doi:10.1016/j.jmb.2004.06.091 J. Mol. Biol. (2004) 342, 635–649

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Local Propensities and Statistical Potentials of Backbone Dihedral Angles in Proteins

Marcos R. Betancourt* and Jeffrey Skolnick

University at Buffalo Center of Excellence in Bioinformatics 901 Washington St., Suite 300 Buffalo, NY 14203, USA

*Corresponding author

The following three issues concerning the backbone dihedral angles of protein structures are presented. (1) How do the dihedral angles of the 20 amino acids depend on the identity and conformation of their nearest residues? (2) To what extent are the native dihedral angles determined by local (dihedral) potentials? (3) How to build a knowledge-based potential for a residue's dihedral angles, considering the identity and conformation of its nearest residues? We find that the dihedral angle distribution for a residue can significantly depend on the identity and conformation of its adjacent residues. These correlations are in sharp contrast to the Flory isolated-pair hypothesis. Statistical potentials are built for all combinations of residue triplets and depend on the dihedral angles between consecutive residues. First, a low-resolution potential is obtained, which only differentiates between the main populated basins in the dihedral angle density plots. Minimization of the dihedral potential for 125 test proteins reveals that most native α-helical residues (89%) and a large fraction of native β-sheet residues (47%) adopt conformations close to their native one. For native loop residues, the percentage is 48%. It is also found that this fraction is higher for residues away from the ends of α or β secondary structure elements. In addition, a higher resolution potential is built as a function of dihedral angles by a smoothing procedure and continuous functions interpolations. Monte Carlo energy minimization with this potential results in a lower fraction for native β-sheet residues. Nevertheless, because of the higher flexibility and entropy of β structures, they could be preferred under the influence of non-local interactions. In general, most α-helices and many β-sheets are strongly determined by the local potential, while the conformations in loops and near the end of β-sheets are more influenced by non-local interactions.

Keywords: protein folding; knowledge based potentials; protein secondary structure; backbone dihedral angles; Ramachandran plots

Introduction

The relation between the structure of a protein and its energy is the central problem in protein modeling. The overall structure of proteins can be described by their backbone dihedral angles. Regular patterns in dihedral angles are indicative of the protein's secondary structure such as α-helices and β-sheets.[1] A way to visualize the allowed values of the dihedral angles and their populations is by generating Ramachandran plots,[2] which show the correlations between the φ, ψ dihedral angles. Recently, Hovmoller *et al*.[3] studied the Ramachandran plots of native proteins for each individual amino acid. This calculation was made possible by the abundant number of available protein structures that have been determined experimentally. The results show marked differences between the plots of different amino acid residues.

The type of secondary structure and dihedral angles the residues adopt are determined from a balance between local interactions (those close in sequence) and non-local ones. The population of dihedral angles in a Ramachandran plot is a partial reflection of the local interaction free energies.

Abbreviations used: PDB, Protein Data Bank; RSSD, residue secondary structure depth; KBPs, knowledge based potentials.

E-mail address of the corresponding author: mbetancourt@mailaps.org

Effective statistical potentials can be extracted from these populations. In recent quantum mechanical computations on peptides, a remarkable agreement was found between the peptides' (relative) conformational energies and the statistical potentials retrieved from their crystal structures.[4] Such local potentials could be combined with non-local potentials to study protein folding. We can expect that the role of a local potential is to reduce the accessible conformational space of a polypeptide by restricting and biasing the dihedral angles to values preferred by the local sequence arrangements.[5]

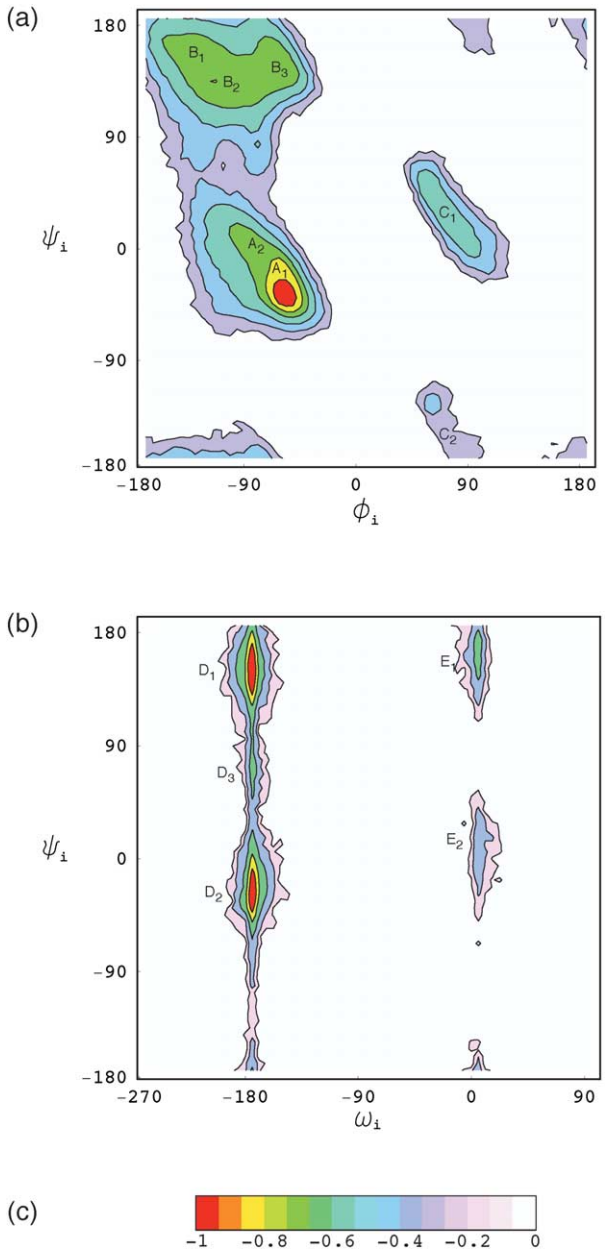

Figure 1. (a) Combined ϕ, ψ free energy contour plot for all residues (except Pro). High energy regions were eliminated for clarity. The data for this and all other free-energy plots are interpolated by the contour plot algorithm. The basins are defined by the outer contour lines and the following conditions: A ($130 \le \phi < 360$, $-100 \le \psi < 50$); B ($130 \le \phi < 360$, $50 \le \psi < 260$); C1 ($0 \le \phi < 130$, $-100 \le \psi < 100$); and C2 ($0 \le \phi < 130$, $100 \le \psi < 260$). (b) ω, ψ density plot for Pro. The basins are defined by the outer contour lines and following conditions: D falls in the ($-100 \le \omega < 90$) range and E falls in the ($-450 \le \omega < -100$) range. The ψ angle boundaries are: D1 ($90 \le \psi < 230$), D2 ($-130 \le \psi < 30$), D3 ($30 \le \psi < 90$), E1 ($90 \le \psi < 230$), and E2 ($-130 \le \psi < 90$). (c) Normalized scale.

The growing number of protein structures deposited in the Protein Data Bank (PDB) allows the study of the free energies involving higher dihedral angle correlations. Ramachandran plots for a particular residue can be affected by the nature of the adjacent residues due to their own torsional propensities and side-chain interactions. It has been argued on the basis of the conformational enumeration of polyalanyl chains[5] and molecular dynamic simulations of monomers, dimers, and trimers[6] that the Ramachandran basin populations are affected by their nearest neighbors, in contrast to the Flory isolated-pair hypothesis. Other molecular dynamic simulations on polyalanyl chains[7] have suggested that non-local interactions of more than five residues away do not violate the Flory isolated-pair hypothesis, but local ones do. The populations are affected, in particular, by the neighbor's conformation and their identity.[6] It can be expected that correlations between a residue's conformation and that of its neighbors are responsible for cooperative effects that enhance the formation of regular secondary structures.

The purpose of this work is to study the effects of the nearest neighbors' identity and conformation on the backbone dihedral angle populations of protein residues in their native state and to derive corresponding dihedral angle statistical potentials that include these effects. Pairwise correlations between the ϕ, ψ angles (or ω for proline) are determined for all combinations of three residues. They include angle correlations between different residues, such as the ψ angle of a residue with the ψ angle of its neighbors. These correlations are used to build low and high-resolution dihedral angle statistical potentials. The low-resolution potentials differentiate mainly between coiled and extended states. Low-resolution free energies are compared to the ones obtained independently of the adjacent residue types and are used to determine to what extent they determine the native secondary structure of proteins. The high-resolution potential, involving a finer dihedral angle partition, is computed and fitted to continuous periodic functions. The results of the minimization of this potential for a group of proteins are described.

