PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Local interactions in protein folding determined through an inverse folding model

Ugo Bastolla,[1]* Markus Porto,[2] and Angel R. Ortíz[1]

[1] Centro de Biología Molecular "Severo Ochoa," (CSIC-UAM), Cantoblanco, 28049 Madrid, Spain

[2] Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstr. 8, 64289 Darmstadt, Germany

ABSTRACT

We adopt a model of inverse folding in which folding stability results from the combination of the hydrophobic effect with local interactions responsible for secondary structure preferences. Site-specific amino acid distributions can be calculated analytically for this model. We determine optimal parameters for the local interactions by fitting the complete inverse folding model to the site-specific amino acid distributions found in the Protein Data Bank. This procedure reduces drastically the influence on the derived parameters of the preference of different secondary structures for buriedness, which affects local interaction parameters determined through the standard approach based on amino acid propensities. The quality of the fit is evaluated through the likelihood of the observed amino acid distributions given the model and the Bayesian Information Criterion, which indicate that the model with optimal local interaction parameters is strongly preferable to the model where local interaction parameters are determined through propensities. The optimal model yields a mean correlation coefficient $r = 0.96$ between observed and predicted amino acid distributions. The local interaction parameters are then tested in threading experiments, in combination with contact interactions, for their capacity to recognize the native structure and structures similar to the native against unrelated ones. In a challenging test, proteins structurally aligned with the Mammoth algorithm are scored with the effective free energy function. The native structure gets the highest stability score in 100% of the cases, a high recognition rate comparable to that achieved against easier decoys generated by gapless threading. We then examine proteins for which at least one highly similar template exists. In 61% of the cases, the structure with the highest stability score excluding the native belongs to the native fold, compared to 60% if we use local interaction parameters derived from the usual amino acid propensities and 52% if we use only contact interactions. A highly similar structure is present within the five best stability scores in 82%, 81%, and 76% of the cases, for local interactions determined through inverse folding, through propensity, and set to zero, respectively. These results indicate that local interactions improve substantially the performances of contact free energy functions in fold recognition, and that similar structures tend to get high stability scores, although they are often not high enough to discriminate them from unrelated structures. This work highlights the importance to apply more challenging tests, as the recognition of homologous structures, for testing stability scores for protein folding.

Proteins 2008; 71:278–299.

Key words: protein folding; statistical potentials; fold recognition; structure alignment; threading.

INTRODUCTION

The fundamental tenet of protein folding is that the native structure of a protein is determined uniquely by its amino acid sequence.[1] On the other hand, structural classifications of proteins[2–4] show that a large number of protein sequences have similar structures, although they have low pairwise sequence identity.[5] Therefore, the inverse folding problem, which consists in determining the sequences that have a target structure as ground state, should be formulated, in a spirit similar to statistical mechanics, as the search of the statistical properties of the sequences compatible with a given structure.

We adopt here an analytic model of inverse folding,[6–8] based on a model of protein sequence evolution with

Grant sponsors: I3P program of the Spanish CSIC, the European Social Fund, the Ramon y Cajal program; Grant sponsor: Spanish Ministery of Education and Science; Grant number: FIS2004-05073-C04-04; Grant sponsor: Deutsche Forschungsgemeinschaft; Grant number: PO 1025/1-1; Grant sponsor: Deutscher Akademischer Austauschdienst; Grant number: D/06/12848.
*Correspondence to: Ugo Bastolla, Unidad de Bioinformatica, Centro de Biologia Molecular "Severo Ochoa," Facultad de Ciencias, Universidad Autonoma de Madrid, Cantoblanco, 28049 Madrid, Spain. E-mail: ubastolla@cbm.uam.es
Received 28 August 2006; Revised 6 July 2007; Accepted 10 July 2007
Published online 11 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21730

mutations acting at the DNA level and selection enforcing the stability of the native state with respect to unfolding and misfolding.[9–11] Stability is assessed through an effective folding free energy based on contact interactions,[12] plus interactions local along the chain, which influence the secondary structure of the protein. In this work, we derive these local interaction parameters self-consistently, by fitting the model of inverse folding to site-specific amino acid distributions in the Protein Data Bank (PDB).

The usual approach to determine local interaction parameters from a statistical analysis of protein structures and sequences is based on the propensity of amino acids for secondary structures,[13] and it has been applied by several groups.[14–18] Propensity is defined as the conditional probability of observing an amino acid of type a at a secondary structure position of type s, normalized times the probability to observe a at a generic position. The logarithm of this propensity is interpreted as a statistical potential for local interactions,

$$L_{\mathrm{Prop}}(s,a) = -\log\left(\frac{P(a|s)}{P(a)}\right). \qquad (1)$$

Nevertheless, the terms $P(a|s)$ are influenced by the combined effect of local interactions and other types of interactions, most notably hydrophobic interactions. Since different secondary structures s have different tendencies to be buried or exposed, hydrophobic interactions influence them differentially. This can produce systematic errors in the local energy parameters estimated through the above formula, in which correlations between local and hydrophobic interactions are not taken into account. This scenario is consistent with the experimental finding that amino acid propensities are affected by the environment, in particular by the type of solvent, and by the sequence context.[19,20]

In contrast, the approach that we present in this article takes into account the correlation between local and hydrophobic interactions, and it can substantially reduce these systematic errors. The local interaction parameters derived through this model are strongly correlated with, but significantly different from, the parameters obtained from the standard propensity measure, Eq. (1). The resulting inverse folding model provides an excellent fit of the observed amino acid distributions.

We then combined the local interaction parameters derived in this work with contact interaction parameters, and tested through threading experiments their capacity to recognize structures similar to the native against unrelated structures aligned with gaps. The scoring function is based only on an effective stabililty score, and it does not consider sequence similarity. The recognition of the native structure is already quite good using only contact interactions (98% success rate), and it improves to 100% through the addition of local interactions. In contrast, the recognition of structures similar to the native has a success rate of 52% if only contact interactions are used, and it improves to 61% when local interactions are taken into account. Although they have not been optimized for such a goal, the local interaction parameters derived in this work perform better than local interaction parameters derived from the standard propensity measure for this difficult task of fold recognition, and they improve substantially the fold recognition capability of contact interactions.

Background: inverse folding model with contact interactions

For completeness, we review here our previous inverse folding model,[6–8] which did not consider local interactions. In this model, the sequence profile of single-domain globular proteins is predicted from a structural profile, the effective connectivity (EC) (see Materials and Methods, (Effective connectivity section) for more details).

We assume that folding stability is governed by contact interactions, so that the effective folding free energy of a protein with sequence **A**, folding into a structure with contact matrix **C**, is given by $E(\mathbf{C},\mathbf{A}) = \sum_{i<j} C_{ij} U(A_i,A_j)$, where the terms $C_{ij}U(A_i,A_j)$ represent the free energy in units of $k_B T$ gained with a contact between amino acids of type A_i and A_j. We then consider the hydrophobic approximation, in which the contact interaction matrix $U(A_i,A_j)$ is approximated through the main component of its spectral decomposition, $U(A_i,A_j) \approx \varepsilon_H h(A_i) h(A_j)$. The 20 values $h(a)$ obtained through the main eigenvector of the contact interaction matrix are related to hydrophobicity,[21,22] and they define a hydrophobicity scale that we call interactivity.[23] The vector $h_i \equiv h(A_i)$ is called the hydrophobicity profile (HP) of the sequence **A**.

For this simple model, we can define the optimal HP for a given structure as the HP having minimal hydrophobic energy, $E(\mathbf{C},\mathbf{A}) \approx \varepsilon_H \sum_{i<j} C_{ij} h(A_i) h(A_j)$, at fixed mean and mean square hydrophobicity. Minimizing the energy enforces high stability against unfolding, whereas constraining the mean hydrophobicity enforces high stability against misfolding.

Thus, the optimal HP is the vector h_i that maximizes the quadratic form $\sum_{ij} C_{ij} h_i h_j$ with fixed mean and mean square component. The optimal HP so defined is parallel to a structural profile, the EC. The EC is the vector c_i that maximizes the quadratic form $Q = \sum_{ij} C_{ij} c_i c_j$, with the constraint of fixed mean and mean-square component. In this sense, c_i measures the global connectivity of a site, since sites with large EC are connected to as many sites as possible with large EC. These sites tend to be buried in the core of the protein, whereas sites with small EC components tend to be exposed. We normalize the EC in such a way that its mean component is one. Therefore, the optimality condition provides a strong

relationship between a sequence dependent profile, the HP, and a structural profile, the EC.

