Relationships Between Amino Acid Sequence and Backbone Torsion Angle Preferences

O. Keskin, D. Yuret, A. Gursoy, M. Turkay, and B. Erman*
College of Engineering and Center for Computational Biology and Bioinformatics, Koc University, Istanbul, Turkey

ABSTRACT **Statistical averages and correlations for backbone torsion angles of chymotrypsin inhibitor 2 are calculated by using the Rotational Isomeric States model of chain statistics. Statistical weights of torsional states of $\phi\psi$ pairs, needed for the statistics of the full chain, are obtained in two different ways: 1) by using knowledge-based pairwise dependent $\phi\psi$ energy maps from Protein Data Bank (PDB) and 2) by collecting torsion angle data from a large number of random coil configurations of an all-atom protein model with volume exclusion. Results obtained by using PDB data show strong correlations between adjacent torsion angle pairs belonging to both the same and different residues. These correlations favor the choice of the native-state torsion angles, and they are strongly context dependent, determined by the specific amino acid sequence of the protein. Excluded volume or steric clashes, only, do not introduce context-dependent $\phi\psi$ correlations into the chain that would affect the choice of native-state torsional angles. Proteins 2004; 00:000–000.** © 2004 Wiley-Liss, Inc.

Key words: Ramachandran map; torsion angle correlations; interresidue correlations; intraresidue correlations; knowledge-based potentials; triplet

INTRODUCTION

Backbone configurations of a protein are fully determined by the specification of its ϕ and ψ torsional angles. The native configuration is described by a unique set of the $\phi\psi$ angles. Different levels of correlations among the $\phi\psi$ angles are already identified and studied for native as well as denatured proteins. Ramachandran maps give the correlations among the ϕ and ψ angles of a residue resulting from exclusion of steric overlaps that hold both for denatured and native proteins.[1,2] More recently, improved versions of Ramachandran maps have been proposed and adopted for structure validation.[3–7] Recently, Pappu et al.[8] and Baldwin and Zimm[9] showed that long-range excluded volume interactions (i.e., steric clashes extending beyond the nearest neighbor residues) introduce strong correlations between the $\phi\psi$ angles of different residues and that the Flory isolated pair hypothesis does not hold in general. The Flory isolated pair hypothesis implies that for random configurations and in the absence of long-range interactions, correlations of ϕ_i and ψ_i within the ith residue do not extend to the angles of the neighboring residues, and each residue is statistically independent of others.[10,11] Correlations between residues are also observed in NMR measurements, which show that individual residues have distinct main-chain conformational preferences that are dependent both on the amino acid type and on neighboring residues in the sequence.[12–15]

Torsion angle correlations in native and denatured proteins are of significant importance because they contain the major source of information on the folding and stability of the protein. In the present article, we study correlations in ϕ and ψ angles, both inter- and intraresidue, by using two different approaches for obtaining torsional potentials. In the first approach, we use PDB-derived knowledge-based potentials to characterize the pairwise distributions of $\phi\psi$ angles for residues. The knowledge-based potentials are extracted as averages over the native states of large number, low homology proteins.

In the second approach, we generate a large number of random coil configurations of an all-atom protein chain model with volume exclusion by Monte Carlo and derive the $\phi\psi$ angle potentials for all residues of the protein chain. The Monte Carlo data yield excluded volume potentials, similar to Ramachandran plots, over the random conformations of the protein chains. Our aim here is to see whether there are significant correlations among $\phi\psi$ angles of the randomly coiled chain[13] and to see whether such correlations may serve as the determinant of the three-dimensional structure of the chain.[16,17] We then convert the torsion angle distributions obtained for the two cases into statistical weight matrices and use the rotational isomeric states model (RIS) of chain statistics[10,11] to calculate statistical averages and correlations for torsion angles over all configurations of the chain.

Our analysis of the results shows that 1) based on statistical weights derived from the Protein Data Bank (PDB), strong correlations exist, both inter- and intraresidue, between the torsion angles of the protein; 2) these correlations favor the native-state values of the torsion angles, 3) the correlations are context dependent (i.e., they are determined by the specific amino acid sequence of the protein); and 4) excluded volume or steric clashes, only, based on the all-atom excluded volume model, do not

*Correspondence to: Burak Erman, College of Engineering and Center for Computational Biology and Bioinformatics, Koc University, Sariyer 34450, Istanbul, Turkey. E-mail: berman@ku.edu.tr

Received 14 August 2003; Accepted 15 December 2003

Published online 00 Month 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20100

tapraid5/z7e-protein/z7e-protein/z7e00804/z7e1709d04g | sweigarm | S=6 | 3/2/04 | 5:15 | Art: 03512003

introduce correlations along the chain that enhance the choice of torsion angles of the native state.