Results

General sequence fragments properties

Figure 1(a) shows the Ramachandran free-energy contour plot for all residues excluding proline. The data come from the structures of high quality

non-homologous proteins (see Materials and Methods). The free energy is given by:

$$V(\phi,\psi) = -\ln[1 + N(\phi,\psi)] \quad (1)$$

where $N(\phi,\psi)$ is the number of residues with dihedral angles ϕ, ψ. The ϕ, ψ space was divided in 64×64 bins. Note that adding one to $N(\phi,\psi)$ avoids infinities for bins with no data and sets a reference zero potential. For proline, an analogous Ramachandran plot was generated by replacing ϕ with ω as shown in Figure 1(b). Proline's ϕ angle is nearly fixed while its ω angle (the one preceding ϕ) is more variable and therefore provides more information on the residue backbone conformation. For future reference, we loosely label different regions of the Ramachandran plots as shown in Figure 1. Regions A1 and A2 contain angles predominant in right-handed α-helices. There is a sharp peak at A1, where the regular right-handed α-helices are located. A2 is a smaller peak containing the more deformed helices, such as short helices or helix ends.[3] Loops and some unpaired sheets are also contained in A1 and A2. For convenience, regions A1 and A2 are collectively called A. A similar convention is used for the other regions. B1 and B2 contain parallel and antiparallel β-sheets. B3 is mostly populated by loops and unpaired β strands. C1 contains left-handed α-helices and turns, and C2 is generated mainly by glycine in loops. Smaller regions could be identified,[8] but these suffice for our purposes. For proline (Figure 1(b)), five regions are defined. D1, E1 and D3 correspond to the same ψ range as the B region while D2 and E2 correspond to the A ψ range. In D, proline is in the *trans* conformation while in E it is in the *cis* conformation.

The conformations of 755,973 consecutive three-residue combinations (triplets) were used for obtaining higher dihedral angle correlations. The number of residues of each type are shown in Table 1, which correspond to the proportions commonly observed.[1] The distribution of triplet samples in our set is Poisson-like, ranging from 1 to 449, with a maximum peaking around 40 samples. On an average, there are 94 samples per triplet, which should provide sufficient data for our statistics. Only for a small fraction of triplets, the number of samples becomes inadequate. For example, 8% of the triplets have less than 20 samples. The observed appearance of some residues deviate from what is expected at random, based on the number of residues of each type. Table 2 shows the triplets with lowest and highest samples. As expected, triplets containing Cys and Trp are less common, while those containing Ala and Leu are more common. The Table also shows the deviation from the expected (random) value, C_o in units of their estimated standard deviation ($\sqrt{C_o}$), i.e. the Z-score. C_o is proportional to the product of the single residue probabilities. For the less abundant triplets, the deviation is not too significant, while for the more abundant ones, their deviation is much more significant. Table 3 shows the triplets with extreme deviations from the expected numbers. The triplet with the largest deviation is HHH. Many of these HHH triplets arise as part of histidine tags. A closer examination shows that these tags contain six histidine residues or less and their dihedral angles are spread over the A and B Ramachandran regions with no noticeable biases. The triplets with extreme deviations also show a frequent appearance of proline at the center position. In spite of these extreme cases, the number of repeated triplets is strongly correlated to the expected values, with a correlation coefficient of 0.97 and a linear regression slope of 0.997.

The low sequence identity criteria on the data set does not exclude the possibility of repeated

Table 1. Number of triplets for each amino acid (aa)

aa	Triplets	aa	Triplets	aa	Triplets	aa	Triplets
Ala	57,270	Gln	30,721	Leu	62,916	Ser	44,898
Arg	38,318	Glu	49,775	Lys	45,625	Thr	41,922
Asn	35,573	Gly	52,383	Met	17,315	Trp	11,752
Asp	44,809	His	19,359	Phe	32,288	Tyr	28,888
Cys	12,160	Ile	44,893	Pro	32,863	Val	52,245

The total number of triplets is 755,973 (see Materials and Methods).

Table 2. Triplets with the lowest and highest samples C

Lowest samples				Highest samples			
Triplet	C_o	C	$(C-C_o)/\sqrt{C_o}$	Triplet	C_o	C	$(C-C_o)/\sqrt{C_o}$
CMW	2.94	1	−1.13	AAA	328.68	449	6.64
WCW	4.33	1	−1.60	ALA	361.08	435	3.89
CCC	2.94	2	−0.55	LAA	361.08	434	3.84
CMC	2.94	2	−0.55	AAL	361.08	411	2.63
CWF	3.04	2	−0.60	ALL	396.68	394	−0.13

C_o is the expected samples, estimated from the number of residues of each type. $(C-C_o)\ \sqrt{C_o}$ is an estimate of the samples Z-score.

Table 3. Triplets with extreme deviations from average (Z-score)

Less than expected				More than expected			
Triplet	C_o	C	$(C-C_o)/\sqrt{C_o}$	Triplet	C_o	C	$(C-C_o)/\sqrt{C_o}$
GPP	98.99	30	−6.93	HHH	12.70	74	17.20
KPP	86.22	32	−5.84	NPE	101.82	186	8.34
LIL	310.95	211	−5.67	TPE	119.99	209	8.13
DPP	84.68	33	−5.62	HPD	49.88	104	7.66
PPD	84.68	33	−5.62	LEK	250.01	364	7.21

C and C_0 are the actual and expected triplet samples, respectively.

Table 4. Observed and expected quintuplet samples

Samples	C	C_o	Samples	C	C_o	Samples	C	C_o
1	472,718	498,980	5	1279	350	9	18	0
2	93,646	97,046	6	392	25	10	7	0
3	20,690	16,417	7	157	6	11	0	0
4	5108	2241	8	61	0	12	2	0

The expected numbers are estimated from the residues numbers. One additional observed event with 30 samples for HHHHH was omitted from the Table.

sequence fragments of significant length. However, these are very rare. Analysis of fragments with five residues (quintuplets) gives a good idea of the probability of finding longer repeated sequences in our set. The number of combinations of quintuplets is 3,200,000, which is more than four times larger than the number of quintuplets available. The distribution of quintuplet appearances is sharply peaked around one. Table 4 shows the distribution. Around 99% of the quintuplets appear four times or less. The distribution is very close to the expected one based on the residue abundance. The three most common quintuplets (along with their counts) are HHHHH:30, AAAAA:12, and AALAA:12. Most of the histidine quintuplets belong to histidine sextuplet tags. These results show that the statistics obtained for triplets are not significantly biased by longer homologous sequences and that there is not much general information that can be obtained from longer continuous sets of residues because of their low count. Note that an exception to the latter remark comes from the possibility of using a lower resolution description of the amino acid types.[9]

Dependency of a residue's dihedral angles on its neighbor residue types

The influence of the nearest neighbor's identity on the amino acid dihedral angles can be measured in part by looking at the changes in probability of being in each of the individual basins shown in Figure 1, with and without considering the neighbor's identity. In particular, the basins are defined by the outer contour lines, which contain more than 99% of the cases, and the boundaries described in each Figure caption. The few angles falling outside the outer contour lines are left unclassified. For each residue, the ϕ, ψ basin probability is computed, both independent of and as a function of the adjacent residue types. For each basin, the difference between these probabilities (i.e. neighbor

Table 5. Basin probabilities for the 20 amino acids, independent of the adjacent residue composition

	A	B	C1	Out		A	B	C1	Out
Ala	0.64	0.34	0.01	0.01	Leu	0.59	0.40	0.01	0.00
Arg	0.59	0.38	0.03	0.01	Lys	0.60	0.36	0.03	0.01
Asn	0.49	0.39	0.11	0.01	Met	0.57	0.41	0.01	0.00
Asp	0.54	0.39	0.05	0.01	Phe	0.48	0.50	0.02	0.00
Cys	0.44	0.53	0.02	0.00	Ser	0.50	0.47	0.02	0.01
Gln	0.61	0.35	0.03	0.01	Thr	0.47	0.52	0.00	0.01
Glu	0.66	0.31	0.02	0.01	Trp	0.52	0.46	0.01	0.01
Gly	0.22	0.24	0.31	0.12	Tyr	0.48	0.50	0.02	0.01
His	0.50	0.45	0.04	0.01	Val	0.42	0.58	0.00	0.00
Ile	0.46	0.54	0.00	0.00					

Out is the probability of being outside the basins. In addition, Gly has a probability of 0.12 of being in C2. For Pro, the probabilities are: D1 0.51; D2 0.41; D3 0.02; E1 0.03; E2 0.01; out 0.03.

