Computational basis of knowledge-based conformational probabilities derived from local- and long-range interactions in proteins

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · OCTOBER 2006

Impact Factor: 2.63 · DOI: 10.1002/prot.21206 · Source: PubMed

Attila Gursoy
Koc University
120 PUBLICATIONS **4,244** CITATIONS

Burak Erman
Koc University
217 PUBLICATIONS **5,847** CITATIONS

Computational Basis of Knowledge-Based Conformational Probabilities Derived from Local- and Long-Range Interactions in Proteins

Lerzan Ormeci, Attila Gursoy, Guzin Tunca, and Burak Erman*
College of Engineering, Koc University, Rumelifeneri Yolu, 34450 Sariyer, Istanbul, Turkey

ABSTRACT **The probabilities of the various basins in Ramachandran maps are examined critically. The theoretical basis of probability calculations both from molecular computations and from protein libraries are discussed. The well-defined basins of the Ramachandran maps are treated as rotational isomeric states. Statistical independence and dependence of the states of different residues along the peptide chain are discussed. The Flory isolated pair hypothesis, near neighbor correlations, context effects, and long-range correlations are examined critically. A method of evaluating long-range correlations in helical and extended sequences is introduced in analogy with earlier polymer theory. Three different protein libraries are constructed where data is considered from residues in the (i) coiled regions, (ii) all regions, and (iii) only the helical and extended regions of proteins. Singlet and pairwise dependent probabilities calculated from these libraries are used to predict whether a given sequence is helical or extended. Predictions using pairwise dependence were not better than those using singlet probabilities. Modeling of long-range correlations improved the predictions significantly. Removal of the Chameleon sequences from the data set also improved the predictions, but to a lesser extent. Proteins 2007;66:29–40. © 2006 Wiley-Liss, Inc.**

Key words: rotational isomeric state; partition function; pairwise dependence; Flory isolated pair hypothesis; torsion energies; secondary structure propensities; chameleon sequences

INTRODUCTION

Conformational preferences of amino acids are suitably described by adopting the ϕ–ψ torsion angle representation, and the associated Ramachandran maps. The free energy surfaces constructed over these maps indicate well-defined basins. The occurrence of a residue in a given basin defines the state of that residue. Molecular calculations using suitable energy functions allow the determination of the probabilities associated with these states for different amino acids. Alternatively, knowledge-based approaches count the frequency of occurrence of residues in the various states using databanks of native proteins and determine the associated probabilities. The probabilities obtained by these two approaches differ, however, because the conditions on which they are based are different. We intend to discuss the statistical basis of probabilities obtained from protein libraries and the sources of differences when different subsets of protein libraries are used.

The use of Ramachandran plots to define the states of a given residue agrees with the rotational isomeric state formalism introduced by Volkenstein, Flory, and others.[1–3] The formalism defines the preferred torsion states of chain bonds, similar to the various basins of the Ramachandran maps and uses them to predict especially the spatial dimensions of synthetic, flexible-chain polymers. An advantage of the method is that the specific chemical structure of the chains can be incorporated into the formalism by specifying bond lengths, bond angles, side groups, and all interactions resulting from the interactions of these. Within this context, the formalism should also be useful for studying the dimensions of unfolded proteins, which have heterogeneous sequences and chemical structure is of importance.

The state of a residue in the absence of neighboring residues indicates the intrinsic propensity or the backbone preference of that residue to be in that state. When the residue is embedded in the polypeptide chain, its states may be correlated with those of the neighboring residues (local correlations) along the chain and those distant along the chain (long-range correlations). The Flory isolated residue pair hypothesis assumes that in the random conformational state two neighboring residues along the chain are statistically uncorrelated in the absence of long-range correlations.[1,4,5] This statement is based on the observation that if the chain is kept in its linear conformation and the ϕ_i, ψ_i and the ϕ_{i+1}, ψ_{i+1} pairs are varied over all allowable values given in the Ramachandran maps, no combination of these four rotations will bring the residue i into interaction with residue $i + 2$. If the rest of the chain is not fixed in

*Correspondence to: Burak Erman, College of Engineering, Koc University, Rumelifeneri Yolu, 34450 Sariyer, Istanbul, Turkey. E-mail: berman@ku.edu.tr

Received 17 May 2006; Revised 4 August 2006; Accepted 9 August 2006

Published online 12 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21206

its linear shape when the four bonds are being rotated as stated above, then residue $i + k$, for any $k > 2$, may interact with residue i. An interaction of this type is classified as a long-range interaction. Keeping the rest of the chain in its linear form corresponds to isolating the pair $i, i + 1$.

The energy surface for a single residue may suitably be calculated by adopting a methyl capped dipeptide by inserting the residue X into *N*-acetyl-*N*′-methylamide to form Ace-X-Nme.[6–10] Recent calculations and measurements of nmr coupling constants give insights into intrinsic backbone preferences in dipeptides and longer sequences.[10] Calculations on tripeptides, Ace-X_1-X_2-Nme, or longer sequences show that the Flory isolated pair hypothesis is not strictly true.[8,11,12] Deviations from isolated pair hypothesis are due to near neighbor (NN) effects. More specifically, the NN effect implies that the two sets of the angles ϕ_i, ψ_i and the ϕ_{i+1}, ψ_{i+1} cannot take values independently. Although the origin of the NN effect is not fully understood yet, the electrostatic screening model[13] can explain why the ϕ angles are shifted toward more negative values if the neighboring residues of a given residue X are aromatic or β-branched. An equally plausible alternative explanation is the formation of hydrogen bonds between side chains and the backbone for sequences longer than dipeptides.[14,15] The next interaction beyond NN is the set of long-range interactions. For shorter polypeptides, long-range effects are not consequential.

When probabilities are derived by the knowledge-based approach, several 'environmental' factors contribute to the configurational state of a residue. First, the neighbors of a residue along the chain exist at specific conformations in the native state. For example, a residue in a helical sequence sees a different neighborhood than if it is in a β strand. This effect is referred to as the 'context effect',[16] which may, however, average out if the database is large enough and all possible neighborhoods are available. Second, every protein in the database is in its native compact state, and long-range forces between residues that are spatially close but far apart along the chain contour are dominant. The differences between database statistics and molecular simulations have been addressed in several papers. Hermans and coworkers[9] compared the results of simulations and database statistics for five amino acids and discussed the sources of the differences between the two. The influence of the local amino acid sequence on ϕ–ψ probabilities were investigated by Garnier and coworkers.[17] The ψ angle probabilities estimated from a databank were shown to be context sensitive and position dependent.[18] Serrano used a coil database and identified the real intrinsic propensities independent of context effects.[16] Similarly, Thornton and collaborators determined the intrinsic ϕ–ψ properties of residues from a coil data bank.[19] Coil libraries are constructed from residues in the nonstructured regions of native proteins with the expectation that contributions from the near neighbor and resulting context effects are as small as possible.

Long-range effects, that is correlations between a residue i and $i + k$, $k > 2$, may average out when the statistics is made over a large dataset. However, since the nature of long-range effects are different in helices and β strands (i.e., a helical sequence makes hydrogen bonds within the sequence and a β strand is hydrogen bonded externally), the long-range interactions may persist in such special cases, as will be discussed on more detail below.

In the present article, we discuss the statistical mechanical features of configurational probabilities based on the rotational isomeric state formalism.[1,2] Using this formalism as reference, we discuss contributions from intrinsic properties, near neighbor and context dependence and long-range interactions. For this purpose, we construct three different database libraries: (i) the full nonredundant PDB set, (ii) the coil library, and (iii) the helix-extended library. The full library considers the conformations of residues irrespective of their secondary structure environment. The coil library consists of data from residues that are in the unstructured regions of proteins. The helix-extended library contains statistics from those residues that are in either a helix or an extended conformation.

As an exercise, using the probabilities obtained from these three libraries, we predict and compare the probabilities of given sequences to be in a helical or an extended configuration. At the end of the article, we introduce a statistical model that evaluates conditional conformational probabilities under long-range effects.

MATERIALS, MODELS, AND METHODS

Computing the Probabilities of a Given Sequence

We represent the state of a residue i by ζ_i, where the state is determined by the ϕ–ψ torsion angle pair for that residue. Figure 1 shows the regions that we define as states.[20] The 11 states (another term for states is 'basins' as used by Freed and coworkers[8]) indicated in this figure are as follows: α_R, α_L, right and left-handed α-helix regions; β_S, region largely involved in β sheet formation; β_P, extended polyproline-like helices; γ and γ', regions forming tight turns known as γ and inverse-γ turns; δ_R, right-handed region commonly referred to as the bridge region; δ_L mirror image of δ_R region; ε, region with $\phi > 0$, $\psi = 180$ that is predominantly observed for gly; ε', mirror image of the ε region; ζ is the region largely associated with residues preceding Pro.