For single-domain globular proteins, the EC almost coincides with the eigenvector of the contact matrix corresponding to the largest eigenvalue, which is called the principal eigenvector (PE) of the contact matrix. For multi-domain proteins, or single-domain proteins that have internal modularity, the optimal HP receives important contributions also from other eigenvectors of the contact matrix.[8,24] In such cases, the optimal HP remains parallel to the EC, which is no longer parallel to the PE.

We simulated sequences folding into the same or similar structure through the structurally constrained neutral (SCN) model of protein evolution,[9–11] in which sequences generated through a mutation process are selected if they maintain the stability of the target structure against unfolding and misfolding above predefined thresholds. Such stable sequences have a HP with a large component in the direction of the optimal HP. Averaging these HP, the components perpendicular to the optimal HP almost vanish, so that the average HP is almost parallel to the optimal one. Consistently, we observed that the evolutionary average of the HP has correlation coefficient equal to one with the EC.[23] This leads to predict the evolutionary average HP as

$$[h_i] \equiv \sum_a h(a) P(a|c_i) \approx \langle [h] \rangle + \sqrt{\langle [h]^2 \rangle - \langle [h] \rangle^2} \left(\frac{c_i - \langle c \rangle}{\sigma_c} \right). \quad (2)$$

Here $h(a)$ is the hydrophobicity of amino acid a, measured through the interactivity scale[23] or some other hydropathy scale, and $P(a|c_i)$ is the conditional probability to observe amino acid a at site i with EC component c_i Square brackets $[\cdot]$ indicate the evolutionary average across sequences belonging to the protein family, angular brackets $\langle x \rangle = \sum_i x_i / N$ indicate the site average across protein sites, and the standard deviation is indicated as $\sigma_x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$.

If we assume that the above condition is the only relevant condition that the site-specific amino acid distributions have to fulfill, and we apply a maximum entropy argument, it follows that the amino acid distribution at sites structurally characterized by EC component c, $P(a|c)$ should be a Boltzmann distribution.[6] The model was further improved by considering the genetic code and a mutation process at the DNA level. Assuming that the mutation process fulfills detailed balance (see Materials and Methods, Computations of site-specific distributions section), we finally obtain the following amino acid distribution[7,8]

$$P(a|c_i) = \frac{w_{\rm mut}(a) \ \exp(-\beta(c_i) h(a))}{\sum_a w_{\rm mut}(a) \ \exp(-\beta(c_i) h(a))}, \quad (3)$$

where $w_{\rm mut}(a)$ is the frequency of the amino acid a expected under the mutation process alone The Boltzmann coefficients $\beta(c_i)$ implicitly represent effective selection for folding stability. They are determined by imposing that the average hydrophobicity satisfies Eq. (2), which leads to the equations

$$\frac{\sum_a w_{\rm mut}(a) \ \exp(-\beta(c_i) h(a)) h(a)}{\sum_a w_{\rm mut}(a) \ \exp(-\beta(c_i) h(a))} = \langle [h] \rangle + \sqrt{\langle [h]^2 \rangle - \langle [h] \rangle^2} \left(\frac{c_i - \langle c \rangle}{\sigma_c} \right). \quad (4)$$

Through Eq (4), each $\beta(c_i)$ is determined from the corresponding structural profile c_i, the mutation parameters $w_{\rm mut}(a)$, and the mean and mean square across protein sites of the evolutionary average HP, $\langle [h] \rangle$ and $\langle [h]^2 \rangle$.

The above prediction, Eq. (3), contains many variables, the 20 $w_{\rm mut}(a)$ and the selection coefficients $\beta(c_i)$, which can be determined by using just five parameters: three stationary nucleotide frequencies (the frequency of the fourth nucleotide is set by the normalization condition), used to determine $w_{\rm mut}(a)$ using a reversible mutation model and the standard genetic code as in Ref. 8 (see Methods section), and the parameters $\langle [h] \rangle$ and $\langle [h]^2 \rangle$, influencing $\beta(c_i)$ through Eq. (4).

In Ref. 8, the parameters $\langle [h] \rangle$ and $\langle [h]^2 \rangle$ were computed according to their definition from all sequences in the PDB, whereas the nucleotide frequencies were determined by a maximum likelihood fit of the model to the site-specific amino acid distributions sampled from the PDB. For each set of parameters, the coefficients $w_{\rm mut}(a)$ and the exponents $\beta(c_i)$ were computed, yielding the likelihood of the observed distribution given the predicted one. The model with maximum likelihood parameters yielded an excellent agreement between observed and predicted distributions.

Equation (3) was also tested very satisfactorily against simulations of the SCN protein evolution model, where selection is applied evaluating folding stability through a contact energy function. In this case, the same nucleotide frequencies were used as parameters both in the evolutionary model and in the inverse folding model, and $\langle [h] \rangle$ and $\langle [h]^2 \rangle$ were obtained from the simulations, so that no free parameters were involved in the comparison.[8]

We would like to make at this point some additional remarks. Equation (3) can be interpreted as the stationary distribution of a mean-field evolution model with independent sites that approximates the SCN model. In the SCN model, the positions of the protein are correlated through the global selective constraints on folding stability. In the mean-field model these constraints are self-consistently recovered through Eq. (4), in which two global quantities, the mean and the mean square HP, are used to determine the coefficients $\beta(c_i)$. These coefficients effectively represent the selective process in that, if

they are nonzero, the amino acid distributions deviate from the frequency expected based on mutation, $w_{mut}(a)$.

In the resulting mean-field evolution model mutation and selection are not independent, since the selective coefficients $\beta(c_i)$ depend on the mutation process through the $w_{mut}(a)$ that enter Eq. (4). This is a consequence of the mean-field approximation, which implicitly implies averages over very long time scales. Therefore, selection and mutation get effectively entangled on long time scales, despite in the SCN model they are independent processes.

Site-specific distributions of physico–chemical properties such as hydrophobicity were previously modelled as Boltzmann distributions, inspired by the maximum entropy principle and by the pioneering work by Miyazawa and Jernigan,[25] by Goldstein and coworkers,[26,27] and by Dokholyan and Shaknovich.[28,29] Different from these approaches, in our model the Boltzmann coefficients $\beta(c_i)$ are calculated analytically as a function of the EC, by imposing Eq. (4). Our model is also analogous to the sequence design model studied by Kleinman *et al.*,[30] who adopted a folding model based on contact interactions and determined the corresponding parameters through maximum likelihood fit of the observed protein sequences.

RESULTS AND DISCUSSION

Definition of secondary structure types

We distinguish the secondary structure types defined by the DSSP program[31]: Strands (E), Coils (C), Helices (H), Turns (T), and Bends (S). Beta-bridges (B) are here clustered with strands, and three-helices (G) and five-helices (I) are clustered with alpha helices.

We distinguish the first four positions along an helix (H1–H4), and the last one (HY). All other helical positions are clustered together (H). If the positions before and after the helix are classified by DSSP as coils, we separate them in two new classes, indicated as H0 and HZ respectively. Therefore, an helix is labelled as H0-(H1, H2, H3, H4, H, . . . , HY)-HZ.

For beta structures, we distinguish the first position (E1), positions from the second on (E), and the last position (EY). If there is only one residue (beta bridge), it is classified as EY. If the two positions before and after the strand are classified by DSSP as coils, we put them in two new classes, indicated as E0 and EZ, respectively. Therefore, a beta structure is labelled as E0-(E1, E, . . . , EY)-EZ. This makes a total of 16 secondary structure types.