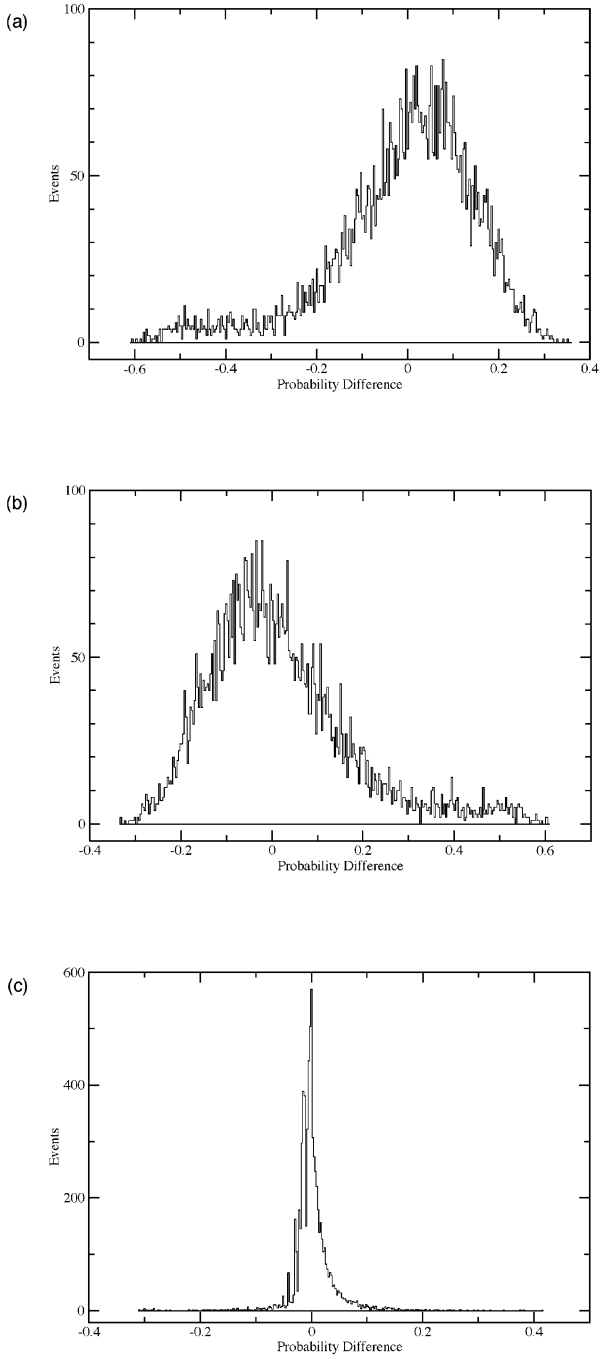

Figure 2. Probability difference between the single residue and the triplet probabilities for basins (a) A, (b) B, and (c) C1. All residue types were used except for Pro as the center residue.

dependent probability−neighbor independent probability) is computed, giving an indication of how much the basin probability changes when the neighbor identities are considered. Note that this does not consider the conformations of the neighbors. Table 5 shows the basin probabilities for the 20 amino acids, computed independently from the neighbors' identities. Most residues have a higher probability of being in the A basin, except for Val, Ile, and Cys, which have a dominant probability of being in the B basin.

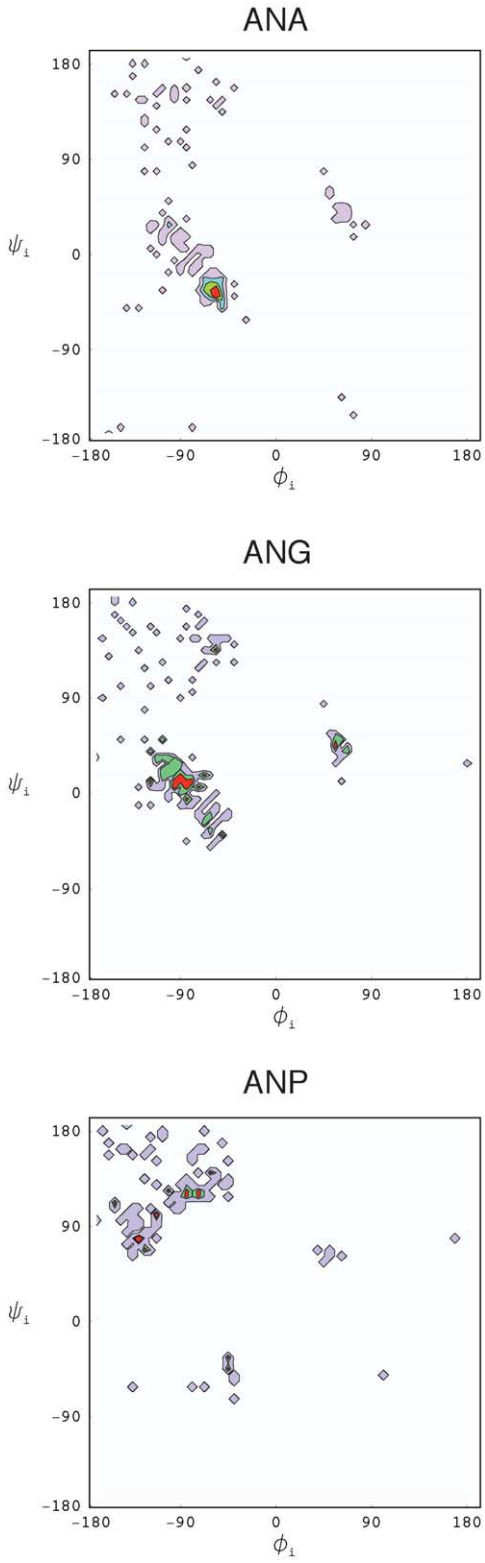

Figure 3. ϕ, ψ free energy plot examples for triplets composed of Ala and Asn, followed by different residues. The residues are labeled as $i-1$, i, $i+1$ and the angles belong to the center residue i.

Plots with the distribution of probability differences for basins A, B, and C1 are shown in Figure 2(a)–(c), respectively. Only triplets with 20 events or more were used in this calculation. In addition, proline residues occupying the triplet center were excluded. Figure 2(a) shows a significant deviation, especially below zero. The distribution has a long tail consisting of triplets with reduced probability. Many of these triplets are residues followed by a proline. It is well known that proline residues affect the conformation of their adjacent residue on their amino-terminal side.[10,11] As the data show, they reduce the probability of being in basin A. Many other triplets in this tail contain Val and Ile residues. On the other side of the distribution, many triplets are surrounded by Ala and Leu, but no long tail is observed. No correlation between the probability differences and the number of events for each triplet is found, showing that the deviations from zero are not due to low data statistics (except below 20 events). The distribution for the B basin (Figure 2(b)) shows almost the opposite effect as the one for the A basin probability difference. Figure 2(c), shows the changes in probability of basin C1. The distribution in this case shows a smaller dependence of the probability on the adjacent residues. In fact, the distribution between -0.025 and 0.025 shows a correlation to the number of triplet data points, indicating that the dispersion of probability differences in this range is dominated by statistical errors. Because the C1 basin is much less stable for most residues with the exception of Gly, it is an indication that this conformation is likely to be determined by global interactions. The long tails of this distribution consist mostly of glycine residues, followed in abundance by His, Asn, and Asp, which also contain significant populations in the C1 region.

Figure 3 shows examples of the ϕ, ψ dihedral angle free energy plots for some triplets. The plots show asparagine preceded by alanine and followed by alanine, glycine, and proline, respectively. When followed by alanine, the triplet is most likely to be in