The 11 states are separated by well-defined barriers. For this reason, it may be appropriate to define these as states in analogy to the well known 'rotational isomeric states' of polymer statistics, on which detailed statistical mechanical formulations were developed previously.[1,2]

The probability P of occurrence of a specific state $\zeta_1\zeta_2\ldots\zeta_i\ldots\zeta_n$ of a sequence of n residues in a given data set Ω is described in its full generality by $P = [(\alpha_1,\zeta_1)(\alpha_2,\zeta_2),(\alpha_3,\zeta_3)\ldots\ldots(\alpha_{i-1},\zeta_{i-1})(\alpha_i,\zeta_i)\ldots\ldots(\alpha_{n-1},\zeta_{n-1})(\alpha_n,\zeta_n)]$. Here, α_i is the ith residue type and ζ_i represents the state of the ith residue, which may be 1 of the 11 states shown in Figure 1. Presented in this manner, the probability function is the joint probability of $2n$ variables, n of which are residue type, represented by α_i, and the remaining n are the states ζ_i. Knowing this probability, one may assess, for example, the secondary structure that

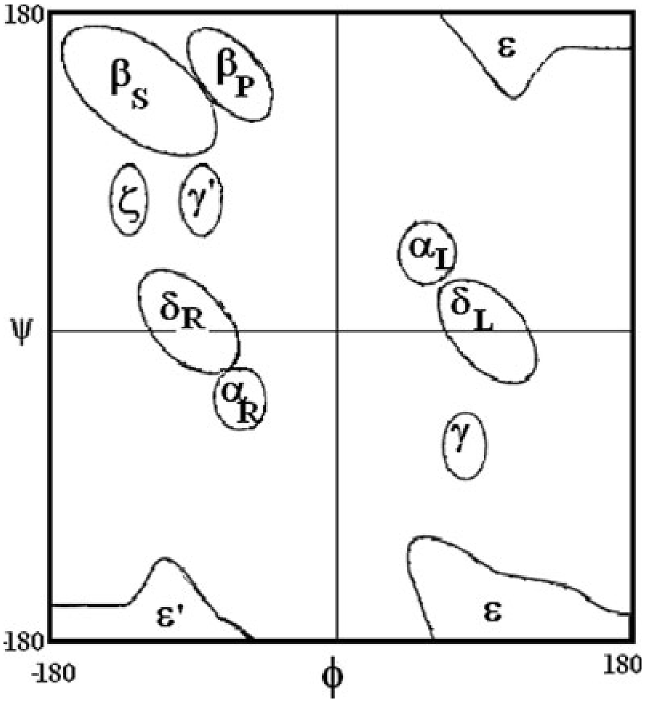

Fig. 1. Allowed states for the φ and ψ torsion angles of amino acids.

the sequence will prefer. As the number n of residues increases in the sequence, computational methods become limited by computational time, and the database methods become limited by scarcity of data points. Therefore, simplifying assumptions are necessary to construct the probability function. The simplest one is the independent residue assumption, where the probability is

$$P = p^{\alpha_1}_{\zeta_1} p^{\alpha_2}_{\zeta_2} \ldots p^{\alpha_i}_{\zeta_i} \ldots p^{\alpha_n}_{\zeta_n} \tag{1}$$

where $p^{\alpha_i}_{\zeta_i}$ is the 'singlet' probability that the ith residue α_i is in state ζ_i. The second level of approximation is based on the pairwise dependence assumption, according to which

$$P = p^{\alpha_1}_{\zeta_1} \prod_{i=2}^{n} q^{\alpha_{i-1}\alpha_i}_{\zeta_{i-1}\zeta_i} \tag{2}$$

where, $q^{\alpha_{i-1}\alpha_i}_{\zeta_{i-1}\zeta_i}$ is the conditional probability that residue α_i is in state ζ_i given that the $i-1$st residue is α_{i-1} and is in state ζ_{i-1}. Conditional probabilities are defined as

$$q^{\alpha_{i-1}\alpha_i}_{\zeta_{i-1}\zeta_i} = \frac{p^{\alpha_{i-1}\alpha_i}_{\zeta_{i-1}\zeta_i}}{p^{\alpha_{i-1}}_{\zeta_{i-1}}} \tag{3}$$

where $p^{\alpha_{i-1}\alpha_i}_{\zeta_{i-1}\zeta_i}$ is the joint probability that the residues $i-1$ and i are of types α_{i-1} and α_i and are in states ζ_{i-1} and ζ_i, respectively.

Long-range effects may perturb the statistics of a chain either by introducing new states, or by changing the relative probabilities of known states. Since the states shown in the Ramachandran maps are separated by steric hindrances, it is plausible to assume that new states, not shown in the original Ramachandran maps, are not created. Long-range effects, therefore, can change only the conditional probabilities defined by Eq. (3). For specific types of long-range correlations, such as α and β structures, correlations may be incorporated into the conditional probabilities following the scheme proposed by Mattice.[2] According to this scheme, the conditional probability $q^{\alpha_{i-1}\alpha_i}_{\mathrm{HH}}$ that residue α_i is in state H when residue α_{i-1} is in state H is perturbed because of its nonlocal interactions. The conditional probability $q^{\alpha_{i-1}\alpha_i}_{\mathrm{EE}}$ for the extended state will be perturbed because of its nonlocal interactions also. The differences between the two interactions will originate from the nonlocal interaction energies E_{H} and E_{E} of the ith residue with the succeeding residues along the chain. The perturbed conditional probabilities $q^{*\alpha_{i-1}\alpha_i}_{\mathrm{HH}}$ and $q^{*\alpha_{i-1}\alpha_i}_{\mathrm{EE}}$ are then expressed as

$$q^{*\alpha_{i-1}\alpha_i}_{\mathrm{HH}} = \frac{e^K q^{\alpha_{i-1}\alpha_i}_{\mathrm{HH}}}{e^K q^{\alpha_{i-1}\alpha_i}_{\mathrm{HH}} + q^{\alpha_{i-1}\alpha_i}_{\mathrm{EE}}} \tag{4}$$

$$q^{*\alpha_{i-1}\alpha_i}_{\mathrm{EE}} = C \frac{q^{\alpha_{i-1}\alpha_i}_{\mathrm{EE}}}{e^K q^{\alpha_{i-1}\alpha_i}_{\mathrm{HH}} + q^{\alpha_{i-1}\alpha_i}_{\mathrm{EE}}} \tag{5}$$

where C is a constant of proportionality, and

$$K = -(E_{\mathrm{H}} - E_{\mathrm{E}})/(kT) \tag{6}$$

Substitution of Eqs. (4) and (5) into Eq. (2) yields the probability of the given segment to be in either a helical or an extended state. Equating the long-range interaction energies to zero leads to $q^{*\alpha_{i-1}\alpha_i}_{\mathrm{HH}} = q^{\alpha_{i-1}\alpha_i}_{\mathrm{HH}}$ and $q^{*\alpha_{i-1}\alpha_i}_{\mathrm{EE}} = q^{\alpha_{i-1}\alpha_i}_{\mathrm{EE}}$, and Eq. (2) is recovered in its unperturbed form. The derivation of the form of Eqs. (4) and (5) is presented in the Appendix.

Configurational States of a Protein may be Studied by the Rotational Isomeric States Formalism

Individual residue

The energy of state η of residue type α in the chain may be defined as E^{α}_{η}. For an isolated residue, identified by the subscript 0 throughout this section, the joint probability $p^{\alpha}_{0\eta}$ that the residue is of type α and is in state η is defined as

$$p^{\alpha}_{0\eta} = \frac{e^{-E^{\alpha}_{\eta}/RT}}{z} \tag{7}$$

The singlet partition function z is defined as

$$z = \sum_{\alpha} \sum_{\eta} e^{-E^{\alpha}_{\eta}/RT} \tag{8}$$

The energies that appear in Eq. (7) may be written in terms of a reference state independent of residue type

$$E^{\alpha}_{\eta} = \overline{E}_{\eta} + \delta E^{\alpha}_{\eta} \tag{9}$$

Here, $\overline{E}_{\eta}$ is the energy of state η that is common to all residue types and in this sense it is 'reference' energy for this state.