We also defined a more refined model, in which we distinguish parallel (E) from antiparallel (e) beta bridges based on their hydrogen bonding pattern. We consider that a residue i forms a parallel beta bridge if it is hydrogen bonded to residues j and $j + 2$, or if residues $i + 2$ or $i - 2$ are forming a parallel beta bridge, whereas it forms an antiparallel bridge if it has two hydrogen bonds with the same residue j, or if residues $i + 2$ or $i - 2$ are forming an antiparallel beta bridge. Residues in beta structures that do not fulfill any of these requirements are regarded as antiparallel. Making this distinction, we obtain a total of 21 secondary structure types, with five additional types, ordered according to their position along the strand, and denoted as e0-(e1, e, . . . , eY)-eZ. In the following, we will refer to the simpler classification when not otherwise stated, since we need also nonlocal information to determine whether a beta bridge is parallel or antiparallel.

Inverse folding model with local interactions

We adopt here a structural representation of proteins based on the contact matrix C_{ij} and on the secondary structure profile s_i, as defined by the DSSP program,[31] and further distinguished as explained earlier.

Our starting point is a protein folding model in which the effective free energy in units of k_BT of a protein with sequence $\mathbf{A} = \{A_i\}$, folded into a structure with contact matrix $\mathbf{C}$ and secondary structure $\mathbf{S}$ and measured with respect to the unfolded state, is approximated as the sum of a contact interaction term plus interactions that are local along the sequence and depend only on the secondary structure,

$$E(\mathbf{C},\mathbf{S},\mathbf{A})/k_BT = \sum_{i<j} C_{ij}U(A_i,A_j) + \varepsilon \sum_i L(s_i,A_i) + F_{str}(\mathbf{C},\mathbf{S}) + F_{seq}(\mathbf{A}). \quad (5)$$

The terms $C_{ij}U(A_i,A_j)$ represent the free energy in units of k_BT gained with a contact between amino acids of type A_i and A_j. The terms $\varepsilon L(s,a)$ represent the free energy in units of k_BT gained when amino acid a takes the local conformation corresponding to secondary structure s. The terms $F_{str}(\mathbf{C},\mathbf{S})$ and $F_{seq}(\mathbf{A})$ depend only on the structure and on the sequence, respectively. The coefficient ε is needed to combine the two kinds of energy contribution, since we use for $U(a,b)$ the numerical values reported in Ref. 12, which are in arbitrary units, while the normalization of the parameters $L(s_i,A_i)$ is fixed as explained later. Adding a term $u(a)$ that depends only on the amino acid a to all local interaction parameters $L(s,a)$ only contributes to the sequence term, and adding a term $v(s)$ that depends only on the secondary structure only contributes to the structure term. Therefore, we are free to choose these additive terms in such a way that, in analogy with the propensity parameters derived through Eq. (1), the following normalization conditions hold for all amino acids and all secondary structure types,

$$\sum_s P(s)\ \exp(-L(s,a)) = 1, \quad (6)$$

$$\sum_a P(a)\ \exp(-L(s,a)) = 1. \quad (7)$$

Therefore, if S is the number of secondary structure types, the local interactions can be expressed as a function of only $19 \times (S-1)$ independent parameters [Eq. (6) impose $S + 20$ constraints, of whose $S + 19$ are independent].

We interpret the terms $L(s,a)$ as the result of interactions that are local along the protein chain, such as steric interactions shaping the accessible volume for each secondary structure type and electrostatic interactions, in particular hydrogen bonding and N-capping in alpha helices. In contrast, the contact (hydrophobic) interactions responsible for the burial of the residues depend on the global conformation of the protein chain.

We generalize Eq. (3) as

$$P(a|c_i,s_i) = \frac{f(a)\ \exp(-\beta(c_i)h(a) - L(s_i,a))}{Z(c_i,s_i)}, \quad (8)$$

$$Z(c_i,s_i) = \sum_a f(a)\ \exp(-\beta(c_i)h(a) - L(s_i,a)). \quad (9)$$

The term $f(a)$ represents the frequency of amino acid a that result from the mutation process and from selection for properties not represented in the present model of protein stability. Different from Eq. (3), where we set $f(a) = w_{\mathrm{mut}}(a)$, which depend only on three nucleotide frequencies and on the genetic code,[8] we consider here the frequencies $f(a)$ as free parameters, since the deviations of amino acid frequencies from the values based on nucleotide frequencies can be quite significant,[32] and we want to avoid that they influence our estimate of the parameters $L(s,a)$. The normalization condition $\sum_a f(a) = 1$ implies that only 19 frequencies are independent.

The exponents $\beta(c_i)h(A_i) + L(s_i,A_i)$ represent the effective site-specific selection for folding stability at site i, dependent on the EC c_i and on the secondary structure s_i. This formula can be justified through the following reasoning: we define structural profiles associated to each secondary structure type s, and the corresponding sequence profiles $L(s,a)$, analogous to the hydrophobicity profile $h(a)$. In analogy with Eq. (3), we then assume that the site-specific effective selection processes have a stationary distribution that is an exponential function of the sequence profile with exponents depending on the structural profile, that is $\beta(c_i)h(A_i) + \hat{\beta}(s_i)L(s_i,A_i)$. This assumption is motivated through a maximum entropy argument. We assume that the site-specific effective selection process acting on hydrophobicity depends only on the structural profile c_i, that is $\beta_i = \beta(c_i)$. Similarly, we assume that the selection process acting on local interactions only depends on the corresponding secondary structure. Without loss of generality, we can choose the parameters $L(s,a)$ so that the corresponding coefficients $\hat{\beta}(s)$ take the value one, thus obtaining Eq. (8).

The coefficients $\beta(c_i)$ must be computed imposing that the average site-specific hydrophobicity for sites with EC equal to c_i and secondary structure type s_i, $[h_i] = \sum_a h(a)P(a|c_i,s_i)$, takes the predicted value. However, we can not follow the strategy leading to Eq. (2), based on the predicted optimal HP, since this can not be calculated analytically in the present case, in which the local interaction parameters $L(s,a)$ are not an analytic function of hydrophobicity, and we have to resort to heuristics. For such a purpose, we assume that Eq. (2) is valid averaged over secondary structure types, so that $[h_i]$, averaged over sites with the same effective connectivity c_i and different secondary structure, has correlation coefficient equal to one with c_i. This implies that

$$\sum_a h(a)P(a|c) = \langle [h] \rangle + \sqrt{\langle [h]^2 \rangle - \langle [h] \rangle^2}\left(\frac{c - \langle c \rangle}{\sigma_c}\right). \quad (10)$$

This prediction is in very good agreement with the data sampled from the PDB. The correlation coefficient between the mean hydrophobicity conditioned to the effective connectivity c, $\sum_a h(a)P(a|c)$, and the effective connectivity is $r = 0.9979$ (see Fig. 1), without any free parameter.

Substituting Eq. (8) into Eq. (10), we get an implicit equation for $\beta(c)$ as a function of the mean and mean square hydrophobicity $\langle [h] \rangle$ and $\langle [h]^2 \rangle$, the amino acid frequencies $f(a)$, and the distribution of secondary structure, $P(s|c)$,

$$\sum_s P(s|c)\frac{\sum_a h(a)f(a)\ \exp(-\beta(c)h(a))\ \exp(-L(s,a))}{Z(c,s)} = \langle [h] \rangle + \sqrt{\langle [h]^2 \rangle - \langle [h] \rangle^2}\left(\frac{c_i - \langle c \rangle}{\sigma_c}\right). \quad (11)$$

The parameters $\beta(c)$ are determined by solving numerically the above equation This equation complements Eq. (8), and together with it it represents the main result of this article.

Propensities measures are affected by hydrophobicity

Most approaches to derive local interaction parameters are based on the propensities of amino acids for secondary structures.[13] This is defined as the ratio between the conditional probability to observe amino acid a with secondary structure s and the overall probability to observe a, and the logarithm of the resulting quantity, Eq. (1), is used to estimate local interaction parameters.