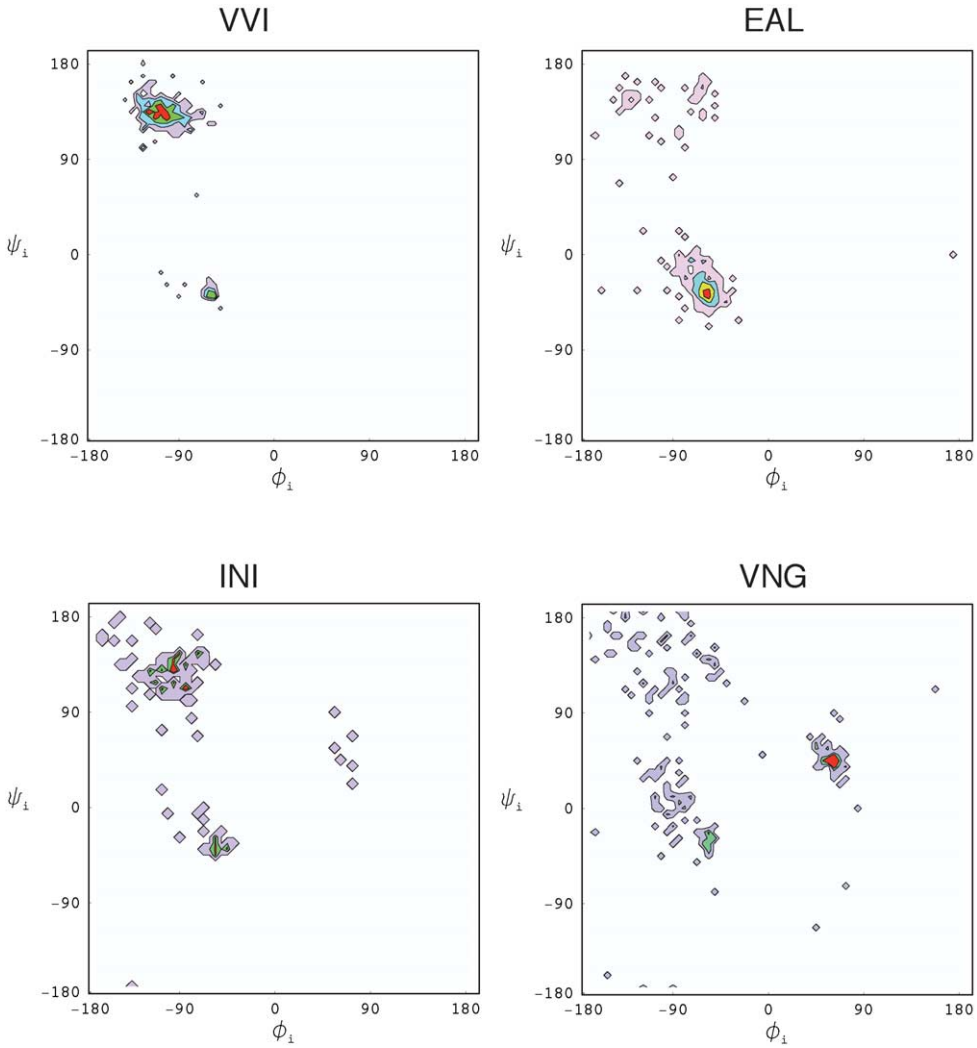

Figure 4. ϕ, ψ free energy plot examples for various triplets. The VVI plot shows a B basin dominated free energy, while the EAL shows an A basin dominated free energy. The bottom plots show other examples of triplets with Asn in the center, one (not followed by a Pro) in which the B basin dominates (INI) and the other in which the C1 basin dominates (VNG).

the A basin. With glycine, the plot shows two noticeable differences. First, the probability of being in the C1 region is slightly larger. Second, the peak in the A basin has shifted from A1 to A2. Because A1 is a typical α-helical region and A2 more common in loops, we should expect the triplet ANG to be found in or near loops most of the time. When asparagine is followed by a proline, the population shifts almost entirely to the B basin.

Figure 4 shows other examples. The top two cases are strong deviations from the single residue case for Val and Ala. VVI shows a high probability of being in the B1 region, typical of β structures. Individually, valine and isoleucine residues have a slight preference to be in the B basin, but when they are together the preference is much stronger showing a cooperative behavior. Conversely, EAL is most likely (88% of the time) to be in the A1 (α-helical) conformation. As expected, Ala, Glu, and Leu are among the residues with the highest probability of being in the A basin (64%, 66%, and 59%, respectively). The lower two cases are some extreme cases for Asn. FNI is a case (not involving Pro) with significant lower probability for the A basin than for the B basin. The probability in A is reduced from 49% for asparagine in general to 21% for INI. The VNG triplet shows a high tendency of being in the C1 basin. With the exception of Gly, Asn is the residue with the highest probability of being in the C1 basin (11%). But in VNG, the probability increases to 29%.

Dependency of dihedral angles on neighbor residue conformations

The dependence of an amino acid conformation on the conformations of its adjacent residues involves too many variables to be captured in a single probability density function for the available data. Instead, we divided the probability densities in individual terms involving pairs of angles. In particular, for a triplet with indices $i-1$, i, $i+1$, we looked at the density plots involving (ϕ_{i-1}, ϕ_i), (ψ_{i-1}, ϕ_i), (ψ_{i-1}, ψ_i), (ϕ_i, ϕ_{i+1}), (ψ_i, ϕ_{i+1}), and (ψ_i, ψ_{i+1}). As before, if the residue is Pro, ϕ is replaced by ω. The only difference between the $i-1$, i and the i, $i+1$ cases is that in the former the probability density depends on the composition of residue $i+1$ and in the latter on the composition of residue $i-1$. Using two angle distributions makes the approximation of the same order as the ϕ_i, ψ_i probability density, which is important in the combination of the free energy terms.

To illustrate the basins occupied by these angles, we show in Figure 5(a)–(c), the free energy of the combined densities for all amino acid pairs (not involving proline). For proline, three more cases (not shown) need to be considered, (ω_i, ϕ_{i+1}), (ϕ_i, ω_{i+1}), (ψ_i, ω_{i+1}). The plots show the correlation between the conformations of consecutive residues. The (ψ_i, ψ_{i+1}) case clearly distinguishes between basins A and B, and gives a higher probability for two consecutive residues to be in the same basin. In

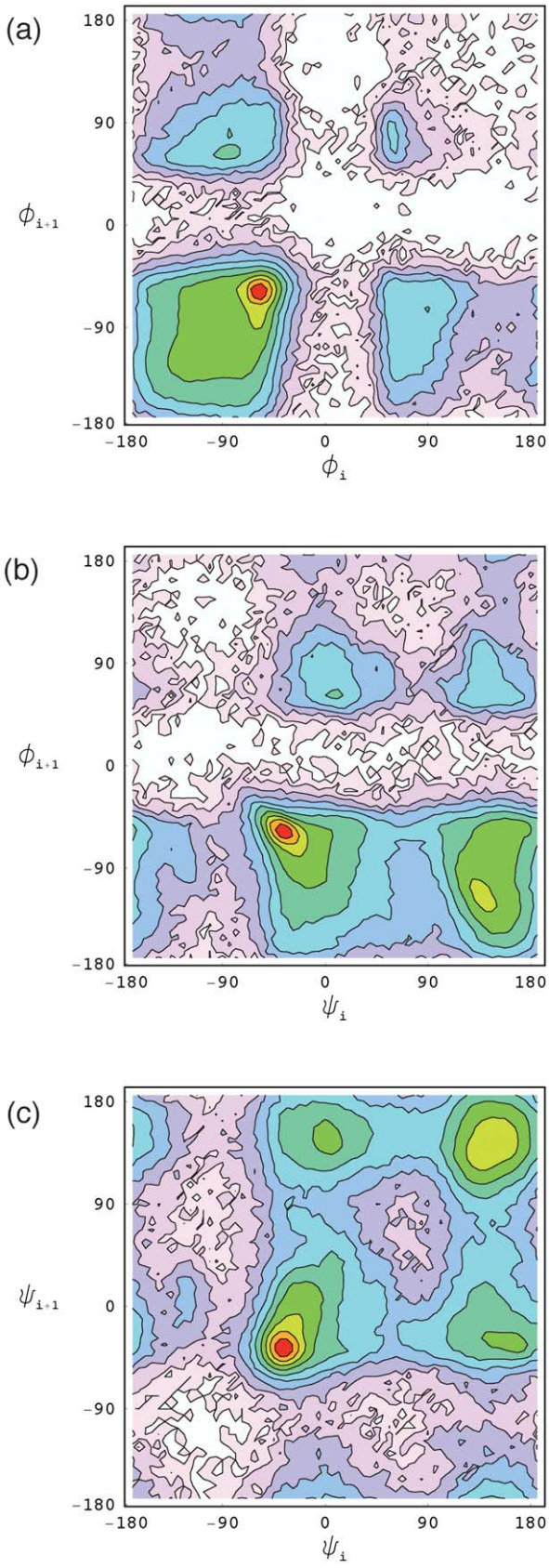

Figure 5. Free energies for dihedral angle pairs between consecutive residues. The angles are indicated in the axes. All the amino acid residues are included in the probability densities except proline. Only residue pairs are used in this illustration, therefore the i, $i+1$ cases are equal to the $i-1$, i cases.