δE^{α}_{η} is the residue-specific part. In the absence of correlations with neighbors, that is NN, context, and long-range effects, it consists of contributions from the intrinsic energy of residue type α. Effects of correlations in addition to those of the intrinsic effects will be discussed in the next section. The energy $\overline{E}_{\eta}$ of a state η common to all amino acids contains no useful information relating to the type of a residue[21] and may be removed. Its values differ for different states though, and its removal distorts the energy surface for the ϕ–ψ maps. One may define probabilities $p_{0\eta}$ of the states η irrespective of the type of a residue according to the relation

$$p_{0\eta} = \frac{z(\eta)}{z} \tag{10}$$

where

$$z(\eta) = \sum_{\alpha} e^{-E^{\alpha}_{\eta}/RT} \tag{11}$$

Using Eqs. (7) and (10), the part $\Delta E^{\alpha}_{0\eta}$ of the energy of the state η of the amino acid type α relative to the reference state is expressed as

$$\Delta E^{\alpha}_{0\eta} = -RT \ln\left(\frac{p^{\alpha}_{0\eta}}{p_{0\eta}}\right) \tag{12}$$

Isolated pair

Energy levels of isolated neighboring pairs are required for the Markovian approximation of chain conformations. The energy of residue α in state η and residue β in state ζ is represented by $E^{\alpha\beta}_{\eta\zeta}$. It consists of the following terms:

$$E^{\alpha\beta}_{\eta\zeta} = E^{\alpha}_{\eta} + E^{\beta}_{\zeta} + \delta E^{\alpha\beta}_{\eta\zeta} \tag{13}$$

Here, E^{β}_{η} and E^{α}_{ζ} are the singlet energies defined by Eq. (7), and $\delta E^{\alpha\beta}_{\eta\zeta}$ is the correlation energy of the two residues when they are in states α and β, respectively.[6]

The joint probability $p^{\alpha\beta}_{0,\eta\zeta}$ that the $\alpha\beta$ pair is in state $\eta\zeta$ is given by

$$p^{\alpha\beta}_{0\eta\zeta} = \frac{e^{-E^{\alpha\beta}_{\eta\zeta}/RT}}{z_2} \tag{14}$$

where

$$z_2 = \sum_{\alpha,\beta}\sum_{\eta,\zeta} e^{-E^{\alpha\beta}_{\eta\zeta}/RT} \tag{15}$$

is the partition function for the pair of residues.

When the Flory isolated pair assumption holds and NN effects are absent, $\delta E^{\alpha\beta}_{\eta\zeta}$ equates to zero in Eqs. (13) and (14) reduces to the product of singlet probabilities, $p^{\alpha}_{0\eta}p^{\beta}_{0\zeta}$.

To define a reference state for the pair of residues, we introduce the partition function $z(\eta,\zeta)$ as

$$z(\eta,\zeta) = \sum_{\alpha,\beta} e^{-E^{\alpha\beta}_{\eta\zeta}/RT} \tag{16}$$

The probability of the doublet being in states η and ζ, irrespective of the residue type is then

$$p_{0\eta\zeta} = \frac{z(\eta,\zeta)}{z_2} \tag{17}$$

and the energy of the residue pair α,β relative to the reference state at η and ζ is

$$\Delta E^{\alpha\beta}_{0\eta\zeta} = -RT \ln\left(\frac{p^{\alpha\beta}_{0\eta\zeta}}{p_{0\eta\zeta}}\right) \tag{18}$$

Context effects

The singlet and pair probabilities defined by Eqs. (7) and (14) will be modified when a residue is embedded into the chain because the probability of occurrence of a given state for the *i*th residue will depend on the type and states of the residues around it along the chain. In this section, we elaborate on the method of determining the probabilities $p^{\alpha_i}_{\zeta}$ and $p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}$ when the singlet or the pair of residues is embedded in a sequence.

We assume that the pair $\alpha_{i-1}\alpha_i$ is embedded into a sequence S of n residues represented by $\alpha_1\alpha_2\alpha_3\ldots\alpha_{i-1}\alpha_i\ldots\alpha_{n-1}\alpha_n$. First, we evaluate $p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}$ for the case where the residues of the sequence S are fixed but their conformations may take all the allowable values subject to the pairwise probabilities. The $\eta\zeta$'th element of the statistical weight matrix for this pair is defined as $U^{\alpha_{i-1}\alpha_i}_{\eta\zeta;i} = e^{-E^{\alpha_{i-1}\alpha_i}_{\eta\zeta;i}/RT}$. We note here that the energies $E^{\alpha_{i-1}\alpha_i}_{\eta\zeta}$ of the states are the same as those for the isolated pair given in Eq. (7). The full partition function Z is

$$Z = \sum_{\Omega} J^{*}\left[\prod_{i=2}^{n} U^{\alpha_{i-1}\alpha_i}_{i}\right] J \tag{19}$$

where J^{*} and J are the row and column vectors of order n with all elements equal to unity and the first summation is over the full set Ω of residue types of the sequence.

The fraction of occurrence where the *i*th residue α_i is in state ζ and the $i-1$st residue α_{i-1} is in state η averaged over all conformations of the remaining residues of the fixed sequence S is

$$p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}(S) = Z^{-1}J^{*}\left\{\prod_{k=2}^{i-1} U^{\alpha_{k-1}\alpha_k}_{k}\, U'^{\alpha_{i-1}\alpha_i}_{i} \prod_{k=i+1}^{n} U^{\alpha_{k-1}\alpha_k}_{k}\right\} J \tag{20}$$

Here, $U'^{\alpha_{i-1}\alpha_i}_{i}$ is obtained by equating all elements to zero except the $\eta\zeta$'th, and the argument S of the probability denotes that it is calculated for the fixed sequence.

The fraction $p^{\alpha_i}_{\zeta}(S)$ where the *i*th residue α_i is in state ζ, averaged over all the states of the remaining amino acids of the fixed sequence, is given by

$$p^{\alpha_i}_{\zeta}(S) = \sum_{\eta} p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}(S) \tag{21}$$

Pair probabilities averaged over neighboring residue types

If Ω represents a library of sequences and N is the size of Ω, that is the number of different sequences in Ω, then the pair probability $p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}$ that two neighboring residues α_{i-1} and α_i are in states η and ζ, respectively, may be obtained by summing $p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}(S)$ over all possible neighboring residues types. Thus,

$$p^{\alpha_{i-1}\alpha_i}_{\eta\zeta} = \sum_{\{S\}'} p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}(S) \tag{22}$$

where the summation is over the full set[22] of sequences in which the residue types α_{i-1} and α_i are kept fixed. Thus, $p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}$ becomes the probability of the pair in the library of sequences.

The singlet probability $p^{\alpha_i}_{\zeta}$ may be obtained from Eq. (22) according to

$$p^{\alpha_i}_{\zeta} = \sum_{a_{i-1}} \sum_{\eta} p^{\alpha_{i-1}\alpha_i}_{\eta\zeta} \tag{23}$$

Finally, the energies ΔE^{α}_{η} and $\Delta E^{\alpha\beta}_{\eta\zeta}$of states relative to a reference state that is common to all residues is defined as

$$\Delta E^{\alpha}_{\eta} = -RT \ln \left(\frac{p^{\alpha}_{\eta}}{p_{\eta}} \right) \qquad \Delta E^{\alpha\beta}_{\eta\zeta} = -RT \ln \left(\frac{p^{\alpha\beta}_{\eta\zeta}}{p_{\eta\zeta}} \right) \tag{24}$$

where the probabilities in the denominators in Eq. (24) are obtained by summing up for all residues and correspond to those given by Eqs. (10) and (17) for the isolated singlets and doublets. The ratios in Eq. (24) are conditional probabilities. More explicitly, the ratio in the first equation gives the probability that a residue, which is known to be in state η is of type α.

Relationships Between Calculated and Database Derived Potentials

One can now relate the various expressions defined in the preceding sections to those obtained by calculations and from protein libraries. Probabilities given by Eqs. (7) and (14) obviously correspond to those obtained by computational means over isolated molecules. Probabilities given by Eqs. (22) and (23), in contrast, contain context effects, and therefore should be closer to those computed from protein libraries by counting the frequencies of occurrence. Consequently, Eqs. (24) form the basis of energy surface computations from protein libraries.[21] Calculations based on data from the library of full set of native proteins involve both intrinsic, context, and long-range contributions. Those obtained using Eqs. (22) and (23) involve all except long-range contributions. Since they are averaged over all members of Ω, their values depend on the nature of Ω. If Ω or the protein database is large enough, every amino acid will be embedded in different sequences and will be found several times at different environments such as helix, β, etc. If the statistics is performed over a large set then the NN, context, and long-range effects may be averaged out as suggested by previous work.[23] On the computational side, this would then imply that the probabilities $p^{\alpha_i}_{\eta}$ and $p^{\alpha_{i-1}\alpha_i}_{\eta\zeta}$ will be close to $p^{\alpha_i}_{0\eta}$ and $p^{\alpha_{i-1}\alpha_i}_{0\eta\zeta}$.

The energies and the associated probabilities are easily evaluated for a dipeptide by the use of force fields and computational energy minimization tools, either by classical or quantum mechanics. Results of database calculations by the group of Hermans[9] for the residues Ala, Asn, Asp, Gly, and Val and their comparison with computations show that context and long-range effects persist in the database calculations, irrespective of the protein library used. However, the computational approach has its shortcomings also. Freed and collaborators evaluated conformational properties of residues using seven different force-fields and showed that the results depend significantly on the chosen force-field.[7,8] Their results show that further work is needed along this direction.