MATERIALS, MODELS, AND METHODS

Knowledge-Based Potentials and A Priori Probabilities for the $\phi\psi$ Angles

Short-range interactions along with the long-range ones are responsible for the stabilization of the native structures in globular proteins. Short-range, or local, interactions refer to those taking place between near-neighbor amino acids along the main-chain. They result from both steric exclusions and pairwise attractive interactions of atoms within the same or neighboring residues. These energies favor certain values of the ϕ–ψ angles over others. These two contributions are also present to a large extent in the denatured state.[17] Long-range contributions result from interactions between residues that are farther apart along the chain or between a residue and the environment of the protein. These result mainly from 1) hydrogen bonding, 2) Lennard–Jones type and Van der Waals forces, 3) electrostatic interactions, and 4) disulfide bridges.

There have been a number of studies investigating the effect of both short- and long-range interactions in proteins.[18–26] Statistical treatments of protein conformations generally consider the interdependence of the backbone torsion angles adjacent to the peptide bond (Ramachandran plots) and neglect higher order interdependences between bond dihedral angles. Nishikawa and Matsuo[26] studied the conformational states and energies of tripeptides to develop a scoring function in fold prediction. Systematic analyses of torsional angles for different types of amino acids lead to efficient scoring functions for sequence alignments and fold predictions. In addition, regular secondary structure motifs, such as α-helices and β-strands, result from the repetition of well-defined rotations along the main chain.[1] These observations suggest that more precise preferences for particular secondary structures can be obtained by considering consecutive residues along the chain. The rapid increase in the number of PDB structures allows for a statistically reliable analysis of triplet conformations, incorporating the triplewise interdependence of adjacent amino acids.

In the present study, we investigate, as a specific example, the torsional bond angle correlations in the 64-residue protein chymotrypsin inhibitor 2 (CI2). Our analysis consists of two parts. In the first part, we partition the CI2 molecule of n residues (n = 64) into successive triplets, n−2 in number. We identify each triplet as XYZ, where X is the i−1st, Y is the ith, and Z is the i+1st residue along the sequence, respectively. Each residue has three torsional angles, ϕ_i, ψ_i, and ω_i. The torsional angle ω_i is fixed due to the partial double bond nature of the corresponding backbone bond. For each triplet of CI2, we derive PDB-based torsional potential energies for the $\phi_i\psi_i$ and $\psi_i\phi_{i+1}$ pairs. For this purpose, we use 1646 nonredundant PDB structures that are nonhomologous and representative of PDB structures.[27] This nonredundant list of proteins is retrieved from the PDBSelect. The resolution of the 1216 X-ray structures range between 0.54 and 3.00 Å and 430 of the structures are NMR solutions. (The list of PDB structures is provided as supplementary material). Residue-specific conformational potentials are developed on the basis of the probabilities of observations, $P_{XYZ}(\phi_i, \psi_i)$ and $P_{XYZ}(\psi_i, \phi_{i+1})$, of the configurations of each triplet XYZ. Here, $P_{XYZ}(\phi_i, \psi_i)$ is the probability of observing the middle residue (type Y) to be in state (ϕ_i, ψ_i) and $P_{XYZ}(\psi_i, \phi_{i+1})$ is the probability of observing residue type Y to be in state (ψ_i) and Z to be in state (ϕ_{i+1}). $P_{XYZ}(\phi_i, \psi_i)$ is a measure of the intraresidue correlation of torsional angles, and $P_{XYZ}(\psi_i, \phi_{i+1})$ is a measure of the interresidue correlation of two successive torsional angles. We use the discrete state formalism for the torsion angles. Each ϕ and ψ angle is divided into intervals of 30°; thus, 12 representative torsional states are assigned to each bond.[21] The statistics for the $\psi_i\phi_{i+1}$ and $\phi_i\psi_i$ pairs could as well be obtained from pairs of residues, YZ without incorporating the i−1th residue, X, along the sequence. This would improve the statistics in the nonredundant database significantly. However, energy maps obtained for doublets and triplets exhibit significant differences. Most importantly, energy maps based on doublets contain different numbers of minima than those based on triplets. We show in the final section that predictions based on doublets are not as accurate as those of triplets.

The evaluation of the probabilities $P_{XYZ}(\phi_i, \psi_i)$ and $P_{XYZ}(\psi_i, \phi_{i+1})$ is outlined in the Appendix (see Eq. A1). The probabilities calculated in this manner are the a priori probabilities, because the condition that the triplet under consideration is embedded into the given specific amino acid sequence has not been used.

The conformational energies $E_{XYZ}(\phi_i, \mu\psi_i)$ and $E_{XYZ}(\psi_i, \mu\phi_{i+1})$ are defined as a function of $P_{XYZ}(\phi_i, \psi_i)$ and $P_{XYZ}(\psi_i, \mu\phi_{i+1})$, respectively, using Eq. A2 in the Appendix.