Substituting in Eq. (1) the model represented in Eq. (8), we see that the local parameters derived from propensity combine both local and nonlocal interactions,

$$L_{\mathrm{Prop}}(s,a) = L(s,a) - \log\left(\frac{H(s,a)}{\sum_{s'} H(s',a)P(s')\ \exp[-L(s',a)]}\right), \quad (12)$$

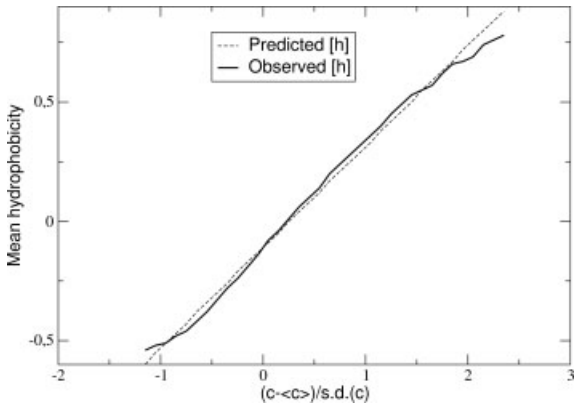

Figure 1

Site-specific mean hydrophobicity [h_i], observed (full line) and predicted (dotted line) according to the r.h.s. of Eq. (10), versus the normalized EC. The hydropathy scale used is the ZC scale, Eq. (14).

where $H(s,a)$ is given by

$$H(s,a) = \sum_c \frac{\exp(-\beta(c)h(a))P(c|s)}{Z(c,s)}. \qquad (13)$$

In other words, secondary structure propensity can be decomposed in two components of different origin:

1. *Local interactions*, expressed in our notation through the term $L(s,a)$. They arise mainly from entropic terms and steric constraints,[33] although enthalpic contributions, in particular hydrogen bonding, have been recently shown to be important as well.[34]
2. *Nonlocal interactions*, which favor the burial or exposure of secondary structure elements.

Nonlocal interactions contribute to the secondary structure propensity of amino acids because secondary structure elements differ in their propensities for solvent exposure. Coiled structures are found very seldom in the core of the protein, where intra-chain hydrogen bonds must be formed, and very frequently on the surface of the protein, where entropic factors favor them. Therefore, hydrophylic amino acids, which are unlikely to be buried, are more likely to be found in coiled structures. Conversely, hydrophobic residues are more likely to occur in beta structures, which tend to be buried. Because of these burial propensities of secondary structures, the effective interaction parameters $L_{\mathrm{Prop}}(s,a)$ derived using the amino acid propensity yield a biased estimate of the local interactions. They overestimate the local interactions favoring the E state for hydrophobic residues, which tend to occupy buried positions where E structures are more frequent, and overestimate local interactions favoring the C state for hydrophylic residues.

Notice that, in the absence of hydrophobic interactions ($h(a) \equiv 0$), and with $f(a) \approx P(a)$, the propensity parameters would be equal to the local interaction parameters, $L_{\mathrm{Prop}}(s,a) \equiv L(a,a)$. In fact, assuming that $h(a) \equiv 0$, and using the normalization condition $\sum_a P(a) \exp(-L(s,a)) = 1$ with $f(a) \approx P(a)$, we find from Eq. (9) that $Z(c,s) = 1$. Using this into Eq. (13), we then get $H(s,a) = 1$. Inserting this result into Eq. (12), and using the normalization condition $\sum_s P(s) \exp(-L(s,a)) = 1$, we finally find $L_{\mathrm{Prop}}(s,a) \equiv L(s,a)$.

Determination of the local interaction parameters

We derived the effective local interaction parameters $L(s,a)$ by fitting the complete inverse folding model, Eqs. (8–10), to the site-specific amino acid distributions observed in the PDB. In this way we take into account the influence of hydrophobicity on the secondary structure propensity, as discussed in the previous section.

The fit was performed by maximizing the sum of the log-likelihood of the observed distributions given the model represented by the predicted distributions, see Materials and Methods section. The observed distributions are calculated by clustering together sites in different proteins with the same secondary structure s_i and similar EC component c_i, and measuring the amino acid distribution in each cluster. When doing so, we cluster together structurally equivalent sites found in different proteins. This is justified as long as the coefficients $\beta(c)$ are the same at sites with equal c_i located in different proteins. We see from Eq. (11) that the coefficients $\beta(c)$ depend on the mean and mean square hydrophobicity $\langle [h] \rangle$ and $\langle [h]^2 \rangle$, on the mutation process, which influence the parameters $f(a)$, and on the distribution of secondary structure across sites with given EC component c_i.

Hydrophobicity is strongly constrained. It can not be too small, otherwise the native state would not be stable against unfolding, and it can not be too large, otherwise the native state would not be stable against misfolding. As a consequence, its mean $\langle h \rangle$ and mean square $\langle h^2 \rangle$ are narrowly distributed for different proteins.[8] We therefore assume that the parameters $\langle [h] \rangle$ and $\langle [h]^2 \rangle$ can be taken as roughly constant across different proteins.

The coefficients $f(a)$ in Eq. (8) are influenced by the mutation process. It is known that mutation patterns depend on the organism, on the DNA strand where the gene is located and, in eukaryotes, they can vary broadly from one region of the genome to another one. Nevertheless, the parameters $f(a)$ mainly depend on the genetic code, and we assume here that they can be taken as roughly constant across the different proteins stored in the PDB.[8]

Finally, clustering together positions with the same c in different proteins requires that the conditional proba-

Table I
Scores Obtained with Various Parameter Sets

Model	Freq.	Hydro	Local	Fit par.	Def. par.	R	l	$\langle\mathcal{L}\rangle$	$\langle$BIC$\rangle$
1	Fit	0	0	19	0	0.593	−5.336	−445	614
2	1	ZC	0	0	20	0.453	−6.061	−496	684
3	Fit	ZC	0	19	20	0.771	−2.802	−267	368
4	Fit	0	Fit	304	0	0.812	−2.333	−234	318
5	Fit	ZC	Prop	19	305	0.943	−0.636	−115	158
6	Fit	ZC	Fit	304	20	0.957	−0.498	−105	150

bility of secondary structure s given the effective connectivity c, $P(s|c)$, is roughly the same for all such positions. This assumption is violated for proteins that belong to different classes of secondary structure content,[2] such as all alpha, all beta, $\alpha + \beta$ and α/β. Nevertheless, we did not distinguish proteins with different secondary structure content, since this would reduce the sample size and complicate the analysis.

Under these assumptions, sites of different proteins having the same structural indicators have the same predicted amino acid distribution.

We obtained from the fit the twenty amino acid frequencies $f(a)$ (19 independent parameters) and the local interaction parameters ($19 \times (S - 1)$ independent parameters). The hydrophobicity parameters were first obtained from the literature. We tested 12 published hydrophobicity scales (see Materials and Methods, Parameters of the model section), obtaining the best results with the buriability scale by Zhou and Zhou[35] and with the scale CH presented in Ref. 23, and the worst results with the scales by Roseman[36] and Kyte and Doolittle.[37]

We then derived a new interactivity scale ZC from the joint distribution of amino acids and EC components. This scale is defined as the average EC component of the sites occupied by the given amino acid,

$$h^{(\mathrm{ZC})}(a) = \sum_c cP(c|a) \qquad (14)$$

(see Materials and Methods, Database section). We then normalize the values of $h^{(\mathrm{ZC})}(a)$ in such a way that the average hydrophobicity is zero, $\sum_a P(a)h^{(\mathrm{ZC})}(a) = 0$, and the average square hydrophobicity is one, $\sum_a P(a)\left(h^{(\mathrm{ZC})}(a)\right)^2 = 1$· With these constraints, the ZC scale has only 18 independent parameters.

The $h^{(\mathrm{ZC})}(a)$ scale is very strongly correlated with other hydropathy scales, since hydrophobic amino acids tend to occupy sites with large EC component. For instance, the correlation coefficient is $R = 0.92$ with the Fauchere and Pliska scale,[38] $R = 0.96$ with the Palliser scale,[39] $R = 0.96$ with our previous interactivity scale derived from the contact interaction matrix,[23] and $R = 0.97$ with the buriability scale by Zhou and Zhou.[35]

The distributions calculated with the new scale ZC fit the observed distributions better than for any other hydropathy scale. Results obtained with this scale will be reported, when not otherwise stated.

To evaluate the contributions of the various components of the model, we tested different models where some of the parameters are set to zero, or to a default value, and the other ones are fitted. They are listed and discussed later, and summarized in Table I. The parameters that we obtained from the fits will be discussed in next section.