Example Problem

Determination of probabilities of helical and extended sequences from different protein libraries

In this section, we compute probabilities from three different protein libraries and compare predictions from these libraries. To simplify the presentation, instead of considering the 11 states shown in Figure 1, we consider only helical (H) and extended (E) sequences (states) and ignore all other states. Following previous practice,[8,9,11,24–26] we construct three protein libraries from the databases DSSP and PDBFIND2[27,28] for assigning the proteins to H and E states using the 2485 nonhomologous proteins listed in the October 2004 release of PDBSELECT25.[29,30]

The three libraries are: (i) the coil library, (ii) the full library, and (iii) the helix-extended library. The coil library contains the set Ω_C of residues that are taken from the unstructured regions of proteins. There were 45,500 residues in this set. Their ϕ–ψ angles were determined and those corresponding to the α_R and β_S regions of Figure 1 were considered. The frequencies of occurrence of HH and EE pairs in the coil library are presented in the first two blocks of Table I. The full library contains the set Ω_F of all residues in the DSSP whose ϕ–ψ angles fall into the α_R and β_S regions of Figure 1. There were 202,032 residues in this set. The frequencies of occurrence of HH and EE pairs in the full library are presented in the third and fourth blocks of Table I. While the coil library consisted of sequences that are neither helical nor extended, the full library contained residues from the full protein library. It should be noted that a residue might be in the α_R region, for example, even though it is not embedded into and stabilized in a helical sequence. The helix-extended library contains the set Ω_{HE} of all residues that are in either a helical or an extended structure. There are 10,644 helices and 15,014 extended strands in the data bank. The frequencies of

TABLE I. Frequencies of Occurence of Residue Pairs in the Coil, Full and HE Libraries

Coil library

HH frequencies

	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	4	4	9	7	14	6	0	5	9	55	20	19	15	13	7	21	11	15	18	9
M	6	1	4	2	5	4	1	3	8	23	14	17	6	16	7	5	1	5	13	7
F	8	5	9	7	17	14	5	5	20	42	22	24	15	16	21	35	13	15	21	13
I	6	6	5	16	23	15	7	10	21	51	39	36	15	27	27	43	19	19	33	23
L	17	5	32	34	45	29	6	14	43	117	55	60	27	44	33	71	17	27	52	26
V	14	8	11	27	38	11	6	12	25	63	36	30	15	31	24	56	14	33	45	21
W	4	8	3	5	9	13	2	6	7	14	1	15	3	13	6	16	3	4	9	5
Y	11	1	18	9	20	9	5	10	13	45	22	30	11	19	11	23	5	11	12	16
A	17	13	20	10	39	31	4	15	33	71	39	54	25	30	34	51	13	25	35	21
G	15	15	18	30	30	29	8	12	44	54	40	52	26	30	35	45	20	35	37	21
T	17	17	28	17	46	30	11	19	35	129	43	61	20	48	37	57	19	29	43	21
S	16	16	19	24	61	33	7	19	45	122	48	112	25	42	40	46	22	34	42	40
Q	9	2	13	15	41	19	6	10	24	40	18	26	14	25	21	34	15	24	23	9
N	4	6	26	22	37	28	5	18	35	47	28	34	8	54	28	36	13	16	32	26
E	11	13	21	19	46	26	13	19	29	68	44	45	18	45	37	43	17	29	27	21
D	17	14	37	37	59	27	14	19	33	61	39	44	16	36	37	42	10	23	37	23
H	6	12	17	11	16	14	2	7	15	25	18	16	8	11	7	15	9	14	16	17
R	13	7	24	27	27	22	10	13	14	40	23	32	28	20	23	33	12	26	29	14
K	20	8	24	32	54	26	10	32	40	67	29	34	23	42	47	62	22	27	50	10
P	15	8	22	18	30	27	8	15	31	51	37	56	18	22	39	32	7	26	30	18

EE frequencies

	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	17	8	11	27	24	23	1	9	29	31	63	54	30	56	24	45	13	23	39	83
M	10	4	11	25	23	33	3	9	27	31	64	80	19	34	24	32	10	20	22	78
F	15	17	21	36	46	53	7	19	46	55	90	92	34	72	52	156	24	46	86	121
I	34	19	29	50	67	69	10	36	70	93	150	173	62	95	107	182	41	75	114	266
L	36	22	31	60	85	69	17	22	102	123	193	219	76	123	102	193	43	87	147	396
V	41	27	36	77	112	90	24	43	103	114	167	185	62	138	119	199	39	105	141	284
W	14	4	7	14	20	13	2	5	17	25	26	41	23	23	24	33	16	15	17	24
Y	21	13	22	27	31	40	12	20	43	59	80	77	39	55	70	97	39	47	59	101
A	28	21	26	62	131	104	15	48	92	95	128	115	82	104	62	118	21	56	88	253
G	29	26	26	54	45	32	6	37	72	142	116	129	33	61	151	163	28	41	138	243
T	43	20	60	81	97	78	26	37	72	72	90	92	45	68	49	92	28	44	68	198
S	48	23	53	69	99	75	17	57	105	152	111	175	45	58	70	93	27	48	65	203
Q	22	36	30	41	64	58	13	22	66	41	53	53	20	56	45	55	20	36	43	102
N	27	17	59	28	71	65	18	46	40	57	56	36	35	56	34	51	15	24	50	74
E	23	16	50	67	109	93	18	31	57	77	121	85	41	63	67	83	17	59	76	142
D	33	20	56	69	114	74	35	43	60	65	63	57	24	48	37	65	26	41	42	96
H	20	18	30	19	22	30	26	14	26	39	69	40	17	26	22	27	39	24	25	68
R	32	17	39	66	74	59	40	30	72	59	97	80	44	40	63	70	26	67	54	126
K	30	45	43	107	114	93	22	34	107	63	108	89	27	67	80	101	35	58	124	218
P	58	46	66	89	164	145	26	58	131	92	141	144	73	86	96	96	35	98	118	212

Full library

HH frequencies

	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	91	75	101	135	250	142	37	79	248	185	145	201	173	122	234	149	103	204	216	128
M	81	116	123	243	468	215	33	140	431	196	197	232	220	217	315	213	117	267	361	80
F	101	157	287	348	756	398	98	260	606	321	294	380	261	291	523	377	167	394	434	174
I	146	256	323	500	954	464	121	283	1039	444	434	608	525	488	808	619	199	556	769	212
L	261	427	595	906	1776	969	211	536	1816	938	863	1087	947	733	1579	1063	359	1090	1554	362
V	182	306	368	554	1068	586	128	262	1081	461	485	646	524	420	865	692	193	607	781	174
W	35	79	137	145	303	158	44	81	240	141	84	138	119	130	195	153	67	179	226	80
Y	122	127	291	310	653	317	83	248	455	263	253	291	270	284	411	305	157	350	368	149
A	284	434	704	972	1887	1061	261	546	2005	838	762	987	830	675	1376	963	315	1023	1162	270
G	142	207	324	458	695	459	127	244	628	342	383	465	308	341	513	417	190	382	491	296
T	161	195	331	474	971	581	144	252	748	590	456	545	405	378	627	486	187	422	612	281
S	163	271	405	554	1122	604	183	371	928	565	531	830	565	434	832	647	269	588	767	368

TABLE I. (Continued)