In Figure 1(a), we show the contour torsional energy map of the $\phi\psi$ angle pair for the Ala depicted from all Ala residues in the database. In Figure 1(b), the energy map, obtained for the central residue as Ala in all the triplets of Glu-Ala-Gln is shown. A color range from black to white is used to show the minima and maxima of the figures, respectively. Comparison of the two figures clearly indicates the effects of first neighbors in the torsional potentials. The first plot is a classical Ramachandran plot for Ala. The usual peaks for α-helix, β-strand and L-α-helix are observed in the plot (darkest regions, as labeled in the figure). The second plot also displays the classical peaks for these three regions. However, there are significant additional peaks elsewhere in the figure. The extra peaks most probably do not show in the first graph, because in general, proteins are dominated by the regular secondary structures of α-helices and β-strands. So the regions for these structures are observed to be more populated when one considers all conformations of Ala. As seen from the second graph, Ala might prefer to be in different states other than the classical regions when it is in a Glu-Ala-Gln triplet; otherwise, these states would be averaged out. And most of the regions are almost unpopulated (white regions). This also shows that there are regions in space where Ala can never access in this specific triplet.

F1

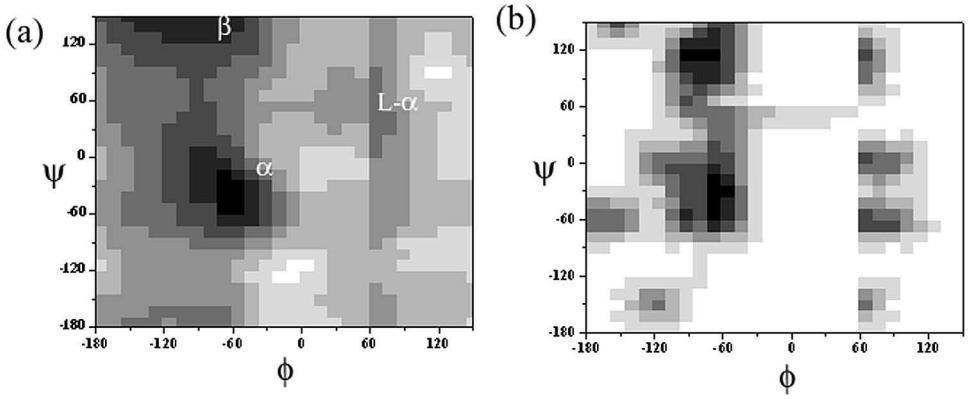

Fig. 1. Energy contour maps for the ϕψ torsional angles for alanine. The torsional angle space is divided into bins of 30°. Black regions correspond to the highly populated conformations, whereas white regions are the unfavored conformations. **a:** The PDB-derived Ramachandran map extracted from all available conformations of alanine residues. **b:** The same type of map derived for only Ala in triplets of Glu-Ala-Gln, Ala being confined between Glu and Gln.

The PDB knowledge-based potentials are derived from an ensemble of triplets belonging to 1646 proteins in their native states as explained above. The average occurrence of triplets, which we sampled from the PDB, is 110. The occurrence of the least observed triplet is 7, whereas the most frequent one is 268, and no discrimination is made for whether the triplet belongs to a helix, strand, or loop. For the large number of samples investigated, the various intermolecular interactions of the triplet leading to the observed potentials may be assumed to average out.[13,28] Therefore, they may be regarded as potentials of mean force operating in the native state and possibly in the denatured state.[29]

In addition to obtaining energy maps from PDB, we also investigate the effects of excluded volume and steric clashes on bond torsion angle energies by an all-atom model of CI2 where bond lengths and bond angles are fixed and torsional changes take place over all ϕ and ψ angle combinations. We generate random configurations of CI2 by Monte Carlo simulations. When steric clashes are excluded, the random configurations give data to obtain the Ramachandran plots. The present calculation considers the steric clashes between all pairs of atoms of CI2 and, therefore, gives a generalized version of the Ramachandran plots.

The generalized Ramachandran potentials based on the chain model with excluded volume were generated by running an all-atom simulation using random torsion angle moves. The simulation started at a random configuration of the CI2 molecule with no overlaps between atoms. The bond lengths and angles are fixed at their native state values. The contact radii used for the element types were H = 1.0, O = 1.3, N = 1.4, C = 1.5 Å. These radii correspond to ~0.75 of the minimum energy VDW radii used in CHARMM22. Two atoms were considered to be in collision if they are $<0.95*(r_1 + r_2)$ distance apart. Exceptions were made for potential hydrogen bonds, which allow O atoms to come closer than the contact distance to H and N atoms. The collision test was performed for atom pairs separated by more than four bonds.

At each step of the simulation, one random ϕ or ψ angle was rotated at random amount within [−.1, .1] radians. Moves that resulted in collisions were simply rejected. The positions reached were sampled at every 100,000 move attempts. A total of 1000 configurations are used for the calculation of potentials. The pairwise dependent ϕψ potentials were then obtained similar to those for the PDB-based potentials described above. The radii of gyration of the generated configurations are in the range of 200–600 Å, whereas the radius of gyration of the native CI2 is 125 Å. Therefore, generated random coil configurations are not compact denatured configurations close to those in the native state.

