...)\ P(\psi_i|R_i\varphi_{i+1}...)... \quad (3)$$

In formula 3, $P(\varphi_i)$, for example, has the same significance as DPDF(φ_i): namely, the probability of φ_i dihedral occurrence in the limit of a small vicinity around its considered value. The right-hand side of formula 3 contains probabilities of coil configuration values conditioned by all their succeeding values in the set. We will consider explicitly the three probabilities shown in eq 3. In the first one, from its succeeding values, R_{i-1} is correlated only with φ_i and ψ_i, according to assumption 2 above and with residue type R_i (assumption 3). In the second probability, φ_i is correlated only with ψ_i, according to assumption 1 and with residue

types R_i and R_{i+1} from assumption 2. Finally, in the third probability, ψ_i is correlated with φ_{i+1} (assumption 1) and with residue types R_i and R_{i+1} (assumption 2). Under the above assumptions, formula 3 can be rewritten as follows:

$$P(...R_{i-1}\varphi_i\psi_iR_i\varphi_{i+1}\psi_{i+1}R_{i+1}...) = ...P(R_{i-1}|\varphi_i\psi_iR_i)\,P(\varphi_i|\psi_iR_iR_{i+1})\,P(\psi_i|R_i\varphi_{i+1}R_{i+1})...$$

The three right-hand-side probabilities can be expressed in terms of DPDFs and residue frequencies that we calculated from the coil data set, as follows:

$$P(\mathrm{Str}+\mathrm{Seq}) = ...\frac{P(\varphi_i\psi_i|R_{i-1}R_i)\,P(R_{i-1}R_i)}{P(\varphi_i\psi_i|R_i)\,P(R_i)}\frac{P(\varphi_i\psi_i|R_{i-1}R_i)\,P(R_iR_{i+1})}{P(\psi_i|R_iR_{i+1})}\frac{P(\psi_i\varphi_{i+1}|R_iR_{i+1})}{P(\varphi_{i+1}|R_iR_{i+1})}... \quad (4)$$

To reveal the meaning of formula 4 we can rearrange these probabilities as follows:

$$P(\mathrm{Str}+\mathrm{Seq}) = ...\frac{P(\varphi_i\psi_i|R_{i-1}R_i)\,P(\varphi_i\psi_i|R_iR_{i+1})}{P(\varphi_i\psi_i|R_i)}\frac{P(\psi_i\varphi_{i+1}|R_iR_{i+1})}{P(\varphi_{i+1}|R_iR_{i+1})\,P(\psi_i|R_iR_{i+1})}\frac{P(R_{i-1}R_i)}{P(R_i)}...$$

Here, the first fraction represents the triplet residue DPDF as shown in formula 2. The second fraction characterizes the correlation between a pair of dihedrals belonging to successive residues. In the case of no correlation, the DPDF for the dihedral pair from the numerator is equal to the product of the two single-dihedral DPDFs in the denominator, and the fraction is equal to 1. The deviation from unity of this ratio is a measure of correlation between the two dihedrals. This correlation function is dependent on the nearest-neighbor interactions, and we call it the nearest-neighbor dihedral correlation (NNDC). The complete resultant expression of the coil configuration probability is given below.

$$P(\mathrm{Str}+\mathrm{Seq}) = \prod_{i-1}^{n} P(\varphi_i\psi_i|R_{i-1}R_iR_{i-1})\times \prod_{i-0}^{n}\mathrm{NNDC}(\varphi_{i+1}\psi_i|R_iR_{i+1})\times\prod_{i-1}^{n+1}P(R_{i-1}R_i)/\prod_{i-1}^{n}P(R_i) \quad (5)$$

Formula 5 includes configuration values corresponding to the flanking residues R_0 and R_{n+1}. Although they are not part of