Q	143	211	308	433	982	485	131	261	892	345	383	443	520	345	739	442	212	525	597	133
N	106	154	301	450	762	442	142	293	686	267	368	404	360	321	560	384	176	375	468	311
E	207	413	572	875	1642	889	279	460	1476	495	708	682	795	596	1471	908	334	821	1259	217
D	179	265	524	717	1144	680	214	469	1063	433	563	594	411	385	895	596	224	578	741	436
H	101	94	177	229	478	213	70	149	348	183	194	222	178	142	260	210	120	202	234	149
R	161	212	376	473	1061	509	135	375	930	352	399	543	509	418	1026	655	262	576	662	178
K	209	296	470	659	1121	621	177	484	1308	491	670	621	584	555	1356	757	335	657	1028	194
P	97	114	231	249	558	353	136	252	613	321	382	525	338	310	875	553	187	327	450	135
EE frequencies																				
	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	63	30	62	101	127	120	18	62	103	119	127	160	74	94	108	137	47	102	114	197
M	27	22	54	92	101	147	16	47	84	80	129	126	54	75	74	85	37	71	82	142
F	84	55	112	228	235	310	56	148	213	244	308	265	132	180	210	283	93	160	243	285
I	121	89	208	343	401	456	72	215	312	360	434	414	175	271	332	389	148	278	334	411
L	121	103	239	367	434	497	87	209	324	320	468	470	222	294	342	403	148	282	401	604
V	149	135	278	525	597	671	105	282	428	497	573	478	237	354	431	474	151	359	488	483
W	37	22	47	67	99	92	18	52	71	72	88	109	69	75	84	73	34	68	83	59
Y	71	54	134	200	212	254	53	122	179	235	268	219	129	152	197	217	107	187	194	234
A	109	89	172	323	346	468	55	195	292	391	379	321	173	242	218	288	113	220	270	452
G	75	78	256	348	232	477	139	226	257	321	442	379	85	193	308	303	143	257	382	586
T	143	90	288	423	517	537	111	209	309	343	293	279	159	180	223	248	93	187	264	469
S	116	83	248	325	421	421	77	229	370	498	291	435	167	209	273	246	83	201	270	573
Q	54	66	125	181	218	262	46	80	147	176	147	134	82	115	121	122	51	95	142	223
N	58	49	157	220	281	268	62	123	158	194	142	115	100	144	142	130	51	97	140	328
E	71	77	194	325	369	416	66	136	176	351	244	178	87	164	220	209	62	156	229	319
D	61	74	171	281	331	311	75	153	241	237	196	198	91	115	218	148	62	142	206	454
H	36	50	114	120	140	144	54	73	116	120	151	103	56	75	66	92	70	74	78	181
R	104	60	153	286	342	326	78	132	196	238	197	175	108	106	179	153	63	170	167	248
K	126	95	169	428	391	521	82	152	255	356	241	204	74	177	213	229	75	169	273	383
P	97	87	164	232	368	415	58	136	318	466	235	270	155	174	242	216	88	203	266	360
HE library																				
HH frequencies																				
	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	39	40	60	75	182	84	23	50	139	38	59	78	98	60	121	50	51	112	116	9
M	36	101	92	186	364	170	25	97	303	93	114	115	150	94	227	131	73	181	228	29
F	60	114	207	293	653	317	66	186	437	158	191	225	208	154	393	235	114	275	266	45
I	113	181	237	412	763	375	83	213	795	275	289	395	396	306	622	384	151	467	529	104
L	166	318	428	727	1444	802	162	396	1431	467	548	683	671	469	1204	664	284	898	1087	154
V	100	188	288	388	807	452	82	209	859	262	310	375	366	301	676	379	133	472	548	84
W	25	54	88	105	235	115	27	48	154	55	62	67	100	64	130	87	42	111	136	27
Y	58	86	196	232	512	232	53	149	318	112	139	170	189	149	297	168	84	247	217	44
A	192	317	529	798	1506	882	194	389	1516	372	487	520	574	353	985	523	207	721	805	109
G	71	109	175	279	444	244	65	113	334	110	156	149	148	116	260	137	73	180	193	66
T	77	121	216	307	668	325	83	169	483	149	201	187	247	135	317	185	86	239	284	48
S	68	135	236	366	668	342	71	180	490	146	228	230	246	145	423	244	107	280	305	27
Q	72	150	223	321	712	338	81	178	668	146	241	244	391	199	509	254	140	383	397	39
N	52	88	169	251	443	234	54	152	397	70	169	165	170	109	305	148	77	215	223	10
E	107	289	422	676	1203	645	160	324	1143	204	423	405	567	317	1076	468	189	646	823	53
D	73	133	308	426	687	415	116	245	607	111	247	236	221	144	494	238	107	286	326	22
H	41	65	126	160	322	124	47	79	229	57	87	102	130	73	167	116	75	133	126	8
R	85	158	251	414	790	406	95	214	732	152	232	304	391	225	730	395	158	434	455	56
K	90	197	282	505	844	439	101	276	908	169	369	344	375	294	836	384	165	432	648	60
P	18	47	94	121	250	143	36	79	248	72	101	95	133	37	308	115	54	102	134	20
EE frequencies																				
	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	51	27	88	114	162	154	28	83	94	81	89	97	44	42	77	38	38	92	84	27
M	25	31	61	116	121	178	20	56	77	48	91	75	45	44	74	59	40	73	72	44
F	82	61	163	329	317	443	62	202	239	177	315	262	143	140	250	175	100	175	225	66
I	137	135	340	585	568	715	108	322	442	340	505	453	186	243	394	308	163	329	324	160
L	133	143	320	541	620	811	118	353	354	220	485	364	259	222	399	289	189	329	323	118
V	209	172	471	794	856	1047	148	435	554	374	697	463	285	242	518	396	203	440	530	195

TABLE I. (Continued)

W	38	26	71	93	120	138	25	87	72	58	93	57	49	60	82	56	29	77	79	9
Y	87	55	209	310	311	376	69	191	203	152	267	193	113	110	192	156	91	198	158	70
A	96	85	237	425	376	572	70	212	256	176	274	202	100	116	168	104	92	153	173	59
G	78	69	207	294	259	351	85	187	168	142	229	172	83	93	129	76	68	128	135	65
T	116	92	343	526	578	690	106	253	263	202	249	184	102	84	136	111	78	163	139	81
S	92	67	249	349	325	420	76	192	211	169	179	144	68	68	88	68	64	94	96	49
Q	49	33	134	230	232	326	49	92	98	77	115	89	52	52	80	44	50	72	70	37
N	39	39	118	212	178	253	35	77	81	54	77	48	35	26	46	31	28	53	57	24
E	74	79	229	431	377	535	73	156	164	106	153	87	65	65	136	80	63	124	142	75
D	46	31	102	215	170	262	37	110	101	63	69	64	30	33	44	23	31	42	56	29
H	29	36	99	178	164	203	32	82	90	56	91	61	43	29	42	52	43	52	49	32
R	103	57	210	368	383	468	59	147	145	120	154	117	69	50	118	82	63	120	112	83
K	98	75	189	419	395	610	62	153	202	127	183	143	61	79	121	70	47	120	139	83
P	25	13	48	113	88	163	10	42	49	31	37	31	27	19	27	11	25	22	28	4

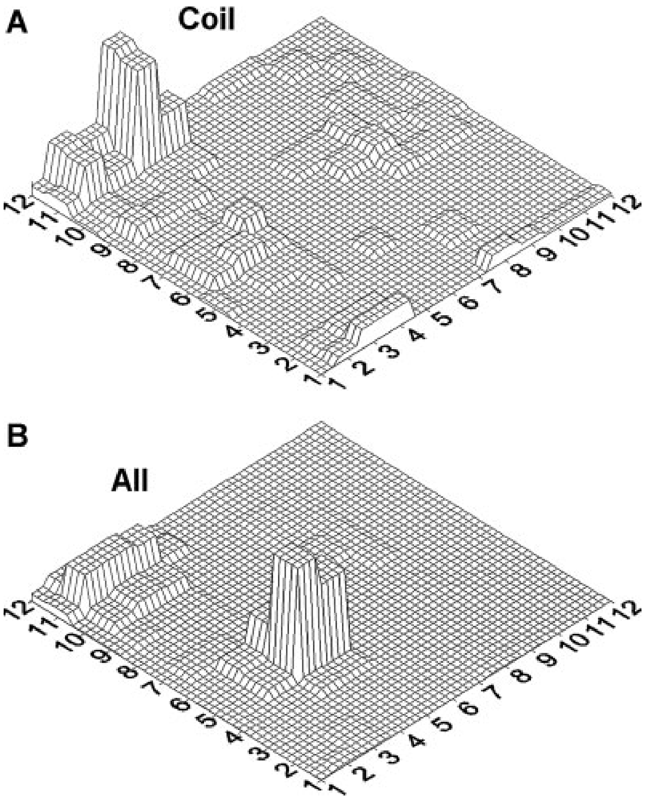

Fig. 2. ϕ–ψ frequency map for lysine preceded by Isoleucine. Data from the coil (**a**) and full libraries (**b**) are used.

occurrence of HH and EE pairs in the helix-extended library are presented in the fifth and sixth blocks of Table I.

Significant differences exist in probabilities when different libraries are used, as may be seen from the different sets of data presented in Table I. As an illustrative example, we present in Figure 2, the frequency ϕ–ψ map for lysine preceded by isoleucine by using the data from the coil and full libraries. The ordinate values, not indicated in the figure, are numbers of observations for each state normalized by the total number of observations of the isoleucine–lysine pair. While mostly the extended conformations are available to lysine in the coil library, helical states are predominant in the full-library.

For each of the libraries, we counted the frequencies, $f_{\mathrm{H}}^{\alpha_i}, f_{\mathrm{E}}^{\alpha_i}, f_{\mathrm{HH}}^{\alpha_{i-1}\alpha_i}, f_{\mathrm{EE}}^{\alpha_{i-1}\alpha_i}$of occurrence of the 20 residue types in H and E states, respectively. We did not discriminate for the positions of the residues inside the H's and E's. Therefore, the superscript i in these expressions are inconsequential and will be omitted in the sequel. We also constructed the sums $f_{\mathrm{H}} \equiv \sum_{\alpha} f_{\mathrm{H}}^{\alpha}$, $f_{\mathrm{E}} \equiv \sum_{\alpha} f_{\mathrm{E}}^{\alpha}$, $f_{\mathrm{HH}} \equiv \sum_{\alpha,\beta} f_{\mathrm{HH}}^{\alpha\beta}$, $f_{EE} \equiv \sum_{\alpha,\beta} f_{\mathrm{EE}}^{\alpha\beta}$. We define conditional probabilities in analogy to those introduced in Eq. (24). Thus, the four relative probabilities are

$$\hat{p}_m^{\alpha} = \frac{f_m^{\alpha}}{f_m} \quad m \in \{\mathrm{H}, \mathrm{E}\} \tag{25}$$

and

$$\hat{p}_{mm}^{\alpha\beta} = \frac{f_{mm}^{\alpha\beta}}{f_{mm}} \quad m \in \{\mathrm{H}, \mathrm{E}\} \tag{26}$$

Defined in this manner, $\hat{p}_m^{\alpha}$ and $\hat{p}_{mm}^{\alpha\beta}$ are the maximum likelihood estimates of the true probabilities, p_m^{α} and $p_{mm}^{\alpha\beta}$.