The RIS Model and Calculation of A Posteriori Probabilities

Here, we use the RIS formalism to obtain the statistics of the chain based on pairwise dependent torsional angles whose a priori derivation is outlined in the preceding section. The use of the RIS formalism for the 12-state model and the derivation of various bond angle probabilities given by Flory[10] may be briefly summarized as follows:

The torsional energy when bond i−1 is in state η and bond i is in state ζ is $E_{\eta\zeta,;i-1,i}$. Here, η and ζ both take 12 discrete values. The total ϕψ (torsional) energy E of the protein for a given configuration $\{\eta_1\zeta_1\eta_2\zeta_2\eta_3\zeta_3\ldots\}$ is

$$E = E_{12} + E_{23} + E_{34} + \cdots + E_{i-1,i} + E_{i,j+1} + \cdots + E_{2n-1,2n} \quad (1)$$

Here, $E_{i-1,i}$ is the energy corresponding to the joint occurrence of bonds i−1 and i in their respective states where the indices η and ζ indicating the corresponding states are omitted for simplicity. The statistical weight matrix $u_{\eta;\zeta i}$ for a given bond pair i−1 and i is evaluated as a function of the energy using Eq. A3 in the Appendix. The probability $p_{\zeta;i}$ that bond i will be in state ζ and the joint probability $p_{\eta\zeta i-1,i}$ that bond i−1 is in state η and bond i is in state ζ are calculated by using the partition function. (A

tapraid5/z7e-protein/z7e-protein/z7e00804/z7e1709d04g | sweigarm | S=6 | 3/2/04 | 5:15 | Art: 03512003

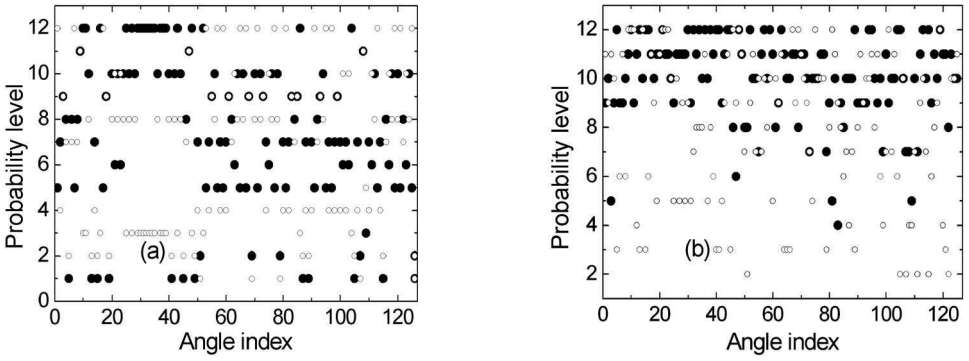

Fig. 2. **a,b:** A priori and a posteriori probability levels chosen by the bonds of the CI2, respectively. The *x* axes are the bond indices (because there are two rotatable bonds for each residue, the number is twice the residue number). The *y* axes represent the 12 probability levels for each of these bonds. Each bond is assigned to one of these levels. The filled and empty circles represent results from PDB database and the random coil chain with excluded volume model, respectively.

detailed description of the derivation of these probabilities is given in the Appendix.)

Correlations between state η of the ith bond and the state ζ of the jth bond are represented by the following relations

$$\begin{aligned} p_{\eta\zeta;i,j} &= p_{\eta;i}p_{\zeta;j} && \text{uncorrelated} \\ p_{\eta\zeta;i,j} &> p_{\eta;i}p_{\zeta;j} && \text{correlated} \\ p_{\eta\zeta;i,j} &< p_{\eta;i}p_{\zeta;j} && \text{anticorrelated} \end{aligned} \qquad (2)$$

Correlatedness/anticorrelatedness implies that the frequency of occurrence of bond i in state ζ and bond j in state η is enhanced/decreased relative to the frequency if these two bonds were independent. Thus, Eqs. A6, A7, and 2 are useful expressions for testing correlations between neighboring torsion angles of the chain. One may suitably define energies $\Delta E_{\eta\zeta;i,j}$ resulting from such correlations in excess of independent bond torsional energies according to

$$\Delta E_{\eta\zeta;ij} = -\ln\left[\frac{p_{\eta\zeta;i,j}}{p_{\eta;i}p_{\zeta;j}}\right] = -\ln\left[\frac{q_{\eta\zeta;i,j}}{p_{\zeta;j}}\right] \qquad (3)$$

The conditional probability $q_{\eta\zeta;i,j}$ in this equation is defined by Eq. A7. Correlations between the states η and ζ of bonds i and j, respectively, decrease the energy of the system, whereas anticorrelations increase the energy.

The probabilities given by Eqs. A5 and A6 are a posteriori probabilities, calculated in the presence of all the other residues of the given protein. Correlations based on the a posteriori probabilities for the $\psi_{i-1}\phi_i$ and $\phi_i\psi_i$ pairs, therefore, reflect the effect of the primary sequence on the $\phi\psi$ propensities. Our calculations reported below for CI2 show that the a priori probability for the occurrence of a given state of the ϕ and ψ pair obtained from the PDB is different from the a posteriori probability for the same state. The difference between the a posteriori and a priori probabilities reflects the effect of context dependence of the $\phi\psi$ propensities. Here, we show that context-dependent a posteriori probabilities extracted from the PDB database are better indicators of native $\phi\psi$ values than the context-independent a priori values.