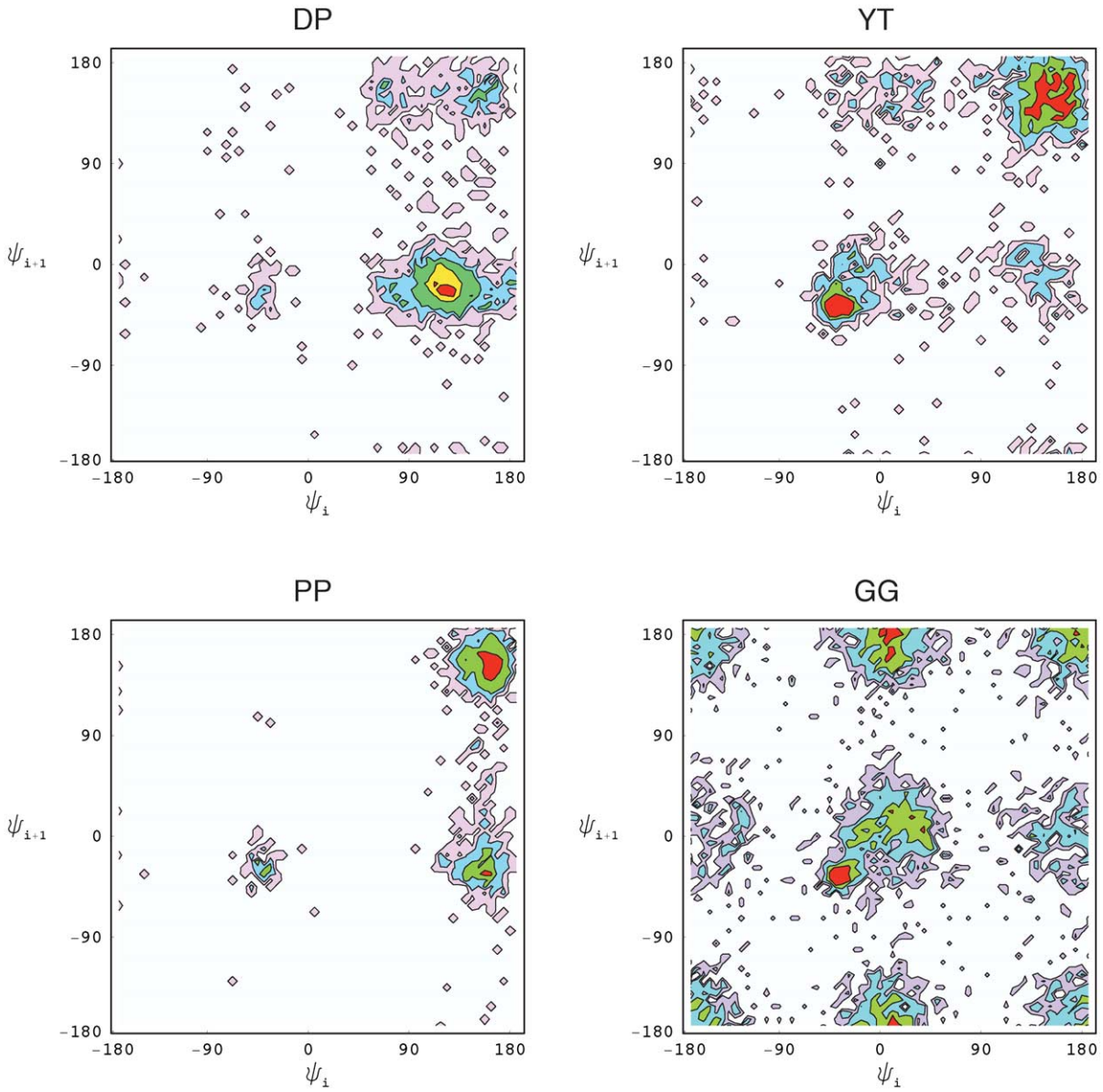

Figure 6. Free energy examples for the dihedral angles ψ_i, ψ_{i+1} between residue pairs. The DP example shows a strong asymmetry. The two deepest minima in YT shows a preference for the pair to be either in a coiled or in an extended conformation. The PP example also shows an asymmetric behavior in spite of being the same residue. In the GG example there is a slight asymmetry opposite to the PP case.

the (ϕ_i, ϕ_{i+1}) case, a distinction is made between the A+B and the C basins. The (ψ_i, ϕ_{i+1}) plot shows some correlations but not as strong as the (ϕ_i, ψ_i) plots.

When residue identities are considered, strong deviations from the general probability densities can be observed. For simplicity, in the following examples we only consider the (ψ_i, ψ_{i+1}) two residue correlations. Figure 6 shows the (ψ_i, ψ_{i+1}) probability densities for four cases: DP, YT, PP, and GG. The DP (Asp, Pro) case shows a clear asymmetry. Evidently, Asp has a higher probability of being in the B basin ($\psi_i > 50$), while Pro prefers to be in the D2 ($\psi_{i+1} < 0$) basin. In the YT example, (Tyr, Thr), the conformations are mostly found with both residues either in the A basin or the B basin, with a slight preference for the B basin. These results vary when the identity of residue $i-1$ is considered. The PP (Pro, Pro) case shows a strong asymmetry. This means, for example that when two proline residues are found together in different basins, the first one is more likely to be in either the D1 or E1 basin (extended) and the second one in the D2 or E2 (coiled) basin. Otherwise, both prefer to be in the D1 or E1 basins. In the GG (Gly, Gly) case, an asymmetry is also noticed, although not as pronounced. This time, the first glycine is more likely to be in the A or C1 basins while the second one

prefers to be in the B basin. Other residue pairs of equal types show more symmetric behavior.

The local free energy and its influence in the native dihedral angles

We now formulate the dihedral angle potential and investigate to what extent it determines a residue native conformation. In the free energy examples shown by us, it is evident that some residues have strong conformational preferences that depend on their nearest neighbors. Therefore, we should expect that a significant fraction of the native backbone conformation is determined locally. This hypothesis is supported by results obtained from the application of information theory to the local sequence-to-backbone structure relationship.[9]

Given a protein sequence, a dihedral angle statistical potential can be defined as:

$$V_{\text{tot}} = \sum_i \{V_{0;\alpha\beta\gamma}(\phi_i, \psi_i) + [V_{1;\alpha\beta\gamma}(\phi_{i-1}, \phi_i) + V_{2;\alpha\beta\gamma}(\psi_{i-1}, \phi_i) + V_{3;\alpha\beta\gamma}(\psi_{i-1}, \psi_i) + V_{4;\alpha\beta\gamma}(\phi_i, \phi_{i+1}) + V_{5;\alpha\beta\gamma}(\psi_i, \phi_{i+1}) + V_{6;\alpha\beta\gamma}(\psi_i, \psi_{i+1})]/2\} \quad (2)$$

where $\alpha\beta\gamma$ represents a sequence of three amino acid residues and each V in the sum is of the form of equation (1) with the appropriate angle dependence. The factor of 2 was introduced because there are two terms for each angle pair, except for V_0. At the protein ends, the V_0 term is substituted by a potential derived from the single residue probability for the given amino acid, and the other terms are derived from the probabilities between two residues. When Pro is present, the corresponding probabilities are replaced by probabilities depending on ω.

To test the role of the local free energy on the native conformations, it is more convenient at first to use a coarse (low-resolution) description of the residue conformations, which can be accomplished by mapping the dihedral angles to the basins described in the ϕ, ψ plots. Similar basins for the dihedral angle combination between neighboring residues can be defined consistent with the ϕ, ψ basins. The total number of events in each basin is used to compute N in equation (1), which in turn is used to build the low resolution version of equation (2). Only the two-dimensional triplet histograms with $N \geq 20$ are used, otherwise the two residue dependent histograms are used. All histograms are renormalized so that the total number of events per histogram is equal to the average (94).

A stable conformational state for a sequence is obtained by minimizing the low-resolution potential with respect to the basins. If only the V_0 terms are considered, minimizing V_{tot} is trivially achieved by minimizing each individual term independently. When the other terms are included, the equation has to be globally minimized. The algorithm to accomplish this is the following:

(1) For each residue $i=1,\ldots,N$ of a sequence of length N, we select a window of consecutive residues of size w_i centered around i. w_i is the same for all i (except at the sequence ends) and it is initially set to one.
(2) Equation (2) is minimized for each window independently of the residues outside the window by searching all possible (low resolution) conformations for the residues in the window. Only the conformation of the center residue i for each window is updated according to the minimum energy conformation of the windows.
(3) Step 2 is repeated by increasing the window size and the process ends when the total energy converges to a minimum, i.e. when the total remains the same after the window size has been increased several (three) times.

This algorithm converges quickly to a very low (but not necessarily the lowest) energy state. Typically, window sizes of no more than seven residues are needed.

The conformations obtained by minimizing the low-resolution potential can be compared to those extracted from the conformations of known protein structures. We arbitrarily selected a test set of non-homologous proteins consisting of a total of 125 chains ranging from 36 to 174 residues. For testing purposes, the dihedral angle potentials were re-derived by eliminating any one of the test proteins appearing in the protein set used to construct the probability distributions. Because of this and the computational requirements in the minimization, the test set was kept relatively small. For each chain, the conformation obtained from the minimization was compared to the native one. For all residues except for Gly and Pro, four states were considered: A, B, C1, and neither. For Gly the C2 state was also considered and for Pro the corresponding states were used. The fraction of identical states was used as a similarity measure.

The results show that around 62% of the residues minimize to the same state as the native one. This can be compared to 60% if only the V_0 term is used, and 55% if V_0 is independent of the adjacent residue identities. Figure 7(a) shows the distribution of the fraction of correctly assigned residues for the test chains. Many of the cases with the lowest fractions correspond to β-rich structures. Conversely, many of the cases with the highest fractions correspond to α-rich structures, although a significant amount of β structure can be found.