We obtained the numerical values of Eqs. (25) and (26) from the full data sets Ω_{F} and Ω_{C}. Predictions were performed on the set Ω_{HE}. For the set Ω_{HE}, we used 50% of the data set chosen randomly as the training set on which the numerical values of Eqs. (25) and (26) were obtained. The remaining half was used to test the predictions.

Calculations of the probabilities of m-sequences

We used Eqs. (1) and (2) for this purpose, that is

$$\hat{P}_m = \hat{p}_m^{\alpha} \hat{p}_m^{\beta} \ldots \hat{p}_m^{\gamma} \ldots \hat{p}_m^{\varepsilon} \tag{27}$$

where, $\hat{p}_m^{\alpha}$'s are the estimates that maximize $\hat{P}_m \cdot \hat{P}_m$ is the estimated probability of observing a strand when the corresponding type is m, which is generally referred to as the likelihood function. We identify the type of the strand as $\hat{m}$ by

$$\hat{m} = \arg\max\{\hat{P}_H, \hat{P}_E\}, \tag{28}$$

where Arg max stands for the argument of the maximum; so in this case $\hat{m} = H$ if $\hat{P}_{\mathrm{H}} \geq \hat{P}_{\mathrm{E}}$ and $\hat{m} = E$ if $\hat{P}_{\mathrm{H}} \leq \hat{P}_{\mathrm{E}}$. In words, we compare the likelihood of observing a given sequence when the underlying type of the strand is H or E, and choose the type that maximizes this likelihood. Here, we note that in this methodology there are two maximization problems; one corresponds to finding the maximum likelihood estimators for p_m^α, while the second is specified by Eq. (28).

Similarly, letting $\hat{q}_{mm}^{\alpha\beta}$ be the maximum likelihood estimate of the conditional probability $q_{mm}^{\alpha\beta}$, we have from Eq. (2):

$$\hat{P}_m = f_m^\alpha \prod_{\beta,\gamma} \hat{q}_{mm}^{\beta\gamma} \tag{29}$$

We identify the type of the strand as $\hat{m}$ again according to Eq. (28).

Probabilities based on long-range interactions and pairwise dependent frequency of occurrence of amino acids

Let $\hat{K}$ and $\hat{C}$ be the respective estimates of the two parameters K and C associated with Eqs. (4) and (5). The methodology of computing $\hat{K}$ and $\hat{C}$ is different from the maximum likelihood estimators used in Eq. (28). Now our aim is to maximize the percentage of correctly identified type of strands, as opposed to maximizing the likelihood of observing a sequence separately for $m = H$ and $m = E$. When maximizing the likelihood function corresponding to different types of the strand, the maximum likelihood estimators of p_m^α and $q_{mm}^{\alpha\beta}$ for $m = H$ and $m = E$ are computed independently of each other. In this case, however, the parameters K and C appear in the estimators of the perturbed conditional probabilities $q_{\mathrm{HH}}^{*\alpha_{i-1}\alpha_i}$ and $q_{\mathrm{EE}}^{*\alpha_{i-1}\alpha_i}$ as seen in Eq. (28). Hence, we cannot consider the problem of maximizing the likelihood function for each strand m separately. Instead, we need to maximize the percentage of correctly identified type of strands with respect to K and C, and use the same estimators $\hat{K}$ and $\hat{C}$ for both $m = H$ and $m = E$. To state this computation clearly, let

$$\hat{P}_H^* = \hat{p}_H^{\alpha_1} \prod_{i=2}^{n} \hat{q}^{*\alpha_{i-1}\alpha_i}_{\mathrm{HH}} \text{ and } \hat{P}_E^* = \hat{p}_E^{\alpha_1} \prod_{i=2}^{n} \hat{q}^{*\alpha_{i-1}\alpha_i}_{\mathrm{EE}}, \tag{30}$$

be the estimated probability of observing H and E, for the given sequence, where

$$\begin{aligned} q^{*\alpha_{i-1}\alpha_i}_{H_{i-1}H_i} &= \frac{e^{\hat{K}} q^{\alpha_{i-1}\alpha_i}_{H_{i-1}H_i}}{e^{\hat{K}} q^{\alpha_{i-1}\alpha_i}_{H_{i-1}H_i} + q^{\alpha_{i-1}\alpha_i}_{E_{i-1}E_i}} \quad \text{and} \\ q^{*\alpha_{i-1}\alpha_i}_{E_{i-1}E_i} &= \hat{C} \frac{q^{\alpha_{i-1}\alpha_i}_{E_{i-1}E_i}}{e^{\hat{K}} q^{\alpha_{i-1}\alpha_i}_{H_{i-1}H_i} + q^{\alpha_{i-1}\alpha_i}_{E_{i-1}E_i}} \end{aligned} \tag{31}$$

Hence, the maximization with respect to $\hat{K}$ and $\hat{C}$ involves the probabilistic features of $m = H$ and $m = E$ at the same time. Then, we can identify the type of the strand as $\hat{m}$ for fixed $\hat{K}$ and $\hat{C}$ by Eq. (28).

We find the values of $\hat{K}$ and $\hat{C}$ by iterating the solution for a wide range of combinations of these two parameters. Although the two new parameters C and K introduced in Eqs. (4) and (5) may appear as fitting coefficients, they can be interpreted consistently in terms of the physics of the problem. The hydrogen bonds that stabilize the helices and extended strands apply forces that extend beyond local effect as discussed in the Appendix. The range of the long-range forces exerted in this manner is different for a helix and an extended strand. In the helix, the force on the ith residue comes from (i+k)th residue that also lies within the same helix, in general. However, in an extended strand the force on the ith residue comes from residue i+k that does not belong to the extended strand in consideration. From a mechanistic point of view, the constraining effects of these two types of hydrogen bonds should be different, where the stabilizing effect is expected to decrease as the number of residues in between increases. Accordingly, a helical conformation should be more strongly constrained than the extended strand. Therefore, the conditional probabilities that determine whether the structure is a helix or an extended strand should be reevaluated in the presence of these long-range forces. More precisely, to include the effect of these long-range interactions, we consider two concepts: (i) the allowed region in the ϕ–ψ map for an extended structure is about six to eight times larger than that for the helix, as may be verified from Figure 3, and (ii) the difference between the energy levels of helix and extended strands shows the difference between the strengths of interactions within the strands due to the hydrogen bonds. The parameter C in Eqs. (4) and (5) accounts for the difference of the size of the allowed region in the ϕ–ψ map for helical and extended strands, and the constant K in Eq. (5) accounts for the energetic differences as discussed in the Appendix.

For the three libraries, the values of $\hat{K}$ and $\hat{C}$ that gave the best prediction were 1.34 and 8, respectively. From Eq. (6), the value of 1.34 for $\hat{K}$ corresponds to an energy difference of 0.75 kcal/mol between the helical and extended sequences.

Role of chameleons

A chameleon sequence is one that may exist either in a helical or an extended configuration in a databank.[31,32] This is an indication that the overall probability of occurrence of a chameleon sequence as an H or an E is close to each other. Our approach, therefore, cannot differentiate their occurrence in either structure. For this reason, we removed the chameleon sequences from Ω_{HE}, and applied our analysis to the remaining sequences. Specifically, we removed the secondary structure sequences that have a chameleon subsequence of length equal to or greater than the half-length of the sequence. When we considered sequences of length four or more,

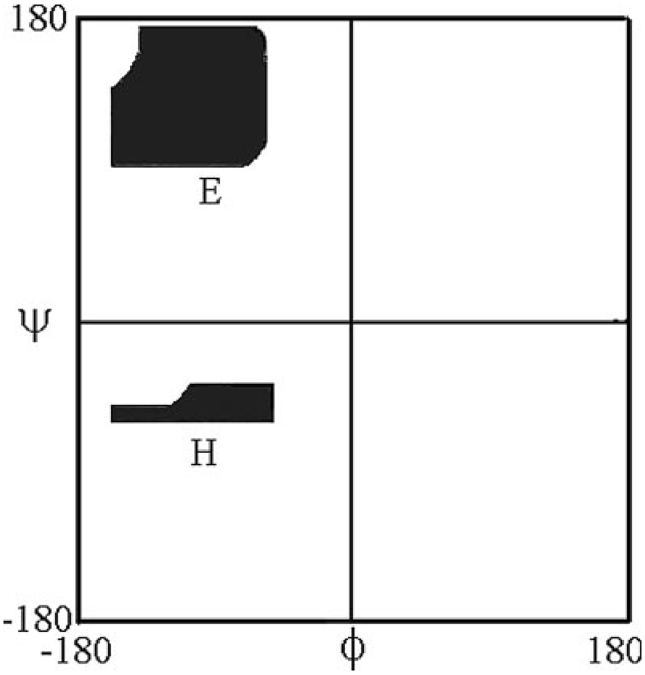

Fig. 3. The extended and helical regions of the Ramachandran map.