Comparison of Calculated $\phi\psi$ Propensities With Native-State Values for CI2: Effects of Context Dependence

Equation A5 gives the a posteriori probabilities $p_{\zeta;i}$ of the 12 states of the ith bond. The subscript ζ identifies the state of the bond. We refer to these 12 values of $p_{\zeta;i}$ as the a posteriori "probability levels" of the ith bond. We order these probability levels from 1 to 12, the latter indicating the state with the highest probability. The torsion angle of the ith bond of the native CI2 corresponds to one of the calculated probability levels. If, for example, the torsion angle of the ith bond in native CI2 corresponds to the state with highest probability, we say that bond i has chosen the highest probability level (i.e., level 12) in the native state. The a priori probability levels are defined in a similar way by using Eq. A1. The probability levels for the angles ψ_i are calculated from the first of Eq. A1 by summing over all states of ϕ_i. Similarly, the levels for ϕ_i are calculated from the second of Eq. A1 by summing over all states of ψ_i. In Figure 2(a), the a priori probability levels chosen by the bonds of CI2 in the native state are shown by the filled circles as a function of bond index. F2

One sees from the figure that 20 native-state bonds have chosen the highest a priori probability level, and 12 bonds have chosen the lowest a priori probability level. If the a priori probabilities correctly reflected all realistic features of the native state, then the native structure would choose the highest probability level for the complete set of bonds. In the worst case scenario, if the a priori probabilities were based on random choice, the distribution of the filled circles in Figure 2(a) would be random with a mean probability level of 6.5 and a mean-square deviation of 3.45 about this mean. (The mean-square deviation value of 3.45 is calculated according to $\Sigma_1^m (p-\bar{p})^2/m$, where p is the randomly distributed probability level, $\bar{p} = 6.5$ is the average over the intervals (ranging between 1 and 12), and m is the number of points in the graph). The mean and the root-mean-square deviation (RMSD) for the points in Figure 2(a) are 7.48 and 3.34, respectively. In Figure 2(b), native torsion angle propensities are shown by the filled

tapraid5/z7e-protein/z7e-protein/z7e00804/z7e1709d04g | sweigarm | S=6 | 3/2/04 | 5:15 | Art: 03512003

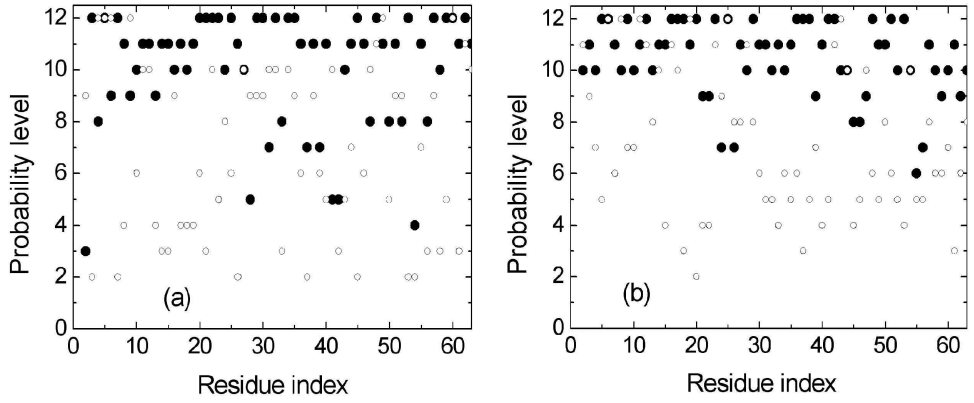

Fig. 3. **a,b:** Conditional probability levels chosen by the ith torsional bond of CI2, given that bond i−1 is in the native state. a: The intraresidue pairs $\phi_i\psi_i$. The filled circles show results for the PDB based potentials; empty circles are from our model with steric clashes only. Figure 3(b) is for the interresidue pairs $\psi_{i-1}\phi_i$.

circles, based on a posteriori probabilities calculated from Eq. A5, the use of which now incorporates the effect of context dependence. The chain in its native state now chooses the top probability level for 28 of its 128 bonds. None of the bonds chooses any of the lowest three levels. The mean and RMSD of the points are 10.1 and 1.74, respectively. There is a significant improvement in the mean and RMSD when a posteriori probabilities, and hence the effect of sequence, are considered.

The probability levels shown by the empty circles in Figures 2(a) and (b) are obtained from torsional energies for the random coil chain in the presence of steric clashes only. The distribution of the ordinate values of the points in Figures 2(a) and (b) is close to random. The mean and RMSD of the empty circles are 6.57 and 3.53 for the a priori and 7.88 and 3.11 for the a posteriori probabilities, indicating that the effect of excluded volume is not sufficient to improve torsion angle preferences for the native state.