To better understand the nature of the preceding result, we can divide the residues according to their native secondary structure, which can be determined by using the DSSP algorithm.[12] For simplicity, we adopt the three code description for the secondary structure, i.e. α, β, and loops for everything else. The DSSP code is reduced to a three

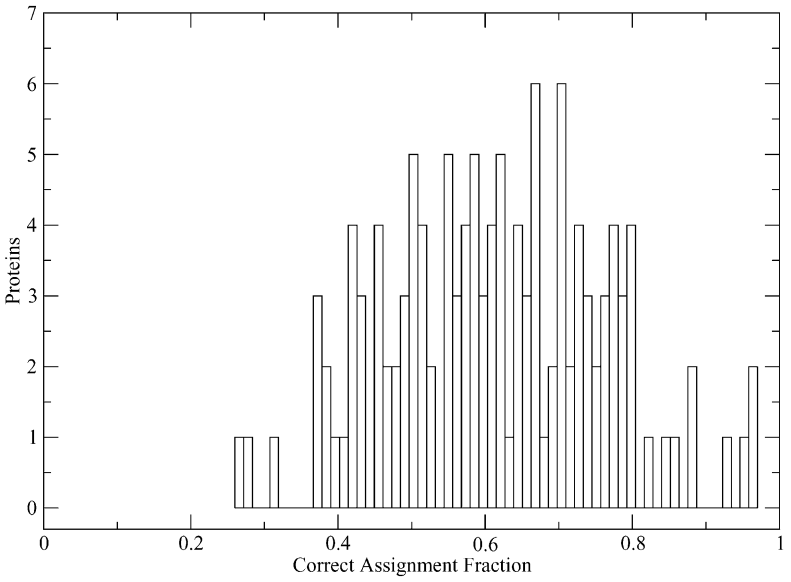

Figure 7. Results of minimizing the low resolution dihedral angle potential for 125 test proteins. The correct assignment fraction is the fraction of residues correctly assigned to their respective basins for each protein.

secondary structure code as follows: {H,G,I} → α; {E}→β; and {B,T,S,_}→L. In terms of this classification, the fraction of residues that adopt a native conformation after minimizing the dihedral angle potential is α: 0.89, β: 0.47, and L: 0.48. Clearly, the conformations of α-helices are strongly determined by the local interactions. The local potential influence in β-sheets is much less than for α-helices, but nevertheless significant. Notice that if the backbone potential simply classifies residues to be in helix and extended basins at random, then if the probability to be in the helix basin is 0.89, we would expect the probability to find a residue in a β structure to be no more than 0.11. Instead, the average probability is nearly four times larger. The fraction obtained for the loop residues is similar to that of residues in β-sheets. Nevertheless, this fraction is somewhat smaller than expected, considering that 43% of the residues in loops are in coiled and 49% in extended states. Therefore, the influence of non-local interaction may play a more important role in loops than in β-sheet residues.

The correlations between conformations of adjacent residues suggest that the local energy influence of residues in α-helices and β-sheets may depend on how deep the residue is buried in the structural element. To this end, we define the residue secondary structure depth (RSSD) as one plus the number of consecutive residue pairs on both sides of a residue in the same secondary structure element as the residue. The results of the backbone energy minimization for each secondary structure type can be divided as a function of the RSSD. The results are plotted in Figure 8. The error bars represent the inverse of the square root of the number of cases. For all three secondary structure types, there is a clear dependence on the RSSD. For α-helices, the fraction grows from 85% to near 92%

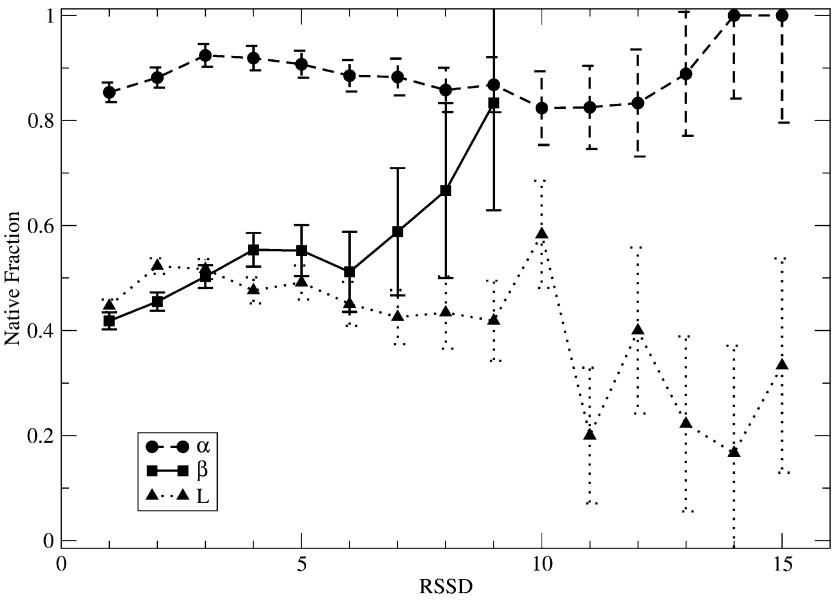

Figure 8. Dihedral angle potential influence in the residue conformation as a function of residue secondary structure depth. The statistics were obtained from 125 test proteins after minimizing the low-resolution dihedral angle potential. The secondary structures are divided into α-helices, β-sheets, and loops (L), according to the protein native structure. The error bars are indicative of the number of data points used in each average point and are calculated as the inverse of the square root of the number of points. Only average points with more than six data points are shown.

from one to three residues deep. Then it gradually decreases to 82% around ten residues deep and increases to 100% soon after. The rapid increase in native fraction near the edge of an α-helix suggests a cooperative behavior that favors their formation. It is not clear why this fraction decreases further into the structure. For β-sheets, it grows from 41% to near 83% after a RSSD of 9. The cooperative effect around the edge of the structure (from one to three residues in depth) is also evident. The loops, on the other hand, show an increase from one to two residues in depth, but a decrease to values below 40% as the RSSD increases further.

Backbone dihedral potential for folding

The dihedral angle probability distributions contain more information than the one provided by coarse potentials described above. Cases such as those in Figure 3 show the advantage of refining the probability distribution beyond a few broad domains. Higher resolution knowledge based potentials (KBPs) can be constructed from these probabilities, which in combination with non-local interactions, can be used in reduced protein model simulations. In this section, we construct continuous functions that broadly approximate the KBPs derived from the triplet probability distributions. This is convenient because of the low data count and higher noise to signal ratio in triplet histograms. In addition, continuous function fits are more appropriate in some minimizations and simulation procedures.

The general ϕ, ψ effective free energy (Figure 1(a)) can be roughly approximated by a sum of arbitrary periodic functions localized at the labeled regions. One possibility is:

$$F = \sum_{i}^{L} A_i \left[\cos^{2\lambda_i}\left(\frac{\phi+\psi}{4} - \theta_i\right) \cos^{2\mu_i}\left(\frac{\phi-\psi}{4} - \eta_i\right) + \sin^{2\lambda_i}\left(\frac{\phi+\psi}{4} - \theta_i\right) \sin^{2\mu_i}\left(\frac{\phi-\psi}{4} - \eta_i\right) \right] \quad (3)$$

This function represents a series of L localized functions whose positions are related to θ_i and η_i, their amplitudes to A_i, and their widths to λ_i and μ_i. For Pro, the function is:

$$F = \sum_{i}^{L} A_i \cos^{2\lambda_i}\left(\frac{\omega}{2} - \theta_i\right) \cos^{2\mu_i}\left(\frac{\psi}{2} - \eta_i\right) \quad (4)$$

Equation (3) is used to fit the ϕ, ψ energy plots of each individual residue obtained independently of the adjacent residue identities. Initial guesses are made for the 20 amino acids and then are fitted using the conjugate gradient method by minimizing an error function between equation (3) and the extracted free energies.

Fitting equation (3) to the triplet free energies is complicated by the higher noise to signal ratio. Several steps are taken to improve the estimates of the probability density and to carry out unsupervised fits over the 8000 sets. First, the probability densities are estimated for plots with 20 points or more; otherwise the single residue (neighbor independent) density of the triplet's center residue is used. To estimate the density of sparse data, selecting a constant grid size can generate large errors. Instead, the densities are estimated by using a modification of a recursive density estimation algorithm.[13] This algorithm selects the space partition tailored to the local density, rather than using a uniform partition. A brief description of the algorithm is presented in Materials and Methods. The estimated densities are transformed into free energies and fitted using equations (3) or (4). The initial values are selected to be those resulting from fitting the single residue free energies. This time, only the amplitudes A_i are fitted, keeping the other parameters constant. In this way, the number of minima remains the same, unless an amplitude A_i vanishes.