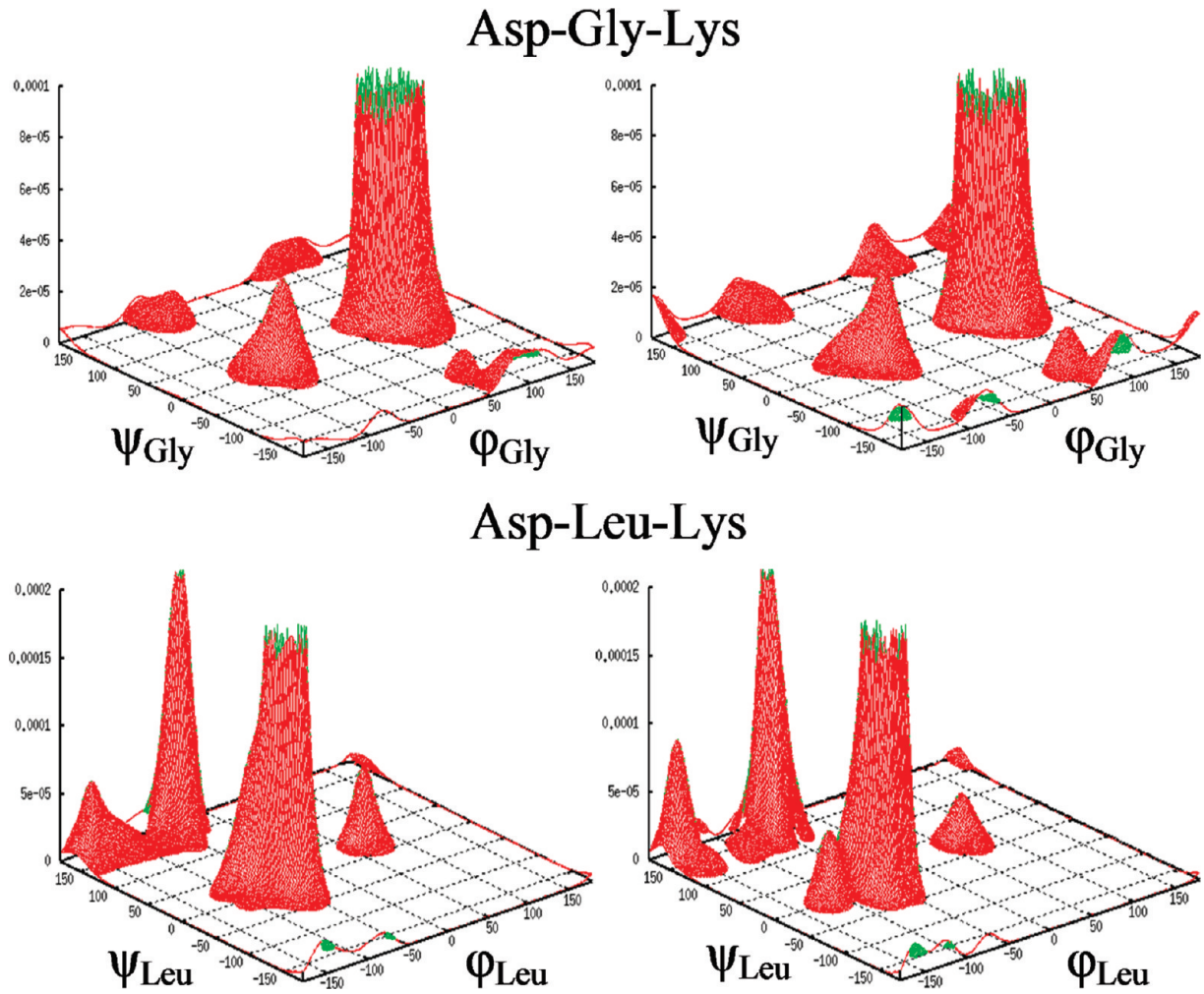

Figure 5. Constructed (formula 2) vs actual data DPDFs (right-hand side) for two residue triplets.

the coil itself and their dihedrals are fixed, they contribute to the coil configuration, probability by the influence they have on their neighbors, R_1 and R_n, respectively. We can prove that the last part of formula 5 represents the occurrence probability of a coil residue combination (sequence) in no given (or general) structural context. We use assumption 3:

$$P(R_0R_1...R_nR_{n+1}) = P(R_0|R_1) \dots P(R_n|R_{n+1})P(R_{n+1}) = \frac{P(R_0R_1)}{P(R_1)} \dots \frac{P(R_nR_{n+1})}{P(R_{n+1})}P(R_{n+1})$$