To assess the quality of the fit, we calculate for each of the 576 site-specific amino acid distribution the logarithm of the likelihood of the observed distribution given the predicted one, $\mathcal{L}_i$, and averaged them over all distributions, obtaining the mean log-likelihood per distribution $\langle\mathcal{L}\rangle$. Likelihood values depend on the size of the data set, therefore they can only be compared for the same data set. To assess the quality of the fit in a way that is less dependent on the size of the data set, we also report here normalized log-likelihood value. This is obtained by dividing the log-likelihood by the absolute value of the maximum possible likelihood, $\sum_i|\mathcal{L}_{i,\max}|$, which is attained when all the predicted distributions are exactly equal to the observed ones, and subtracting one (see Materials and Methods section). We denote this normalized likelihood by the symbol $l = (\sum_i \mathcal{L}_i - \mathcal{L}_{i,\max})/\sum_i|\mathcal{L}_{i,\max}|$. The maximum possible value of this quantity is zero, and its value is less influenced by the size of the data set. To compare models with different number of parameters, we use the Bayesian information criterion (BIC), see Materials and Methods Bayes Information Criterion section, which is averaged over the 576 distributions and reported in Table I.

We also report in Table I the weighted mean correlation coefficient between observed and predicted distributions.

1. *Pure mutation.* Our null model considers mutation and no selection on protein stability, corresponding to setting $h(a) \equiv 0$, $L(s,a) \equiv 0$, $P_i(a|c_i, s_i) = f(a)$. The best fit yields in this case $l = -5.336$, $\langle\mathcal{L}\rangle = -445$ and $R = 0.593$.
2. *Selection on hydrophobicity, equal frequencies.* The second model considers only selection on hydrophobicity and not on local interactions (we set $L(s,a) \equiv 0$). The genetic code is not considered, and the mutation factors are considered equal for all amino acids ($f(a) \equiv 1$). This model does not contain any free parameters

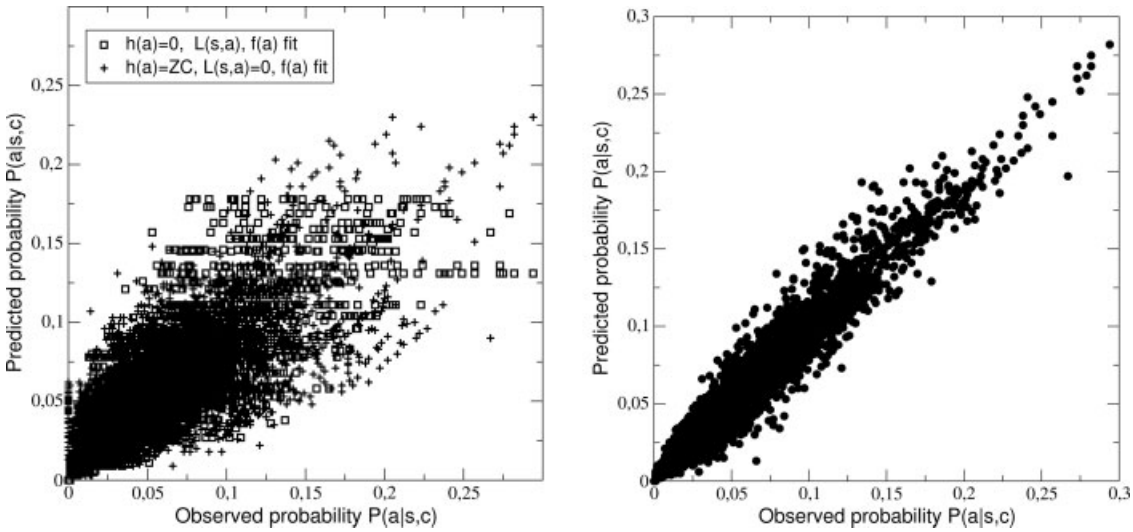

Figure 2

Observed (horizontal axis) versus predicted (vertical axis) amino acid distributions, for all site classes and all amino acid types. Left panel; either local interactions or hydrophobicity set to zero, Right panel; local interactions derived in this work.

except the mean and mean square hydrophobicity that are calculated from their definition rather than fitted. The scale IH provides the best results, $l = -6.06$, $\langle \mathcal{L} \rangle = -496$ and $R = 0.45$, which is not much worse than the mutation model, despite having many fewer free parameters.

If distributions with the same EC and different secondary structure are combined together into structural classes characterized only by the EC component, the model with selection on hydrophobicity performs better than the optimal mutation model, for all scales.

3. *Mutation plus selection on hydrophobicity.* Now, we set again the local interaction parameters to zero, but we we optimize the amino acid frequencies, obtaining $l = -2.802$, $\langle \mathcal{L} \rangle = -267$ and $R = 0.771$ with the ZC scale. Observed versus predicted probabilities obtained with the ZC scale are shown in Figure 2, left plot. This result improves considerably the likelihood both with respect to the model with uniform amino acid frequencies, and with respect to the pure mutation model. Notice that the predicted distribution is the same at all sites with the same EC and different secondary structure. If all these sites are clustered together, as we did in a previous work,[8] the average observed distribution agrees much better with the predicted one ($R = 0.86$). This indicates that sites with similar EC but different secondary structure have significantly different statistical properties.
4. *Mutation plus selection on local interactions.* We now neglect hydrophobicity, setting $h(a) \equiv 0$. We first calculated local interaction parameters from secondary structure propensities as in Eq. (1), and fitted $f(a)$, obtaining $l = -2.333$, $\langle \mathcal{L} \rangle = -234$, $R = 0.81$.

These results do not improve if the interaction parameters are fitted. We obtain in fact $l = -2.333$, $\langle \mathcal{L} \rangle$ -234, $R = 0.81$, confirming our expectation that, if we neglect hydrophobic interactions, the maximum likelihood parameters coincide with the usual propensity parameters.

5. *Mutation plus selection on hydrophobicity and local interactions calculated from propensities.* We set again $L(s,a) \equiv L_{\mathrm{Prop}}(s,a)$, Eq. (1), but now we use the hydrophobicity scale $h(c)$ and fit $f(a)$, obtaining $l = -0.636$, $\langle \mathcal{L} \rangle = -115$, $R = 0.943$, which yields BIC per distribution 158.7.
6. *Mutation plus selection on hydrophobicity and optimized local interactions.* This is the most complete model, having 19×15 additional free parameters with respect to the previous one. Fitting $L(s,a)$ and $f(a)$, we obtain $l = -0.498$, $\langle \mathcal{L} \rangle = -105$, $R = 0.957$. The value of l approaches zero, which is the maximum possible normalized likelihood. The BIC per distribution is 150.1, which shows that the model with fitted local interactions is superior to the one with local interactions derived from propensity.

Remarkably, the mean log-likelihood improves from -234 when we neglect hydrophobic interactions to -105 with only 20 additional parameters, which are not obtained by fitting the data. This improvement is of similar extent to the one obtained using local interaction parameters with respect to setting $L(s,a) \equiv 0$ ($\langle \mathcal{L} = -267$).

Observed and predicted probabilities obtained with the best model are plotted in Figure 2, right plot.

In Figure 2 the observed and predicted probabilities are plotted for each site class and each amino acid type. The plot on the left refers to either local interaction parameters $L(s,a)$ or hydrophobic parameters $h(a)$ set to zero, in the plot on the right we use optimized local interaction parameters and hydropathy scale ZC. In all cases, the amino acid frequencies are fitted. In the plot, site classes with same secondary structure and similar EC component are clustered so that each cluster contains at least 1000 sites.

The quality of the fit between observed and predicted distributions depends on the site class considered, and in particular on the secondary structure. We computed the correlation coefficients between observed and predicted distributions as a function of secondary structure and EC component. To get more homogeneous data, we clustered site classes so that each one contains at least 1000 residues. For most site classes, the correlation coefficients are larger than $R = 0.95$. The only site classes with correlations smaller than $R = 0.90$ are either exposed strands and H4 structures, which are very rare in exposed regions, or buried bends, which are also very rare. Overall, the predicted distributions fit very well the observed ones. We also measured the predicted site-specific mean hydrophobicity, $[h_i] = \sum_a h(a)\ P(a|c_i, s_i)$, which plays a central role in our model. For all secondary structure types, the correlation coefficient between observed and predicted mean hydrophobicity is always larger than $R = 0.95$. However, it is possible to observe some systematic trends: for secondary structures that have a propensity for exposure, like coils, turns, bends, and the positions preceding helices and strands, the predicted hydrophobicity is slightly overestimated, in particular for buried positions, whereas for secondary structures that prefer to be buried the mean hydrophobicity is slightly underestimated, in particular at exposed positions. However, these systematic effects are very small.