The improvement introduced by chain averaging presented by the filled circles in Figure 2(b) results from correlations between angles of the neighboring residues. To see this effect clearly, we calculated the conditional probabilities according to Eq. A7. In Figures 3(a) and (b) we show the conditional probability levels chosen by the ith torsional bond of CI2, given that bond i−1 is in the native state. Figure 3(a) is for the intraresidue pairs $\phi_i\psi_i$. The filled circles show results for the PDB-based potentials; empty circles are from our model with steric clashes only. Figure 3(b) is for the interresidue pairs $\psi_{i-1}\phi_i$. The mean and RMSD of the filled circles in Figures 3(a) and (b) are (10.10, 2.24) and (10.22, 2.13), respectively. For the empty circles, these values are respectively (6.85, 3.33) and (7.16, 2.89). Thus, the PDB-based potentials show that if bond i−1 is already in the native state, there is a strong tendency for bond i to be also in the native state. For the model with steric clashes only, the preferences for such choices are insignificant and close to random.

Relationships among torsional states of two successive bonds may best be analyzed in terms of energy differences resulting from correlations using Eq. 3. In Figures 4(a) and (b), correlation energies are plotted for bonds i−1 and I, both of which are in the native state and have the highest probability level according to the PDB-based potential. Figure 4(a) is for $\phi_i\psi_i$ bonds, and Figure 4(b) is for $\psi_{i-1}\phi_i$ bonds. Dominant negative energies in both figures shown by the filled circles indicate that there are strong positive correlations among neighboring pairs when they are in the native state. Unfilled circles are results from our model with steric clashes only and indicate that excluded volume interactions alone do not enhance correlations that favor the native state for the torsion angles.

Results of calculations based on PDB energy maps for doublets are not as sensitive as those provided by the energies for triplets. The mean and the RMSD values of the torsion angle probability levels are (10.0, 3.91) compared to (10.1, 1.74) obtained by using the triplet energies. Here, the mean probability levels for the doublet and triplet energies are comparable, but the dispersion of the data points for the doublets is significantly higher. Similarly, the doublet-based statistics do not indicate the native angle preferences obtained by using the triplet energies that are presented in Figure 4.

CONCLUSION

Using pairwise dependent inter- and intraresidue torsional angle ($\phi\psi\mu$ energies derived from triplets from PDB and the RIS scheme of polymer statistics, we show that context dependence is significant in establishing the torsional bond angle preferences for the native state.

Calculations based on torsion angle data obtained from random configurations of the all-atom protein model with volume exclusion show that no correlations are introduced that would lead to preferences for the native values of the torsion angles. Therefore, one should search for different factors that cause $\phi\psi$ preferences in the random state. It was previously proposed that the amount of hydrophobic surface and hydrogen-bond formation with the solvent could be responsible for conformational preferences in the random states.[29] In addition, the intrinsic propensities for β-strand, α-helix, and preproline dihedral angles of the 20 amino acids in coil conformations indicate that the side-

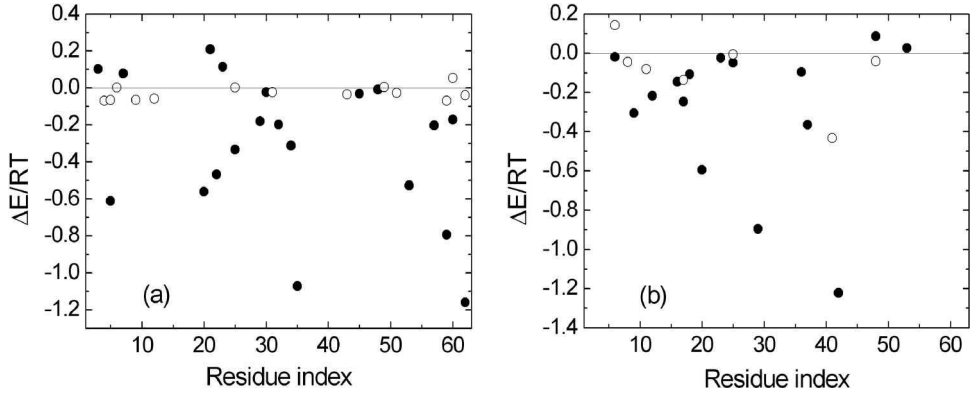

Fig. 4. **a,b:** Correlation energies plotted for bonds i−1 and i both of which are in the native state and have the highest probability level according to the PDB-based potential. a: $\phi_i\psi_i$ bonds. b: $\psi_{i-1}\phi_i$ bonds. Dominant negative energies in both figures shown by the filled circles indicate that there are strong positive correlations among neighboring pairs when they are in the native state. Unfilled circles are results from our model with steric clashes only and indicate that excluded volume interactions alone do not enhance correlations that favor the native state for the torsion angles.

chain of the amino acids determines the relative preferences for the $\phi\psi$ angles.[29]