After reducing $P(R_{n+1})$, we have:

$$P(\text{Sequence}) = \prod_{i=1}^{n+1} P(R_{i-1}R_i) / \prod_{i=1}^{n} P(R_i) \qquad (6)$$

Thus, formula 5 can be written in terms of conditional probabilities:

P(Structure and Sequence) = P(Structure | Sequence) × P(Sequence), with:

$$P(\text{Structure} \mid \text{Sequence}) = \prod_{i\text{-}1}^{n} P(\varphi_i\psi_i|R_{i-i}R_iR_{i+i}) \prod_{i-0}^{n} \text{NNDC}(\varphi_{i+1}\psi_i|R_iR_{i+i}) \qquad (7)$$

The probabilities involved in formula 7 are obtained with data from the entire Coil Library with no explicit reference to the particular global coil configuration that each data point is collected from. In Erman's group,[22,23] a statistical model is used to explicitly adjust the local residue probabilities on the basis of information about each global conformation that the local data are extracted from.

Formula 7 can be used for scoring structures characterized by different backbone dihedral sets for a coil of given residue sequence, but formula 5 (or 4) can, in general, compare configurations with different structures, sequences, or both.

Results

Because our statistical backbone potential is obtained from information about sequence−structure relationship, it depends on both structural (dihedral angles) and sequence (residue types) variables (formula 5). From this observation comes the idea of testing this scoring method performance for two different problems. The first one is the "direct" problem of structure prediction: keeping the sequence of the coil fixed, allow only the structure to vary and evaluate the capability of the potential to select the most probable conformations. The second problem is the "inverse" one of structure design: given a fixed (desired) coil structure, allow only the sequence to vary and estimate the power of the potential to find the suitable residue combinations that are most likely to correspond to the given structure, considering that the native sequence is close to optimal for a given loop structure.[45]

For structure prediction, an experimentally testable evaluation of a potential is based on the thermodynamic hypothesis, which assumes that the native structure is at the absolute minimum of its free energy surface. As a consequence, all other alternative (decoy) conformations can be assumed to score worse on the free energy scale, but we cannot say by exactly how much, and therefore, we cannot reliably order them. We can only say, for example, that from a set of decoys of the same protein coil, our potential function scores the native structure to be better (or more stable) than a certain percentage of all available decoys. The higher this percentage is, the better the scoring function performs. The method relies, of course, on how many and how "good" the available decoys are. This procedure is sometimes also used to optimize the parameters of a scoring function by maximizing the percentage of decoys scored as worse than the native structure. In the assumption that the potential is "transferable" (which is valid for our strictly probabilistic derived function), we can then use other experimentally determined loop structures (from other proteins) with their corresponding decoy sets to obtain additional information for our potential evaluation. As mentioned before, the decoys need to be well-constructed. They have to be stable enough energetically to be competitive with the native structure. They have to cover a sufficiently wide and representative range of conformations around the native structure. The decoys themselves have to be evaluated with some scoring scheme used for their construction and selection. It is important that this scoring scheme does not rely on similar structural features that are used by the tested potential function itself. For testing our statistical backbone potential, we use sets of coil decoys eventually minimized and evaluated using a physical all-atom potential. These Coil Decoy Sets are available online at http://francisco.compbio.ucsf.edu/~jacobson/decoy.htm.[46] They are well-constructed decoys that are used as a benchmark for evaluations of potential functions.