We tested the optimal interaction parameters with a different data set, composed by 4400 protein domains with less than 40% pairwise sequence identity, stored in the ASTRAL40 database. In this set, different from the training set, domains present in the same chain are parsed in different structures. This changes the value of the effective connectivity at the domain interfaces. The fitted local interaction parameters and amino acid frequencies, applied to site-specific amino acid distributions sampled from this set of proteins, yield $l = -1.531$ and mean correlation coefficient $r = 0.947$, which is quite similar to the results obtained with the training set.

Analysis of the local interaction parameters

The optimal local interaction parameters that we obtained with different hydropathy scales are very strongly correlated. We report in Table II the parameters obtained with the hydropathy scale ZC derived in this work, Eq. (14), which yield the best fit between observed and predicted distributions. This scale is reported in the same table. Results reported in this article refer to this set of parameters, when not otherwise stated.

Figure 3 shows four sets of effective local interaction parameters for strands (E), helices (H), coils (C), and bends (S), respectively. In the figure, we compare the parameters obtained in this work with the parameters $L_{\mathrm{Prop}}(s,a)$ obtained through Eq. (1), which are also influenced by hydrophobic interactions. The two sets of parameters are very strongly correlated, with correlation coefficients always larger than or equal to $r = 0.95$, and nevertheless they perform significantly different in fitting the observed distributions. An exception are the local interactions in the coil state, for which the correlation between the two sets of parameters is only 0.85.

E positions are defined as positions in strands excluding the first and the last one. The strongest preferences for these positions are shown by the amino acids Val, Ile, Thr, Tyr, and Cys, in this order. The same amino acids occur in the first position of strands, with Trp and Lys also favoured at this position, and Trp and Cys favoured at the last position in strands.

We label with H the positions in helices excluding the first four and the last one. For these positions, the order of preferences is Ala, Gln, Arg, Glu, Lys, Leu, and Met. For the first position H1, the most preferred residues are Pro and Trp. For the second position H2, Glu, Asp, and Trp show the strongest preferences. For the third position H3, Glu, Asp, Gln, and Ala are preferred. The position H4 shows local interactions very similar to those of the helical positions Hi with $i > 4$, except that there is an increased propensity for the aliphatic amino acids Leu, Ile, and Val, while Glu is disfavored here. The last position in helices, HY, has a strong preference for Leu, Phe, Met, and Tyr.

C positions are positions without any DSSP code, excluding positions flanking helices or strands; In particular, turns (T) and bends (S) are not considered in this class. The strongest preferences are for Pro, Cys, and Phe. Also Trp, His, and Tyr have a slight preference for coil positions. Some of these are aromatic amino acids, which are seldom observed in C positions, because the exposure of these positions does not fit their large hydrohobicity. The result of our calculation shows that aromatic amino acids in coil positions are more frequent than one would expect based on their hydrophobicity and on the exposure. Our model attributes this higher than expected occurrence to the effect of favourable local interactions. Another possibility is that this higher occurrence is due to functional constraints, since many functional residues occur in coil positions, and they are often aromatic amino acids. These local interactions favoring aromatic amino acids in coil positions represent the most important qualitative difference between the parameters derived from the usual propensity measure and the parameters presented here.

Table II
Local Interaction Parameters and Hydrophobicity Parameters Derived in This Work

	V	I	T	Y	A	Q	E	R	K	L	M	P	C	F	W	H	G	D	N	S
Loc(C)	−0.021	−0.201	0.117	−0.031	0.316	0.282	0.466	0.225	0.285	−0.113	−0.070	−0.809	−0.255	−0.323	−0.237	0.004	0.142	0.069	0.088	0.023
Loc(T)	0.682	0.569	0.374	−0.010	0.242	0.164	0.202	0.156	0.097	0.115	0.113	−0.392	0.095	−0.065	−0.089	−0.126	−0.834	−0.135	−0.458	0.079
Loc(S)	0.321	0.268	−0.061	0.027	0.367	0.225	0.257	0.113	0.119	0.160	0.129	−0.090	0.004	−0.079	0.085	−0.117	−0.632	−0.286	−0.289	−0.135
Loc(0)	1.289	1.127	−0.758	0.674	0.764	0.685	0.703	0.511	0.665	0.818	0.523	−0.728	−0.079	0.694	0.776	−0.125	0.237	−1.022	−0.852	−0.938
Loc(1)	0.098	0.100	0.349	0.050	−0.115	0.184	−0.126	0.106	0.091	−0.062	0.154	−1.127	0.557	0.018	−0.310	0.309	0.315	0.198	0.779	0.153
Loc(2)	0.440	0.355	0.284	0.068	−0.191	−0.093	−0.714	0.090	0.072	0.262	0.204	0.106	0.406	0.081	−0.360	−0.114	0.399	−0.387	0.178	−0.129
Loc(3)	0.162	0.359	−0.009	0.189	−0.115	−0.480	−0.717	0.230	0.138	0.136	0.043	0.754	0.238	0.144	0.173	0.051	0.569	−0.552	0.228	0.168
Loc(4)	−0.236	−0.354	0.330	0.093	−0.528	−0.263	0.292	−0.350	−0.150	−0.384	−0.188	6.443	0.240	−0.050	0.041	0.431	0.899	0.602	0.520	0.416
Loc(H)	0.173	−0.027	0.333	0.124	−0.422	−0.335	−0.303	−0.300	−0.295	−0.203	−0.249	1.740	0.212	0.261	0.200	0.133	0.780	0.222	0.124	0.249
Loc(c)	0.458	0.227	0.179	−0.286	−0.192	−0.138	−0.042	−0.128	−0.150	−0.529	−0.287	9.999	0.060	−0.336	−0.002	−0.177	0.963	0.219	−0.047	0.030
Loc(X)	0.350	0.049	0.538	−0.181	0.011	0.191	0.666	0.163	0.411	−0.212	0.126	9.999	−0.502	−0.458	0.239	−0.267	−0.797	−0.066	−0.273	−0.036
Loc(A)	0.224	0.422	−0.162	0.224	0.077	0.064	0.199	0.066	−0.133	0.528	−0.298	−1.028	0.097	0.297	0.031	0.064	−0.096	−0.056	0.004	−0.051
Loc(B)	−0.444	−0.315	−0.375	−0.229	0.376	−0.020	0.098	−0.102	−0.261	0.271	0.092	1.134	0.085	−0.065	−0.203	−0.079	−0.020	0.725	0.398	0.048
Loc(E)	−0.551	−0.297	−0.315	−0.128	0.124	0.103	0.080	0.023	0.043	0.126	0.213	0.710	−0.126	0.083	0.142	0.038	0.102	0.474	0.337	−0.033
Loc(Y)	−0.328	−0.278	−0.180	−0.161	0.346	0.206	0.265	0.097	0.168	0.167	0.181	0.003	−0.233	−0.152	−0.168	−0.111	0.308	−0.145	−0.025	−0.105
Loc(Z)	0.002	0.097	−0.326	0.169	0.145	0.356	0.501	0.285	0.331	0.027	0.153	0.040	−0.595	0.178	0.187	−0.080	−0.178	−0.421	−0.271	−0.249
ZC	0.897	1.409	−0.400	1.252	−0.086	−0.825	−1.314	−0.326	−1.225	1.430	1.131	−0.616	1.170	1.839	1.824	0.169	−0.844	−1.110	−0.738	−0.772

C, structures not classified by DSSP; T, turns; S, Bends; H0, C positions preceding a helix; Hk, k-th position in the helix; HY, last position in the helix; HZ, C position following a helix; E0, C positions preceding a strand; E1, first position in the strand; EY, last position in the strand; EZ, C position following a strand.