The terms involving correlations of the angles between two residues (V_1,…,V_6) are fitted in a similar way. This time, all the free energies are fitted by equation (4). Nine terms ($L=9$) are used when only the ϕ, ψ angles are involved. This number is different if one of the angles is ω. Initially, the three free energies in Figure 5(a)–(c) are fitted using estimates for the initial parameters. In addition, the ϕ_i, ω_{i+1}, ψ_i, ω_{i+1}, ω_i, ϕ_{i+1}, and ω_i, ω_{i+1} cases, involving Pro, are fitted separately. The parameters of these functions are used as initial conditions to fit the corresponding free energies for all residue pairs. That is, the free energies involving the three dihedral angle pairs (replacing ϕ by ω when appropriate) for each one of the 400 residue pair combinations are fitted. Finally, these fitted parameters are used as initial conditions for fitting the amplitudes A_i of the V_1,…,V_6 potentials for the 8000 triplet combinations (containing at least 20 points).

To evaluate the potential, we perform Monte Carlo simulations on a simplified protein model whose details are described elsewhere[14] (unpublished results). This model uses a simplified representation of the side-chain, while keeping a detail description of the backbone. However, for the present simulations, only the backbone dihedral-angle potential is used and the side-chains are irrelevant. The Monte Carlo algorithm works by generating window moves for a group of consecutive residues. These moves efficiently change the values of the flexible dihedral angles while keeping the bond angle and distances constant or within their natural range.

The test set of 125 sequences was used in the energy minimization. As described earlier, the potential was re-derived without the proteins appearing in the test set. Random initial conformations were generated and a finial structure was obtained by minimizing the backbone potential. The minimization was done for 40 independent annealing simulations. For each residue, the dihedral angles between the native and minimized

structures were compared. If the difference between any of the residues' dihedral angles between the minimized and the native structures was more than 90°, the residue was classified as non-native.

The total fraction of residues in the native conformation, separated by native secondary structure type is α: 0.96, β: 0.18, and *L*: 0.44. The native fraction is 0.55 for all residues combined. This high resolution potential gives a larger fraction of residues in helical conformation than the low resolution potential. It comes at the cost of a lower fraction for residues in β structures. In spite of being small, it is still more significant that an α-dominated random assignment (i.e. 96% α and 4% β).

The difference between the high and low potential can be understood by noticing that the dihedral angles for helical structures are more concentrated around the typical α-helices values, making the free energy lower for them than for extended structures, whose dihedral angles are more disperse. This suggests that an equilibrium conformation at a simulation temperature higher than zero could increase the yield of β structures. Carrying out such a calculation at an arbitrary (but low) temperature, and averaging over ensembles of 40 independent minimizations for each of the 125 proteins, we obtained α: 0.82, β: 0.25, and *L*: 0.41. While the fraction in α structures have diminished, the β fraction evidently increased.

The conformational state of a residue is affected by that of its neighbors, which are competing to minimize the global energy. Locally, a residue can be in a higher energy conformation as a result of this competition. This local stability can be measured by computing the difference between the residue energy in the energy-minimized structure and the minimum energy among all other alternate conformations. Residues in alternate conformation are defined as differing from the minimum energy conformation by a dihedral distance $\Delta > 90°$, which is defined as:

$$\Delta^2 \equiv \sum_{i=1}^{3} (\tau_i - \lambda_i)^2 \qquad (5)$$

where τ_i and λ_i are the backbone dihedral angles for a residue in the native and minimum energy structure, respectively. Equation (5) measures the Euclidian distance between the angles of a residue, taking the periodicity into account. The particular choice of metric is not important for our purposes, as long as it gives a measure of distance. The alternate conformation corresponds to a local competing state and the energy difference corresponds to a local energy gap. The local energy gap for a residue can be plotted against Δ to determine the correlation between the local stability and the proximity to its native conformation.

In Figure 9, we have plotted the local energy gap as function of Δ for residues classified by their conformational basin in the minimum energy conformation and by their native secondary structure. Only the A and B basins were considered,

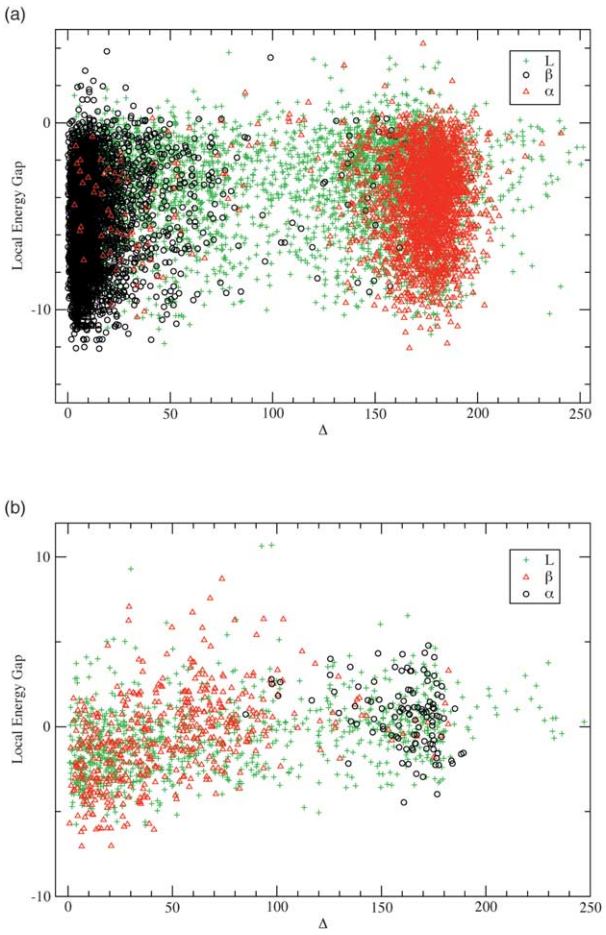

Figure 9. Local energy gap after the minimization of the high-resolution potential, as a function of the residue dihedral angle distance Δ for the 125 test proteins. The local energy gap is the difference between the energy of a residue after energy minimization and a competing conformation with the lowest energy outside a 90° range. The results are separated into two plots according to which basin the residues minimize into. The residues (plot points) are further classified according to their native secondary structure. (a) A basin residues. Residues that form α-helices have on average slightly lower energy gaps than residues that form β-sheets or loops. (b) B basin residues. Residues that form β-sheets or loops have on average significantly lower energy gaps than residues that form α-helices. In addition, the Δ for β-sheets generally increases as the energy gap increases.

corresponding to Figure 9(a) and (b), respectively. Note how the data for α-helices and β-sheets tend to cluster in two groups. In Figure 9(a), most of the residues that minimize to the A basin are also α-helices. On an average, β-sheet residues that minimize to the A basin have slightly higher energy gaps. Residues in loops either minimize to their native basin or an alternate one, judging from the clustering in groups with low and high Δ values. In either case, the energy is also slightly higher on average, as compared to the residues that form α-helices and β-sheets. In particular, the average local energy gaps and their standard deviations are

α: −5.18±2.7; β: −4.10±2.5; and Loop: −3.42±2.5. Figure 9(b) shows the results for residues that minimize to the B basin. This time, the residues that form β-sheets have significantly lower energy gaps, on average, than α-helical residues folding into the B basin. The local energy gaps in this case are α: 0.71±1.9; β: −0.69±2.6; and Loop: −0.72±2.1. The gaps are significantly smaller than for the A basin indicating less stable structures. Notice that a correlation between the energy gap and Δ can be observed for β-sheet residues. The dihedral angles increasingly deviate from their minimum energy values as they become locally unstable.

Curiously, Figure 9(a) shows some points that minimize to the A basin and form β-sheets but shows little change in Δ. This discrepancy is a result of the way the DSSP algorithm classifies residues into secondary structure classes. For example, residue 56 in chain A of the protein with PDB code 1nbc is classified by DSSP as β, in spite of having dihedral angles typical of α-helices ($\phi = -68°$, $\psi = -37°$). The structure shows this residue buried in the middle of a β-strand. In Figure 9(b) some residues that minimize to basin B and form β-sheets have a large Δ. For example, residue 13 of the protein with PDB code 1rip is classified as β in spite of having dihedral angles outside the B basin and closer to the C2 basin ($\phi = 74°$, $\psi = -102°$). The dihedral angles of this particular structure were not accurately determined experimentally. Overall, cases like these are rare and do not significantly change our results.

Discussion

We have investigated how the distribution of dihedral angles of residues in native structures are affected by their neighboring residue identities and conformations. Triplets of residues with all available combinations were examined by looking at the probability density plots between the φ, ψ angles for the central residue (Ramachandran plots) and for other combinations of dihedral angles between adjacent residues. For cases involving Pro, the ω angle replaced the φ angle. φ, ψ Plots revealed strong conformational dependencies on the adjacent residue identities and conformations for many triplet combinations. This agrees with the conclusions of Zaman *et al.*[6] for tripeptide simulations. The well-known effect of Pro on the conformation of preceding residues was clearly observed and quantified. Residues preceding Pro showed a high probability of being in a β-like state (the B basin). In addition, residues surrounded by amino acid residues with greater tendency to be in β-like conformations (such as Ile, Val, and Cys), have an increased tendency to be in β-like conformations. Similarly, residues surrounded by amino acid residues with greater tendency to be in α-like conformations (A basin), are likely to be in α-like conformations. The plots between adjacent residues, such as the ψ, ψ plots, showed cooperative behavior. Most amino acid residues have higher probabilities to be in similar conformations rather than in different ones. One common exception occurs when Pro follows some other amino acid residues. In this case, Pro is more likely to be in a β-like conformation while the preceding residue is more likely to be in an α-like conformation.