The decoy sets are sorted by coil length. We present here the results obtained for the decoy sets from all available coils with 11 residues, the longest coils with large enough available data sets (generally over 1000 decoys available per coil). Before evaluating the coils from the test set, we need to exclude them from the coil data set used for creating the potential itself (the so-called "jackknife" procedure) so there will be no bias in favor of the native structures of the testing coils. Thus, for this problem, we generate special DPDFs from a new data set with the 11 residue test coils excluded. Table 1 presents the results of testing our potential against the sets of decoys for all 11 residue coils from the Coil Decoys Set. The table shows how the native structure compares with the available decoys as scored with our tested potential. For every coil denominated in the first column, the next column contains the percentage of all available decoys for that coil that are worse than the native structure as scored by our potential. The results consistently show that the native structure is placed among a small percentage of best-scored decoys and, for some instances, is even scored as the best structure in the set. This outcome is remarkable considering that our potential is based on only backbone local interactions, as opposed to the decoy constructing method that uses an all-atom potential.

On the basis of the fact that our potential is sequence-dependent as well as structure-dependent, we use the inverse problem of sequence design to develop a testing method described as follows. Keeping the structure of a given coil fixed, we vary only the residue sequence and evaluate all the possible residue type combinations for that coil. We use the assumption that the native sequence is optimized by evolution in relationship with the coil structure, and given the coil structure, its native sequence should be among the most probable ones in terms of the sequence−structure potential. Here, we should mention that although there usually exist mutations that can stabilize a coil, they do it by affecting the structure, as well. But for a given structure, the native sequence is optimized in correlation with that structure and should score better than the vast majority of

TABLE 1: Percentages of Conformations that Score Worse than the Native Structures for the Decoy Sets of the 11 Residue Coils

coil denomination	structures scoring worse than native, %
NLS	100
1IXH	98.5
5PTI	100
1CSE	98.9
2PTH	89.3
1MSI	100
1FUS	99.4
5P21	93.7
1RCF_1	99.8
1RCF_2	97.1
2CTC_1	99.5
2CTC_2	100
1ABA	100
3SEB	99.3
1A2P	100
1A2Y	98.7
1EZM_1	95.8
1EZM_2	100
1MLA	99.8
1RIE	99.7
2ENG	100
1AKZ	100
153 L	100
1A3C	100
1ADS	100
1ARU_1	100
1ARU_2	100
1BTK	99.9
1CVL	100
1DAD	99.7
1DIM_1	100
1DIM_2	99.4
1PPN_1	97
1PPN_2	100
3PTE	100

alternative sequences.[45] Again, this assumption is valid mostly for coils. It is not valid for helices, for example, in which their relatively fixed structure is stabilized by regular hydrogen bonds irrespective of the local sequence to a high extent.

The challenge is to show that the native sequence of the coil is selected by our potential to be among the best-scored ones. One of the problems here is the accuracy by which the given coil structure is determined experimentally. Coils are usually worse-resolved than the rest of the protein because of their higher flexibility. Often, refinement methods are required for the final coil structure, and the dihedral angles get artificially deviated from their native values. In other words, the method is very sensitive to the resolution and the refinement factor of structure determination. That is the reason we chose for testing illustration one of the best-resolved proteins from Protein Data Bank, protein 1US0 (in the PDB nomenclature). Its structure is determined to an accuracy of 0.66 A, with an R-value of 0.09. The protein is still large enough (316 residues) and contains multiple coils of various sizes. Table 2 presents the results obtained for all the coils between four and nine residues long from this protein. The coils have been carefully chosen not to contain secondary structure segments. The jackknife procedure has been applied in this case, too. All the coils used for testing have been excluded from the data set used for extracting the DPDF data. For each coil of size N, all the possible residue combinations the coil can have (20^N), have been evaluated in our potential, and the percentage of sequences that score worse than the native one has been recorded in Table 2. Data in Table 2 show that, indeed, the native sequence of each coil is placed among a small percentage of the best-scored ones. This highlights the capability of our potential to accurately score different sequences for a given structure as well as different structures for a given sequence.