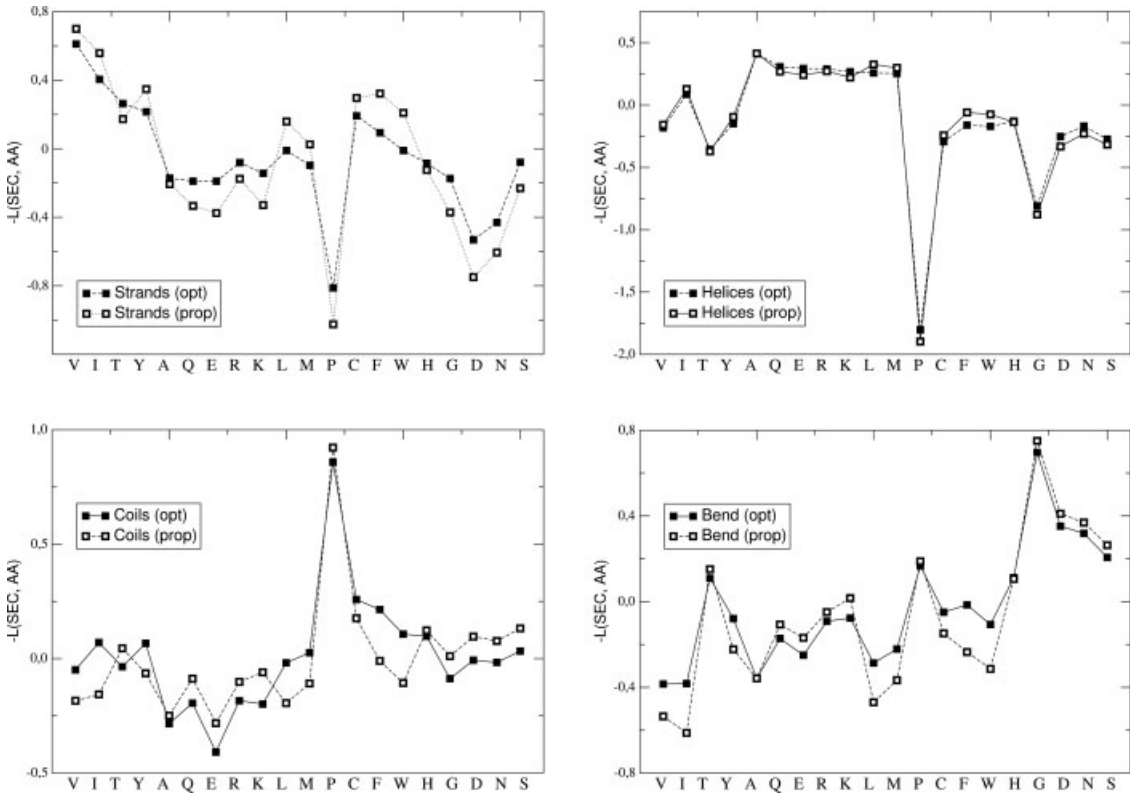

Figure 3

Local interaction parameters, changed of sign so that positive means favoured. The parameters derived in this work are compared with those derived with the propensity measure. Top left, Positions in strands; Top right, Positions in helices; Bottom left, Positions not classified by DSSP; Bottom right, positions in bends.

The small amino acids Gly, Asp, Asn, and Ser show the strongest preferences for positions classified as bends by the DSSP algorithm, labelled with S. Turns (T) positions show very similar local interaction parameters (correlation coefficient $r = 0.86$ with bend interactions), with Gly, Asn, Pro, and Asp the most preferred amino acids. Positions before the helix see a prevalence of the same types of amino acids, namely Asp, Ser, Asn, Thr, and Pro. In positions after helices, Gly, Cys, Phe, Asn, and His prevail. Pro and Met are prevailing before strands, and Cys, Asp, Thr, Asn, and Ser have strong preference to occur after strands.

Structural propensities

As aforementioned, different secondary structure types vary in their propensity for buriedness. In analogy with Eq. (1), these structural propensities can be evaluated as the ratio between the conditional probability to observe a secondary structure s given that the EC component is c, and the overall probability to observe s. Minus the logarithm of the propensity can be interpreted as an effective interaction, expressing the tendency of secondary structure s to occur with EC component c,

$$L_{\mathrm{STR}}(s, c) = -\log\left(\frac{P(s|c)}{P(s)}\right). \qquad (15)$$

The structural propensities are shown in Figure 4. The upper left panel shows the propensities for the secondary structures classified into the three main classes, coil, helix, and strand, as a function of the effective connectivity, which is related to buriedness. As expected, coil propensity decreases almost linearly with increasing buriedness, strands have high propensity to occur in buried positions, and helices have only a slight preference for buriedness over exposure. These propensities are in agreement with previous statistical studies in which the exposure was measured as accessible surface area, and they can be rationalized with simple physico–chemical principles. For instance, residues not in regular secondary structure pay a high enthalpic cost in the interior of the protein, where their polar groups can not form hydrogen bonds with water, and exposed strands pay a high entropic cost with respect to exposed coils.

We investigated these preferences more precisely, distinguishing the secondary structure types more finely. Looking at residues that do not belong to regular second-

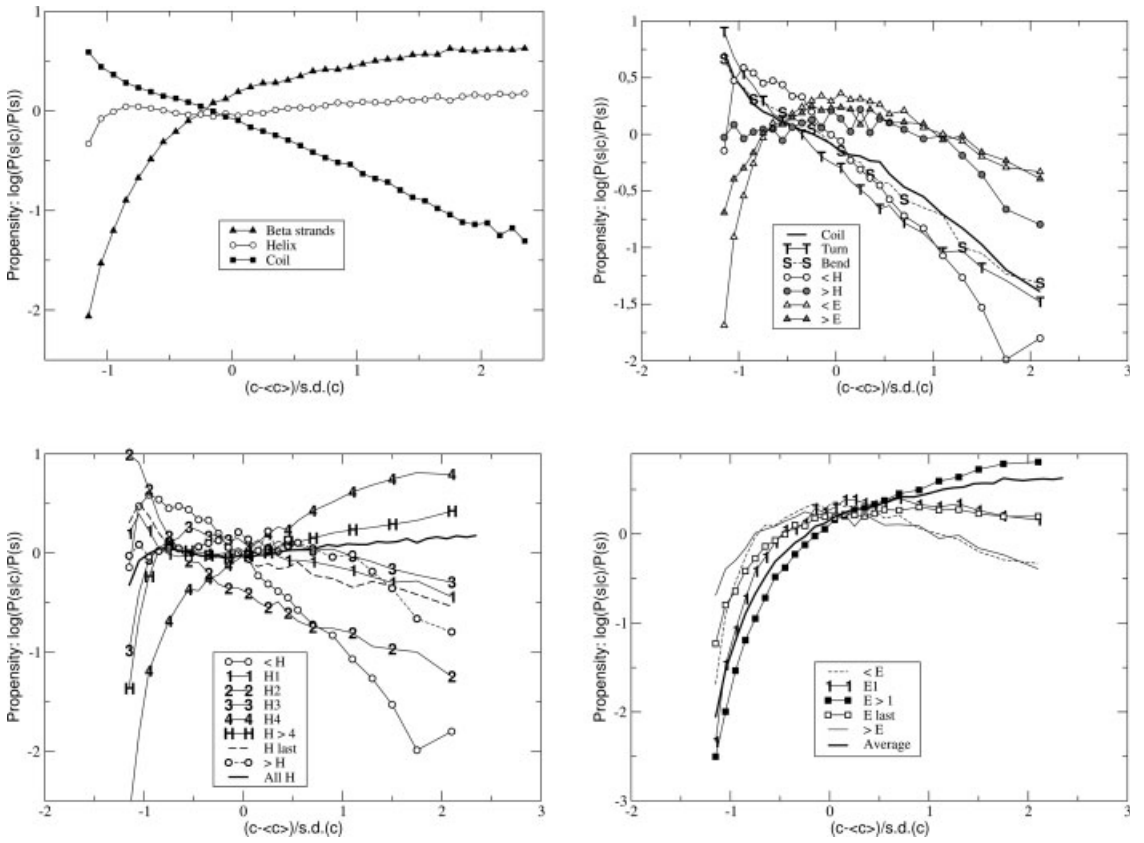

Figure 4

Structural propensities for secondary structure elements. Upper left, Strands, Beta and Coil structures; Upper right, Positions outside regular secundary structure; Lower left, helical positions; Lower right, strand positions.

ary structure elements, we see that turns (T) and bends (S) have a tendency to be exposed more than other coil structures (Fig. 4, upper right panel). As already known, the residue preceding a helix has a strong tendency to be exposed. A similar tendency is also present at the end of the helix, but it is weaker there than for other coil structures. The beginning and the end of a beta strand also tend to be exposed, although much less than other irregular structures, and there is not a significant asymmetry between the beginning and the end, as expected for antiparallel arrangements, for which the beginning of one strand coincides with the end of the other one.