Equation 3 systematically introduces an easy and a straightforward way of calculating energy differences due to correlations between the configurations of successive residues in a protein. Previous mutagenesis experiments have shown that there are significant energy differences in the helical and β-strand propensities of the 20 amino acids.[12,30–36] It has been shown that context plays a major role in determining these energies. We propose that Eq. 3 and the RIS scheme is a straightforward way of determining these energy differences, whether due to favorable or unfavorable interactions with the preceding and following residues within the given primary sequence.[29]

Some time ago Serrano[29] made the plausible argument that if the protein database is large enough including only proteins with a low degree of homology, the context effects may cancel out by averaging over many different environments and that the distribution of the dihedral angles for the different amino acids may follow the Boltzmann distribution.[34,37] The improvement in the $\phi\psi$ propensities—when the full protein sequence is considered in our RIS scheme—results from the context dependence of these variables by considering triplets. There have been several studies aimed at evaluating context dependence (see, e.g., Ref. 29 and the references given therein). Our analysis shows that the RIS scheme is the most straightforward method based on the evaluation of conformational probabilities of chain structures. In the present article, we have given only the general features of the RIS scheme with the CI2 sequence as the specific example. The same method of analysis may be extended to study several protein sequences to extract information on the conformational propensities of the 20 individual amino acids as they appear in α-helices, β-sheets, or any given specific subsequences.

We finally note that the RIS scheme describes the well-established fact that the local sequence-to-structure mapping is not one to one over all sequence space.[38] This finding may readily be observed from Eq. A5, which states that the probability of any state is influenced by the full sequence and not only by the local structure.

REFERENCES

1. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain conformations. J Mol Biol 1963;7:95–99.
2. Ramakrishnan C, Ramachandran GN. Stereochemical criteria for polypeptide chain conformations. II. Allowed conformations for a pair of peptide units. Biophys J 1965;5:909–933.
3. Walther D, Cohen FE. Conformational attractors on the Ramachandran map. Acta Crystallogr D 1999;D55:506–517.
4. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. Structure 1996;4:1395–1400.
5. Gunasekaran K, Ramakrishnan C, Balaram P. Disallowed Ramachandran conformations of amino acid residues in protein structures. J Mol Biol 1996;264:191–198.
6. Pal D, Chakrabarti P. On residues in the disallowed region of the Ramachandran map. Biopolymers 2002;63:195–206.
7. Lovell SC, Davis IW, Arendall WB, deBakker, PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. Proteins 2003;50:437–450
8. Pappu RV, Srivanasan R, Rose GD. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. Proc Natl Acad Sci USA 2000;97;12565–12570.
9. Baldwin RL, Zimm BH. Are denatured proteins ever random coils? Proc Natl Acad Sci USA 2000;97:12391–12392.
10. Flory PJ. Statistical mechanics of chain molecules. New York: Wiley; 1969.
11. Mattice WL, Suter UW. Conformational theory of large molecules. The rotational isomeric state model in macromolecular systems. New York: John Wiley & Sons, Inc.; 1994.
12. Smith CK, Withka JM, Regan LA. Thermodynamic scale for the β-sheet forming tendencies of the amino-acids. Biochemistry 1994;33:5510–5517.
13. Smith LJ, Fiebig KM, Schwalbe H, Dobson CM. The concept of a random coil. Residual structure in peptides and denatured proteins. Fold Design 1996a;1:R95–R106.
14. Smith LJ, Bolin KA, Schwalbe H, MacArthur MW, Thornton JM, Dobson CM. Analysis of main chain torsion angles in proteins. Prediction of NMR coupling constants for native and denatured conformations. J Mol Biol 1996b;255:494–506.
15. Penkett CJ, Redfield C, Dodd I, Hubbard J, McBay DL, Mossakowska DE, Smith RAG, Dobson CM, Smith LJ. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. J Mol Biol 1997;274:152–159.
16. Minton AP. Excluded volume as a determinant of macromolecular structure. Biopolymers 1981;20:2093–2120.