TABLE 2: Percentages of Sequences Threaded on 1US0 Loops that Score Worse than Native Sequences

coil length	starting residue	sequences scoring worse than native, %
4	37	99.8
4	109	99.1
5	101	99.8
5	151	99.999
6	19	99.9
6	260	99.9
6	292	98.4
6	305	99.9
7	65	99.9
7	186	99.998
7	275	99.9
9	172	99.9999

Conclusions and Discussion

In this work, we present a new statistical analysis on a representative data set of known coil structures for the purpose of quantifying the local protein interactions and placing them in the context of a comprehensive approach to correlating structure with sequence. From this approach, we deduce a backbone potential of mean force based on local interactions only, and we show that it can be efficiently used for structure as well as sequence prediction problems in coils. We show that this reduced representation contains enough information to correlate backbone configurations and residue types at a level of reliability that can have utility for protein structure prediction and design. Thus, the sequence−structure correlations at the local backbone level prove to be important determinants of native protein conformations.

This section considers possible applications of our potentials and further work utilizing and assessing the importance of local interactions in proteins.

Structure prediction is an important application of a force field. For structure prediction problems, it is often convenient to use a hierarchical approach, in which the backbone is generated first and then the side chains are added and optimized. Our method can be reliably applied at the level of backbone construction on which all the following steps rely. Furthermore, because our potential is restricted to local interactions, it can be combined with complementary potentials for nonlocal interactions such as statistical potentials that are described by pairwise atomic interactions. The most obvious use of this potential is in prediction of coil structures. For this purpose, we already combined our torsion angle PDFs with the distance-based PDFs produced by the Subramaniam group.[44] We coupled these torsion and distance-based potentials with a search routine to generate trial structures and assessed the ability to predict loop structures (manuscript in preparation). The method can also be useful in structure homology modeling in which fragments of the modeled structure have to be constructed on a given scaffold inferred from a structurally resolved homologue. The missing fragments are generally coil regions, and their backbone construction is fundamental because there is no reliable a priori side-chain information available in this case.

The fact that our potential is able to accurately discern among well-constructed all-atom structures suggests its applicability as a filtering method of the available alternative coil configura-

tions that are not well-separated by the original potential. In other words, it can serve as a complementary method to other potentials that are not based on the sequence−structure information that we use. The method can be used at various stages of structure optimization since it relies on only backbone information.

Another immediate application consists in designing the sequence that would fit a desired backbone structure for a coil or peptide. The potential can be used to assess potential sequences that would conform to that structure. In the Results section, we proved the capability of our potential to select the fittest sequences that are suitable for a given backbone structure. One can chose from the best-scored sequences that also satisfy other required structural conditions for optimizing nonlocal interactions. Again, our approach can be used in combination with other nonlocal interaction potentials for structure design optimization. A workflow can be developed in which sequence and structure are varied iteratively in cycles for protein design. In such a workflow, a structure with a fixed sequence would be varied in a Monte Carlo calculation with the information-based dihedral potentials used as the energy reference. Then the sequence would be varied on the basis of some of the best-scoring structures to find sequences that would best fit that structure. The iteration would continue until scores stopped getting better. The result would be a set of structures and sequences optimized to each other that could be used as a basis for peptide and protein design. These calculations would help in the understanding of the effects of point mutations on structure. Specifically, this could lead to an informatics-based way of predicting $\Delta\Delta G$ for specific mutations.[47]

All the applications discussed above, about coil prediction and design, apply in a similar manner to binding problems of ligands, toxins, or peptide drugs. In many cases, these have irregular, coil-like structures and they bind to proteins through similar interactions that are formed between a coil and the rest of the protein that the coil is part of.

Finally, an important advantage of our potential consists in its high computational efficiency, because it is based entirely on backbone structures. The side-chain identities and structures enter implicitly by the propensities for $\varphi-\psi$ around different amino acids. The fact that they do not need to be entered explicitly is an enormous computational efficiency advantage. This mitigates the cost of energy calculation as a bottleneck in structure or sequence evaluation problems, when many atoms and interactions have to be taken into account. In addition, the potential can always be computed by a few inexpensive operations between DPDF expressions that are already stored in read-up tables.

Acknowledgment. We acknowledge support from NIH grants 5PN2EY016570-06 and 5R01NS063405-02 and from NSF grants 0835718, 0829382, and 0845702.

Note Added after ASAP Publication. This paper was published on the Web on January 13, 2010, with an error in equation 7. The corrected version was reposted on January 15, 2010.