10,297 out of 25,879 structures were eliminated. Results of calculations based on the HE library in the absence of chameleons are presented in Table I as numbers in parenthesis.

RESULTS AND DISCUSSION

The results of calculations outlined in the preceding section are presented in Table II. The values are given as percent accuracy of prediction. The numbers in parenthesis are results obtained after removing chameleon sequences.

The estimates of probabilities from the coil library give the poorest agreement. The doublet probabilities are calculated from Ω_C as $p^{\alpha\beta}_{mm} = \sum_{\{S\}'} p^{\alpha\beta}_{mm}(S)$ in accordance with Eq. (22). Here, S corresponds to all sequences in Ω_C that surround a given pair $\alpha\beta$. These calculations form the training phase. Thus, the system is trained over Ω_C but tested on m-sequences in Ω_{HE}. The nature of the sequences surrounding $\alpha\beta$ in Ω_C are different than those in Ω_{HE}. This discrepancy is perhaps the major cause of the poorness of predictions. Singlets are similarly obtained from doublets, according to $p^{\beta}_{\eta} = \sum_{\alpha} \sum_{\zeta} p^{\alpha\beta}_{\zeta\eta}$. Same arguments on doublets are therefore valid for the singlets.

Comparison of the results for the singlets and doublets for the three libraries shows that considering doublets does not improve the predictions. This observation implies that either (i) there are no correlations between neighboring pairs, or (ii) the correlations that exist between neighboring pairs cannot discriminate between helical or extended conformations. In recent work, the full set of native proteins with 1646 nonredundant PDB structures that are nonhomologous and representative of PDB structures[30] was used as the protein library and the correlations between the states of ψ_{i-1}–ϕ_i angles joining the $i-1$st residue with the ith in the sequence of the protein CI2 were analyzed.[33] Calculations indicated strong correlations in agreement with similar calculations of Freed and coworkers.[8] The presence of near neighbor correlations is now well described by the electrostatic solvation energies and hydrogen bonds.[10,13–15,22,34–36] We therefore conclude that correlations that exist between neighboring pairs cannot discriminate between helical and extended sequences when extracted from protein data banks.

TABLE II. Comparison of Percent Accuracy of Predictions from Different Libraries

	Singlet	Doublet (no long-range)	Doublet (with long-range)
Coil library	61	56	73
All library	72	70	81
H-E library	73 (75)[a]	75 (77)[a]	84 (88)[a]

[a]Numbers in parenthesis are obtained after removing the chameleon sequences

In Table I, predictions based on the coil library yields 61% for the singlets whereas this value drops to 56% for the doublets. In the singlet data set, the data is collected into 12 bins each of 30° intervals whereas in the doublet data set 144 bins are used, resulting in a dispersion of data. Thus, the decrease from 61–56% may be attributed to sparseness of data in the latter.

In the coil library, the states are distributed more uniformly over the 12 bins whereas in the *H–E* library the data is concentrated in the bins that correspond to helical and extended states. Dispersion of data over states in the coil library constitutes another source of sparseness that may be affecting the statistics.

As may be seen from the last column of Table I, introducing long-range correlations in a mean-field sense makes a significant improvement in predictions. According to these, the stability of *H*'s and *E*'s are distinguished from each other by biasing the conditional probabilities of successive pairs using a modified statistical weight used by Theodorou and Suter for evaluating long-range effects in dense polymer systems.[37] Contributions from long-range effects to HH and EE pairs may be compared with each other by eliminating the common denominators in Eq. (31) that leads to

$$\frac{q^{*\alpha_{i-1}\alpha_i}_{H_{i-1}H_i}}{q^{*\alpha_{i-1}\alpha_i}_{E_{i-1}E_i}} = \left(\frac{e^{\hat{K}}}{\hat{C}}\right)\left(\frac{q^{\alpha_{i-1}\alpha_i}_{H_{i-1}H_i}}{q^{\alpha_{i-1}\alpha_i}_{E_{i-1}E_i}}\right) \qquad (32)$$

This equation states that the ratio $q^{*\alpha_{i-1}\alpha_i}_{H_{i-1}H_i}/q^{*\alpha_{i-1}\alpha_i}_{E_{i-1}E_i}$ of the perturbed conditional probabilities is $e^{\hat{K}}/\hat{C}$ times (which is 0.477 according to present calculations) the ratio $q^{\alpha_{i-1}\alpha_i}_{H_{i-1}H_i}/q^{\alpha_{i-1}\alpha_i}_{E_{i-1}E_i}$ of unperturbed conditional probabilities. Thus, long-range perturbations favor extended sequences over helical ones.

It is to be noted that the long-range correlations introduced in a mean-field way in the present article have been addressed recently by Fang and Shortle,[38,39] quantifying sequence correlations present in unfolded proteins, coming from interactions between side-chains, the backbone, and the nearby side-chains. Their goal is to use these sequence correlations to identify the correct native backbone topology that corresponds to the amino acid sequence of a long peptide (e.g., 30 residues).

The removal of the chameleon sequences that contain residues that are 50% or more identical among helices and extended conformations further improved the predictions. The highest improvement was when long-range effects were considered, as will be seen from the last row of Table I. The chameleon test is an indication of the sensitivity of the proposed method in discriminating between helical and extended conformations.

We expect that the discussion of the computational basis of probabilities in this work will serve as a guide in interpreting knowledge based probabilities. However, several key questions brought up are not answered conclusively and awaits further work. Do the context effects average out in calculating probabilities on sufficiently large databases? If so, do we recover the probabilities for the isolated singlets and pairs? The answers to these two questions are important because if they are both affirmative, then the determination of probabilities from isolated singlets and doublets, a relatively easy task that may be carried out computationally, will allow characterization of conformations of full proteins.

ACKNOWLEDGMENTS

The authors gratefully acknowledge helpful discussions with Dr. Ozlem Keskin for, and suggestions by Professor Robert Baldwin that are incorporated into the final version of the manuscript.