AQ: 1
17. Creighton TE. Proteins: structures and molecular properties, 2nd ed. New York: W.H. Freeman; 1993.
18. Sippl M. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
19. Go N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
20. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001; 44:223–232.
21. Bahar I, Kaplan M, Jernigan RL. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Proteins 1997:29:292–308.
22. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
23. Miyazawa S, Jernigan RL. Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. Proteins 1999;36:347–356.
24. Miyazawa S, Jernigan RL. Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space. Proteins 2003;50:35–43.
25. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258: 367–392.
26. Nishikawa K, Matsuo Y. Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. Protein Eng 1993;6:811–820.
27. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
28. Fiebig KM, Schwalbe H, Buck M, Smith LJ, Dobson CM. Towards a description of the conformation of denatured states of proteins. Comparison of a random coil model with NMR measurements. J Phys Chem 1996;100:2661–2666.
29. Serrano L. Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation. J Mol Biol 1995;254:322–333.
30. O'Neil K, DeGrado WA. Thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 1990;250:246–250.
31. Horovitz A, Matthews JM, Fersht AR. α-Helix stability in proteins. II. Factors that influence stability at an internal position. J Mol Biol 1992;227:560–568.
32. Blaber M, Zhang X, Matthews BW. Structural basis of amino acid a-helix propensity. Science 1993;260:1637–1640.
33. Chakrabartty A, Kortemme T, Baldwin RL. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing sidechain interactions. Protein Sci 1994;3:843–852.
34. Munoz V, Serrano L. Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. Proteins 1994;20:301–311.
35. Kim CA, Berg JM. Thermodynamic β-sheet propensities measured using a zinc-finger host peptide. Nature 1993;362:267–270.
36. Minor DL, Kim PS. Context is a major determinant of β-sheet propensity. Nature 1994;371:264–267.
37. Stites EW, Pranata J. Empirical evaluation of the influence of side-chains on the conformational entropy of the polypeptide backbone. Proteins 1995;22:132–140.
38. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci USA 1996;93:5814–5818.

Supplementary Material

Supplementary material used in this article can be found at the following Web page: http://ccbb.ku.edu.tr/pub/torsionanglecorrelations/

APPENDIX

The pairwise dependent probabilities $P_{XYZ}(\phi_i, \psi_i)$ and $P_{XYZ}(\psi_i, \mu\phi_{I+1})$ are calculated according to

$$P_{XYZ}(\phi_i, \psi_i) = N_{XYZ}(\phi_i, \psi_i)/\sum N_{XYZ}$$

$$P_{XYZ}(\psi_i, \phi_{i+1}) = N_{XYZ}(\psi_i, \phi_{i+1})/\sum N_{XYZ} \qquad \text{(A1)}$$

where $N_{XYZ}(\phi_i, \mu\ \psi_i)$ indicates the number of residue triplets observed having the indicated values of the argument. The definition for $N_{XYZ}(\psi_i, \mu\ \phi_{I+1})$ is similar. The term ΣN_{XYZ} in the denominator is the total number of observed triplets of XYZ in all possible states.

We define a conformational energy for a given residue Y (in the triplet XYZ) along the primary sequence of the protein as:

$$E_{XYZ}(\phi_i, \psi_i) = -RT \ln [P_{XYZ}(\phi_i, \psi_i)/P^0_{XYZ}(\phi_i)\, P^0_{XYZ}(\psi_i)]$$

$$E_{XYZ}(\psi_i, \phi_{i+1}) = -RT \ln [P_{XYZ}(\psi_i, \phi_{i+1})/P^0_{XYZ}(\psi_i)\, P^0_{XYZ}(\phi_{i+1})] \qquad \text{(A2)}$$

where $P^0_{XYZ}(\phi_i)$, $P^0_{XYZ}(\psi_i)$, and $P^0_{XYZ}(\phi_{I+1})$ are the uniform distribution probabilities (i.e., those valid when all angles are equally probable). In continuous space, they are equal to $1/2\pi$, and in the discrete state formalism, they are directly proportional[21] to the size of the angular intervals of the states (1/12).

The statistical weight matrix $u_{\eta\zeta;i}$ for a given bond pair $i-1$ and i is given as[10,11]

$$U_i = u_{\eta\zeta;i} = \exp\left(-\frac{E_{\eta\zeta;i-1,i}}{RT}\right) \qquad \text{(A3)}$$

Here the $\eta\zeta$'th element of U_i indicates the statistical weight for bond i when it is in state ζ, whereas the bond $i-1$ is in state η. The partition function, Z, of the chain is obtained according to

$$Z = J^* \left[\prod_{i=2}^{2n} U_i\right] J \qquad \text{(A4)}$$

where $J^* = [1\ 0 \ldots 0]$; J = column$[1\ 1 \ldots 1]$

The probability $p_{\zeta;i}$ that bond i will be in state ζ is given by

$$p_{\zeta;i} = Z^{-1} J^* \left[\prod_{m=2}^{i-1} U_m\right] U'_{\zeta;i} \left[\prod_{m=i+1}^{2n} U_m\right] J \qquad \text{(A5)}$$

Here, $U'_{\zeta;i}$ is the matrix obtained by equating the entries of all of its columns to zero except those of column ζ.

The joint probability $p_{\eta\zeta;i-1,i}$ that bond $i-1$ is in state η and bond i is in state ζ is given by

$$p_{\eta\zeta;i-1,i} = Z^{-1} J^* \left[\prod_{m=2}^{i-1} U_m\right] U'_{\eta\zeta;i} \left[\prod_{m=i+1}^{2n} U_m\right] J \qquad \text{(A6)}$$

The conditional probability $q_{\eta\zeta;i-1,i}$ that bond i will be in state ζ, given that bond $i-1$ is in state η is obtained from

$$q_{\eta\zeta;i-1,i} = \frac{p_{\eta\zeta;i-1,i}}{p_{\eta;i-1}} \qquad \text{(A7)}$$