Appendix

The absence of correlation between the frequencies of any given R_{i-1} and R_{i+1} residue types in the context of any given intervening R_i, φ_i, and ψ_i (assumption 3) can be expressed with the following formula,

$$P(R_{i-1}R_{i+1}|R_i\varphi_i\psi_i) = P(R_{i-1}|R_i\varphi_i\psi_i)\,P(R_{i+1}|R_i\varphi_i\psi_i) \tag{8}$$

which, after rewriting all three conditional probabilities and reducing $P(R_i\phi_i\psi_i)$ becomes

$$P(R_{i-1}R_iR_{i+1}\varphi_i\psi_i) = \frac{P(R_{i-1}R_i\varphi_i\psi_i)\,P(R_iR_{i+1}\varphi_i\psi_i)}{P(R_i\varphi_i\psi_i)} \tag{9}$$

We want to show that from formula 2, we can obtain formula 8 or 9. The conditional probabilities in eq 2 can be written as follows:

$$\frac{P(\varphi_i\psi_iR_{i-1}R_{i+1})}{P(R_{i-1}R_iR_{i+1})} = \frac{P(\varphi_i\psi_iR_{i-1}R_i)\,P(\varphi_i\psi_iR_iR_{i+1})\,P(R_i)}{P(\varphi_i\psi_iR_i)\,P(R_{i-1}R_i)\,P(R_iR_{i+1})} \tag{10}$$

The only way formula 10 is valid for any combination of φ_i, ψ_i, R_{i-1}, R_i, and R_{i+1} is if we separately have

$$P(R_{i-1}R_iR_{i+1}) = \frac{P(R_{i-1}R_i)\,P(R_iR_{i+1})}{P(RR_i)} \tag{11}$$

and $P(\phi_i\psi_iR_{i-1}R_iR_{i+1}) = (P(\phi_i\psi_iR_{i-1}R_i)\;P(\phi_i\psi_iR_iR_{i+1}))/P(\phi_i\psi_iR_i)$, which is equivalent to eq 9.

Formula 11 is actually equivalent to a particular form of assumption 3 for the case of no specifically given dihedral angle context:

$$P(R_{i-1}R_{i+1}|R_i) = P(R_{i-1}|R_i)\,P(R_{i+1}|R_i) \tag{12}$$

References and Notes

(1) Dill, K. A. *Biochemistry* **1990**, *29*, 7133.
(2) Baldwin, R. L. *J. Mol. Biol.* **2007**, *371*, 283.
(3) Ho, B. K.; Brasseur, R. *BMC Struct. Biol.* **2005**, *5*, 11.
(4) Ho, B. K.; Thomas, A.; Brasseur, R. *Protein Sci.* **2003**, *12*, 2508.
(5) Perskie, L. L.; Street, T. O.; Rose, G. D. *Protein Sci.* **2008**, *17*, 1151.
(6) Vijayakumar, M.; Qian, H.; Zhou, H. X. *Proteins: Struct. Funct. Genet.* **1999**, *34*, 497.
(7) Avbelj, F. *J. Mol. Biol.* **2000**, *300*, 1335.
(8) Avbelj, F.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10967.
(9) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.* **1963**, *7*, 95.
(10) Ramachandran, G. N.; Sasisekharan, V. *Adv. Protein Chem.* **1969**, 23.
(11) Anderson, R. J.; Weng, Z. P.; Campbell, R. K.; Jiang, X. L. *Proteins: Struct. Funct. Bioinf.* **2005**, *60*, 679.
(12) Dahl, D. B.; Bohsnnan, Z.; Mo, Q.; Vannucci, M.; Tsai, J. *J. Mol. Biol.* **2008**, *378*, 749.
(13) Hovmoller, S.; Zhou, T.; Ohlson, T. *Acta Crystallogr., D* **2002**, *58*, 768.
(14) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Wiley: New York, 1969; Vol. 30.
(15) Grdadolnik, J.; Grdadolnik, S. G.; Avbelj, F. *J. Phys. Chem. B* **2008**, *112*, 2712.
(16) Xu, C.; Wang, J.; Liu, H. Y. *J. Chem. Theory Comput.* **2008**, *4*, 1348.
(17) Colubri, A.; Jha, A. K.; Shen, M. Y.; Sali, A.; Berry, R. S.; Sosnick, T. R.; Freed, K. F. *J. Mol. Biol.* **2006**, *363*, 535.
(18) Fitzgerald, J. E.; Jha, A. K.; Colubri, A.; Sosnick, T. R.; Freed, K. F. *Protein Sci.* **2007**, *16*, 2123.
(19) Shortle, D. *Protein Sci.* **2003**, *12*, 1298.
(20) Albiero, A.; Tosatto, S. C. E. *Curr. Drug Discovery Technol.* **2006**, *3*, 75.