We have also explored the extent to which local angle propensities determine the conformation of dihedral angles in the native state. A low-resolution knowledge-based potential was derived using dihedral angle basins representing the most common secondary structure conformations such as the coiled and extended state. After minimizing this potential for a series of test proteins, we found that 55% of the residues are assigned to the native basin only when the identity of the central residue, but not that of the adjacent residues, is considered. If the identity of the adjacent residues is considered, the native assignments increase to 60%, and to 62% when the conformations of the adjacent residues are also considered. This number clearly shows a local energy influence on the dihedral angles and is better appreciated when the residues are divided according to their secondary structures in their native states. The local potential influence is larger for α-helices, in which 89% of the residues in native α-helical conformations minimize to the A basin. The influence in β-sheets is smaller (47% minimize to the B basin) but significant, given the high α-helical fraction. The loops (or random coils) yield 48% of residues in the native basin, which is similar to that of β-sheets. However, because loops can be either in coiled or extended conformation, 48% is less of what would be expected given the yield for the α-helices and β-sheets. We also found that the conformation of residues in α-helices or β-sheets is increasingly determined by the local potential as their location moves away from the edge and into the secondary structure element. The increase is more significant for β-sheets.

A higher resolution dihedral potential, derived by fitting periodic localized functions, was obtained and used to minimize the dihedral angles in off-lattice Monte Carlo simulations for a group of test proteins. This time, the simulations showed that the potential favors coiled conformations. The fractions of residues in the native basin after minimization are 96% for α-helices, 18% for β-sheets, and 44% for loops. The discrepancy between these results and the low-resolution ones arises because the dihedral angles of helical structures are more localized in the A basin that the ones of extended structures are in the B basin. This results in a lower free energy in the A basin for some triplets, even when the events in the B basin are higher. The larger entropy of the B basin indicates a higher flexibility for β structures, which could be preferred when non-local interactions are present. Further analysis of the energy of a residue relative to conformations with competing energies (energy gap) revealed that residues that minimize to the B basin generally become β-sheets if the gap is large and negative. This is also true for

α-helices but to a much lesser extent. In general, β-sheet residues that minimize to the B basin and α-helices that minimize to the A basin are more stable than when they minimize to opposite basins. Loops are generally more stable in coiled conformations. Overall, these results indicate that the dihedral angle potential is a significant contribution to the total folding potential and to the determination of the native dihedral angles.

The dihedral angle potential alone can give a good indication of the backbone conformational preference, but it is not intended to be a secondary structure predictor. For that purpose, secondary structure prediction (SSP) algorithms, such as PSIPRED,[15] can determine the state of a residue with an average success rate of nearly $Q_3 = 78\%$, using the three-type secondary structure description (α, β, Loop). Part of the success of the SSP methods is due to the intrinsic local propensity of the residues towards different secondary structure types, as indicated by our results. Another part is the non-local interactions implicitly included in multiple sequence alignment and neural networks method. Without the latter, Q_3 could drop below 70%. While 78% seems to be a significant amount, the structural information contained in a loop prediction is significantly less than that of an α or a β prediction. This is because the loop dihedral angles encompass a wide rage of values, including the extended or the coiled state ones. In our data, the distribution of native secondary structure is 33%, 24%, and 43% for α, β and loops, respectively. Applying the SSP method to our data would result in approximately 45% (57%×78%) of residues with determinable dihedral angles (α or β), which is comparable to what we obtained by minimizing the low-resolution dihedral angle potential for the α and β structures combined.

Materials and Methods

Protein set

The dihedral angle populations were computed from experimentally solved proteins structures deposited in the PDB. From 26,256 structure files, 54,860 polypeptide chains were extracted and clustered to eliminate highly redundant chains. In particular, clusters of chains differing by 95% sequence identity or more, normalized by the shortest chain, were obtained. A complete linkage hierarchical clustering was used, meaning that all sequences from a cluster were at least 95% identical with other members of the cluster and less than 95% identical with all other sequences outside the cluster. The FASTA sequence alignment algorithm was used for the sequence alignment.[16] The best structure from each cluster was selected satisfying the conditions that they should contain at least ten residues with all their heavy atoms, that the structure should not be obsolete, and that the resolution should be 3 Å or better. This process resulted in 7070 high quality chains.

For computing the dihedral angle distributions, it is important to use as many non-redundant high-quality sequences as possible. This is accomplished by further clustering the 7070 chains into groups differing by no more than 20% sequence identity, resulting in 703 clusters. Notice that 20% sequence identity nearly guarantees unrelated proteins although no requirement is imposed on the global structure similarity. If only one sequence from each cluster is used, the data would be insufficient to obtain good statistics. To circumvent this problem, for each of the 8000 triplets (and the seven pairs of dihedral angle combinations used), we use all 7070 chains with the requirement that triplet events come from different clusters differing by 20% identity or more and that the center residue of each triplet is at least two residues away from the chain ends. This allows us to find events for all triplets resulting in a total of 755,973 triplet events and an average of 94.5 events per triplet.

Probability estimation method

A probability density estimation algorithm is applied to the dihedral angle distributions of the residue triplets. It recursively divides the density space by 2^d, where $d = 2$ is the dimension of the space. In our version, the division is made by dividing each coordinate range in half. The four boxes created by this division are subdivided again. Then a χ-square test is made for the two division levels corresponding to a 20% confidence level. The χ-square conditions are:

$$\chi_3^2 = [(16/9)(1/N)\sum_{i=0}^{3}(a_i - N/4)^2] < 1.547 \qquad (6)$$

$$\chi_{15}^2 = [(256/225)(1/N)\sum_{i,j=0}^{3}(b_{ij} - N/16)^2] < 1.287 \qquad (7)$$

where N is the number of points in the box being divided, a_i is the number of points in the ith box of the first division level, and b_{ij} is the number of points in the ijth box of the second division level. If any of these inequalities is not satisfied, the density is assumed non-uniform, and the division procedure is repeated for each of the first level boxes. This process is iterated until no more substructures are found.

Acknowledgements

We thank Adrian K. Arakaki for stimulating discussions. This research was supported in part by NIH Grant GM-37408 of the Division of General Medical Sciences.

References

1. Creighton, T.E. (1996). *Proteins: Structures and Molecular Properties* (2nd edit.). W. H. Freeman: New York.
2. Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Advan. Protein Chem.* **23**, 283–438.
3. Hovmoller, S., Zhou, T. & Ohison, T. (2002). Conformations of amino acids in proteins. *Acta Crystallog. sect. D*, **58**, 768–776.
4. Perczel, A., Jakli, I. & Csizmadia, I. G. (2003). Intrinsically stable secondary structure elements of proteins: a comprehensive study of folding units of

proteins by computation and by analysis of data determined by X-ray crystallography. *Chem. Eur. J.* **9**, 5332–5342.
5. Pappu, R. V., Srinivasan, R. & Rose, G. D. (2000). The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl Acad. Sci. USA*, **97**, 12565–12570.
6. Zaman, M. H., Shen, M., Berry, R. S., Freed, K. F. & Sosnick, T. R. (2003). Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides. *J. Mol. Biol.* **331**, 693–711.
7. Ohkubo, Y. Z. & Brooks, C. L., III. (2003). Exploring Flory's isolated-pair hypothesis: statistical mechanics of helix-coil transitions in polyalanine and the c-peptide from rnase a. *Proc. Natl Acad. Sci. USA*, **100**, 13916–13921.
8. Karplus, P. A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* **5**, 1406–1420.
9. Solis, A. D. & Rackovsky, S. (2002). Optimally informative backbone structural properties in proteins. *Proteins: Struct. Funct. Genet.* **48**, 463–486.
10. MacArthur, M. W. & Thornton, J. M. (1991). Influence of proline residues on protein conformation. *J. Mol. Biol.* **218**, 397–412.
11. Hurley, J. H., Mason, D. A. & Matthews, B. W. (1992). Flexible-geometry conformational energy maps for the amino acid preceding a proline. *Biopolymers*, **32**, 1443–1446.
12. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
13. Fraser, A. M. (1989). Information and entropy in strange attractors. *IEEE Trans. Info. Theroy*, **IT-35**, 245–262.
14. Betancourt, M. B. (2003). A reduced protein model with accurate native-structure identification ability. *Proteins: Struct. Funct. Genet.* **53**, 889–907.
15. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
16. Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **183**, 2444–2448.

Edited by J. Thornton

(Received 23 January 2004; received in revised form 25 June 2004; accepted 28 June 2004)