REFERENCES

1. Flory PJ. Statistical mechanics of chain molecules. New York: Wiley; 1969.
2. Mattice WL, Suter UW. Conformational theory of large molecules. New York: Wiley-Interscience; 1994.
3. Volkenstein M. Configurational statistics of polymer chains (translated from the Russian ed.; Timasheff MJ, Timasheff SN, Translator). New York: Interscience; 1963 (originally published in 1959).
4. Brant DA, Flory PJ. The role of dipole interactions in determining polypeptide configurations. J Am Chem Soc 1965;87:663,664.
5. Brant DA, Flory PJ. The configuration of random polypeptide chains. II. Theory. J Am Chem Soc 1965;87:1175–1184.
6. Jha AK, Colubri A, Freed KF, Sosnick T. Statistical coil model of the unfolded state: resolving the reconciliation problem. Proc Natl Acad Sci USA 2005;102:13099–13104.
7. Jha AK, Colubri A, Zaman MH, Koide S, Sosnick TR, Freed KF. Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. Biochemistry 2005;44:9691–9702.
8. Zaman MH, Shen MY, Berry RS, Freed KF, Sosnick TR. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides. J Mol Biol 2003;331:693–711.
9. O'Connell TM, Wang L, Tropsha A, Hermans J. The "random-coil" state of proteins: comparison of database statistics and molecular simulations. Proteins: Struct Funct Genet 1999;36:407–418.
10. Avbelj F, Grdadolnik SG, Grdadolnik J, Baldwin RL. Intrinsic backbone preferences are fully present in blocked amino acids. Proc Natl Acad Sci USA 2006;103:1272–1277.
11. Hu H, Elstner M, Hermans J. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. Proteins: Struct Funct Genet 2003;50:451–463.
12. Mu Y, Kosov D, Stock G. Conformational dynamics of trialanine in water. II. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. J Phys Chem B 2003;107:5064–5073.
13. Avbelj F, Baldwin RL. Origin of the neighboring residue effect on peptide backbone conformation. Proc Natl Acad Sci USA 2004;101:10967–10972.
14. Fitzkee NC, Rose GD. Sterics and solvation winnow accessible conformational space for unfolded proteins. J Mol Biol 2005;353:873–887.
15. Aurora R, Rose GD. Helix capping. Protein Sci 1998;7:21–38.
16. Serrano L. Comparison between the ϕ distribution of the amino acids in the protein database and NMR data indicates that amino acids have various ϕ-propensities in the random coil conformation. J Mol Biol 1995;254:322–333.
17. Gibrat JF, Robson B, Garnier J. Influence of the local amino acid sequence upon the zones of the torsional angles ϕ and ψ adopted by residues in proteins. Biochemistry 1991;30:1578–1586.
18. Hong SK, Kurochkina NA, Lee B. Estimation and use of protein backbone probabilities. J Mol Biol 1993;229:448–460.
19. Swindells MB, MacArthur MW, Thornton JM. Intrinsic ϕ, ψ propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 1995;2:596–603.
20. Karplus PA. Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci 1996;5:1406–1420.
21. Sipple M. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
22. Avbelj F, Fele L. Role of main-chain electrostatics, hydrophobic effect, and side-chain conformational entropy in determining the secondary structure of proteins. J Mol Biol 1998;279:665–684.
23. Munoz V, Serrano L. Intrinsic secondary structure properties of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales. Proteins: Struct Funct Genet 1994;20:301–311.
24. Zaman MH, Berry RS, Sosnick TR. Entropic benefit of a cross-link in protein association. Proteins: Struct Funct Genet 2002;48:341–351.
25. Ferro D, Hermans J. Semiempirical energy calculations on model compounds of polypeptides. Crystal structures of DL-acetylleucine *N*-methylamide and DL-acetyl-*N*-butyric acid *N*-methylamide. Biopolymers 1972;11:105–117.
26. Sreerama N, Woody RW. Molecular dynamics simulations of polypeptide conformations in water: a comparison of a, b, and poly(Pro)II conformations. Proteins 1999;36:400–406.
27. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
28. Hooft RWW, Sander S, Vriend G. The PDBFINDER database: a summary of PDB, DSSP, and HSSP information with added value. Comput Appl Biosci 1996;12:525–529.
29. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–417.
30. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
31. Crooks GE, Brenner SE. Protein secondary structure: entropy, correlations, and prediction. Bioinformatics 2004;20:1603–1611.
32. Mezei M. Chameleon sequences in the PDB. Protein Eng 1998;11:411–414.
33. Keskin O, Yuret D, Gursoy A, Turkay M, Erman B. Relationships between amino acid sequence and backbone torsion angle preferences in proteins. Proteins: Struct Funct Bioinform 2004;55:992–998.
34. Avbelj F. Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins. J Mol Biol 2000;300:1335–1359.

35. Avbelj F, Baldwin RL. Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: distributions of ϕ. Proc Natl Acad Sci USA 2003;100:5742–5747.
36. Avbelj F, Fele L. Role of main-chain electrostatics, hydrophobic effect, and side-chain conformational entropy in determining the secondary structure of proteins. J Mol Biol 1998;279:665–684.
37. Theodorou DN, Suter UW. Detailed molecular structure of a vinyl polymer glass. Macromolecules 1985;18:1467–1478.
38. Fang Q, Shortle D. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. Proteins: Struct Funct Genet 2005;60:90–96.
39. Fang Q, Shortle D. Enhanced sampling near the native conformation using statistical potentials for local side-chain and backbone interactions. Proteins: Struct Funct Genet 2005;60:97–102.

APPENDIX

According to the theory,[2] the perturbed conditional probability is given according to the relation

$$q^{*\alpha_{i-1}\alpha_i}_{\eta_{i-1}\zeta_i} = \frac{e^{-E_{\zeta_i;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{\eta_{i-1}\zeta_i}}{\sum_{\zeta_i} e^{-E_{\zeta_i;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{\eta_{i-1}\zeta_i}} \qquad \text{(A1)}$$

Here, $q^{\alpha_{i-1}\alpha_i}_{\eta_{i-1}\zeta_i}$ is the conditional probability that the residue α_i is in state ζ_i given that the residue α_{i-1} is in state η_{i-1} when long-range (LR) forces are absent. In the presence of LR forces, an asterisk is appended as a superscript. The energy $E_{\zeta_{i;LR}}$ is the total LR interaction energy between the residue i when it is in state ζ_i and all other residues along the chain. When the $i-1$st residue is in state H, the ith residue may be either in an H state or an O state, where O represents the states other than H. Similarly, the $i-1$st residue is in state E, the ith residue may be either in an E state or an O state, where O represents the states other than E. For H and E states, Eq. (A1) then reads as follows:

$$q^{*\alpha_{i-1}\alpha_i}_{\mathrm{HH}} = \frac{e^{-E_{H;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{\mathrm{HH}}}{e^{-E_{H;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{\mathrm{HH}} + e^{-E_{O;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{HO}} \qquad \text{(A2)}$$

$$q^{*\alpha_{i-1}\alpha_i}_{\mathrm{EE}} = \frac{e^{-E_{E;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{\mathrm{EE}}}{e^{-E_{E;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{\mathrm{EE}} + e^{-E_{O;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{EO}} \qquad \text{(A3)}$$

The energy $E_{\mathrm{O:LR}}$ appearing in Eq. (A2) may differ from that in Eq. (A3). They may be removed from the equations, however, by redefining $E_{H;LR}$ as $E_{H;LR} = E_{H;LR} - E_{O;LR}$ and $E_{E;LR}$ as $E_{E;LR} = E_{E;LR} - E_{O;LR}$, which is obtained by dividing Eqs. (A2) and (A3) by the respective terms $e^{-E_{O;LR}}/RT$. Here, we keep the symbols $E_{H;LR}$ and $E_{O;LR}$ unchanged in order not to introduce new notation upon redefining. Also using the identities $q^{\alpha_{i-1}\alpha_i}_{HO} = 1 - q^{\alpha_{i-1}\alpha_i}_{HH}$ and $q^{\alpha_{i-1}\alpha_i}_{EO} = 1 - q^{\alpha_{i-1}\alpha_i}_{EE}$, Eqs. (A2) and (A3) is written as

$$q^{*\alpha_{i-1}\alpha_i}_{HH} = \frac{e^{-E_{H;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{HH}}{\left(e^{-E_{H;LR}/RT} - 1\right) q^{\alpha_{i-1}\alpha_i}_{HH} + 1} \qquad \text{(A4)}$$

$$q^{*\alpha_{i-1}\alpha_i}_{EE} = \frac{e^{-E_{E;LR}/RT} q^{\alpha_{i-1}\alpha_i}_{EE}}{\left(e^{-E_{E;LR}/RT} - 1\right) q^{\alpha_{i-1}\alpha_i}_{EE} + 1} \qquad \text{(A5)}$$

Defining two parameters K and C as

$$K = -(E_{H;LR} - E_{E;LR})/RT \qquad \text{(A6)}$$

and

$$C = \frac{(e^{-E_{H;LR}/RT} - 1) q_{HH} + 1}{(e^{-E_{E;LR}/RT} - 1) q_{EE} + 1} \qquad \text{(A7)}$$

and substituting into Eqs. (A4) and (A5) leads to the expressions given by Eqs. (4) and (5).

For the simple case where the pairwise probabilities are independent, one may replace q_{HH} by p_{H}, q_{EE} by p_{E} with similar replacements for the perturbed case. With the assumption of independence, equating the ratio of Eqs. (5) to (4) to the ratio of Eqs. (A5) to (A4) leads to

$$C = \frac{e^{-E_{H;LR}/RT} p^*_E/p_E}{e^{-E_{E;LR}/RT} p^*_H/p_H} \qquad \text{(A8)}$$

The long-range correlations of the helical and β strands are essentially through the formation of hydrogen bonds of residues in H or E states with other residues. The hydrogen bonds in a helical sequence are intramolecular, whereas those of a β sequence are formed with residues external to the sequence. The differences between these two are in the depth of the energy well and the range of the ϕ–ψ region to which the hydrogen bond constrains the residue. Thus, the energies $E_{H;LR}$ and $E_{E;LR}$ are to be interpreted as free energies, which may be written as $U_{H;LR} - TS_{H;LR}$ and $U_{E;LR} - TS_{E;LR}$, respectively, where the S terms are the entropic components relating to the size of the region in the ϕ–ψ plane. Thus, $e^{-E_{H;LR}}/RT = A_H e^{-U_{H;LR}}/RT$ and $e^{-E_{E;LR}}/RT = A_E e^{-U_{E;LR}}/RT$ with the front factors relating to the size of the ϕ–ψ plane.

In as much as LR interactions do not introduce new states but change the depth of the energy landscape, the ratios on the right hand side of Eq. (A8) may be approximated by $p^*_H/p_H = e^{-U_{H;LR}}/RT$ and $p^*_E/p_E = e^{-U_{E;LR}}/RT$. Substituting these in Eq. (A8) leads to

$$C = \frac{A_E}{A_H} \qquad \text{(A9)}$$

Interpreted in this manner, C is a measure of the relative sizes of the regions of the ϕ–ψ plane for the extended and helical states.