(21) Amir, E. A. D.; Kalisman, N.; Keasar, C. *Proteins: Struct. Funct. Bioinf.* **2008**, *72*, 62.
(22) Keskin, O.; Yuret, D.; Gursoy, A.; Turkay, M.; Erman, B. *Proteins: Struct. Funct. Bioinf.* **2004**, *55*, 992.
(23) Ormeci, L.; Gursoy, A.; Tunca, G.; Erman, B. *Proteins: Struct. Funct. Bioinf.* **2007**, *66*, 29.
(24) Shortle, D. *Protein Sci.* **2002**, *11*, 18.
(25) Tosatto, S. C. E.; Battistutta, R. *BMC Bioinformatics* **2007**, *8*, 13.
(26) Gilis, D.; Rooman, M. *Theor. Chem. Acc.* **1999**, *101*, 46.
(27) Dehouck, Y.; Gilis, D.; Rooman, M. *Biophys. J.* **2006**, *90*, 4010.
(28) Engin, O.; Sayar, M.; Erman, B. *Phys. Biol.* **2009**, *6*, 13.
(29) Fang, Q. J.; Shortle, D. *Proteins: Struct. Funct. Bioinf.* **2005**, *60*, 90.
(30) Frishman, D.; Argos, P. *Proteins: Struct. Funct. Genet.* **1995**, *23*, 566.
(31) Fitzkee, N. C.; Fleming, P. J.; Rose, G. D. *Proteins: Struct. Funct. Bioinf.* **2005**, *58*, 852.
(32) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.
(33) Sippl, M. J. *J. Comput.-Aided Mol. Design* **1993**, *7*, 473.
(34) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. *Proteins: Struct. Funct. Genet.* **1995**, *23*, 142.
(35) Finkelstein, A. V.; Gutin, A. M.; Badretdinov, A. Y. *Subcell. Biochem.* **1995**, *24*, 1.
(36) Butterfoss, G. L.; Hermans, J. *Protein Sci.* **2003**, *12*, 2719.
(37) Jha, A. K.; Colubri, A.; Zaman, M. H.; Koide, S.; Sosnick, T. R.; Freed, K. F. *Biochemistry* **2005**, *44*, 9691.
(38) Chen, K.; Liu, Z. G.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15352.
(39) Chellgren, B. W.; Creamer, T. P. *Biochemistry* **2004**, *43*, 5864.
(40) Maccallum, P. H.; Poet, R.; Milnerwhite, E. J. *J. Mol. Biol.* **1995**, *248*, 374.
(41) Betancourt, M. R. *J. Phys. Chem. B* **2008**, *112*, 5058.
(42) Lovell, S. C.; Davis, I. W.; Adrendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins: Struct. Funct. Genet.* **2003**, *50*, 437.
(43) Lennox, K. P.; Dahl, D. B.; Vannucci, M.; Tsai, J. W. *J. Am. Stat. Assoc.* **2009**, *104*, 586.
(44) Rojnuckarin, A.; Subramaniam, S. *Proteins: Struct. Funct. Genet.* **1999**, *36*, 54.
(45) Kuhlman, B.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10383.
(46) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins: Struct. Funct. Bioinf.* **2004**, *55*, 351.
(47) Gilis, D.; Rooman, M. *J. Mol. Biol.* **1996**, *257*, 1112.

JP909874G