As it is well-known,[40–42] the initial and final positions of helices tend to be located in exposed regions. Consistently, we see in Figure 4, lower left panel, that the first three positions of an helix, denoted as H1, H2, and H3, have high propensity for low EC components (exposed regions), in particular the second position, whereas the fourth position has a strong tendency to be buried, as well as positions after the fourth one, which alternate between more buried and more exposed positions.

Last, if we examine beta strands, we see that the first and last residues along the strand have less tendency to be buried than residues in the middle, and even less for coil residues flanking the strand, see Figure 4 lower right panel. Distinguishing parallel and antiparallel beta bridges, we also see that parallel bridges have a stronger tendency to be buried than antiparallel ones, see Figure 5. This difference is large for positions in the middle and the end of the strand, and small for positions at the beginning of the strand.

Reduced frustration between local and nonlocal interactions

Local interactions favoring different secondary structure types behave in different ways with respect to hydrophobicity. For instance, interactions favoring strands are positively correlated with hydrophobicity ($R = 0.64$), helix interactions are not significantly correlated with it ($R = 0.21$), and both turn and bend interactions are anticorrelated with it ($R = -0.55$). The low correlation between helical parameters and hydrophobicity is consistent with experimental results.[43]

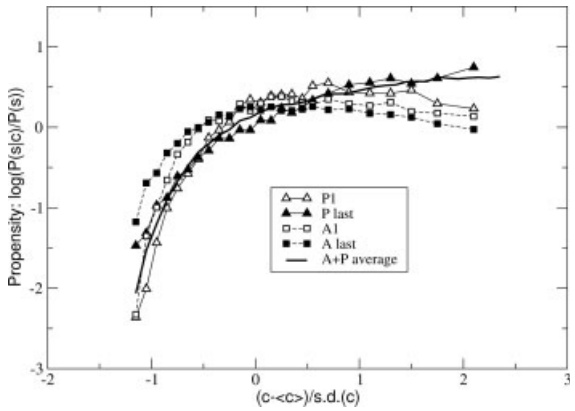

Figure 5

Structural propensities for positions in beta strands. Parallel (P) and antiparallel (A) positions are compared. Notice that parallel strands have larger propensity for buried positions (large effective connectivity c).

These correlations between local and nonlocal interactions, in concurrence with the structural propensities of secondary structures for buriedness or exposure that we discussed earlier, favor the evolutionary optimization of local interactions and hydrophobicity at the same time. For instance, at buried strand positions, amino acids like Val and Ile optimize both local and hydrophobic interactions at the same time. Similarly, at exposed turns or bends, residues as Gly or Asn optimize both kinds of interactions.

In general, we observed that there is a positive correlation between the structural propensity for buriedness of a secondary structure type, that we quantify through the average effective connectivity of this secondary structure, and the correlation between local interactions favoring that secondary structure type and hydrophobicity. We show this in Figure 6, in which we represent for each secondary structure type s the correlation between local and hydrophobic interactions, $r[h(a), L(s,a)]$, versus the average effective connectivity, $\sum_c c\, P(c|s)$.

The fact that the most frequent combinations of secondary structure and buriedness favor the optimization of both kinds of interactions is consistent with the "principle of minimal frustration" in protein energetic enunciated by Bryngelson and Wolynes,[44] and with the "consistency principle" of Go.[45]

Performances in fold recognition

We tested the ability of the effective free energy with optimal local interactions to recognize the native structure of a protein against alternative structures generated from the PDB. We used the numerical values of the contact interactions $U(a,b)$ reported in Ref. 12. Since their energy unit is arbitrary, we multiplied the local interactions $L(s,a)$ times a numerical coefficient $\varepsilon > 0$, which is needed to convert the two sets of energy parameters to the same unit. We consider therefore the energy function

$$E(\mathbf{C},\mathbf{S},\mathbf{A}) = \sum_{i<j} C_{ij} U(A_i, A_j) + \varepsilon \sum_i L(s_i, A_i) + F_{\mathrm{str}}(\mathbf{C},\mathbf{S}) + F_{\mathrm{seq}}(\mathbf{A}). \quad (16)$$

In this way, if we set $\varepsilon = 0$, the energy function lacks local interactions In the following, we will fix the value of ε by optimizing the results of the threading test.

Gapless threading

Our first test involved gapless threading. We used the entire PDB90 database, containing 7494 chains with less than 90% sequence identity. We eliminated from the test set the structures determined by NMR, nonglobular proteins, and membrane proteins (see Materials and Methods, Database section), remaining with 5371 protein chains. The entire database was used to generate decoys, yielding from 84,000 decoys for the longest chains up to 1.5 millions for the shortest ones, a very large number that makes this test rather challenging.

Using only contact interactions ($\varepsilon = 0$), for 92.6% of the protein chains the native structure has the lowest energy among all decoys, with an average normalized energy gap equal to 0.543. Most of the 397 chains that are not recognized belong to multi-chain proteins, so that the failure of recognition may indicate that they are not stable in the isolated form. Only 83 single-chain proteins are not recognized.

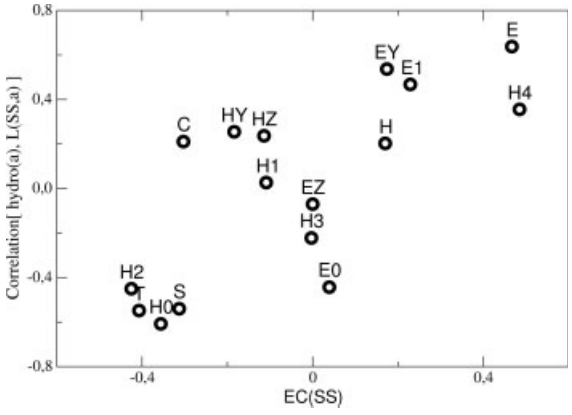

Figure 6

Horizontal axis: Average effective connectivity (EC) conditioned to a secondary structure type, $\sum_c cP(c|s)$, which is related to the propensity for buriedness. Vertical axis: correlation coefficient between local interaction parameters and hydrophobicity, $r[h(a),L(s,a)]$. Each point represents a secondary structure type. Notice that secondary structures that tend to be more buried, like E and H4, are stabilized by local interactions that are positively correlated with hydrophobicity, and secodary structures that tend to be more exposed are stabilized by local interactions that are negatively correlated with hydrophobicity.

We then included in the model the local interaction parameters determined by propensity, Eq. (1). This improves the performances of the energy function up to 94.5% recognition, with only 57 single chain proteins that are not recognized, and a normalized energy gap of 0.64, for $\varepsilon = 0.50$. Finally, we used the local interaction parameters determined in this work. The best recognition rate was obtained with a coefficient $\varepsilon = 0.60$, yielding 94.3% recognition, with only 58 single chain proteins that are not recognized, and a normalized energy gap of 0.66. The performances of both local interaction parameter sets are thus very similar concerning the recognition, where they yield a small but significant improvement with respect to using only contact interactions. The improvement of the normalized energy gap is more significant, and the parameters derived in this work perform better than the propensity parameters under this respect.

Gapped threading

Then, we tested the ability of the effective folding free energy to recognize structures similar to the native by assigning them a high stability score. In this test, pairs of proteins were structurally aligned with the program Mammoth,[46,47] and to each structural alignment we assigned a stability score based on the effective free energy function. This test is expected to be much more challenging than the previous one, because of three reasons. (1) Decoys generated through optimal structural alignments with gaps are in general more challenging than decoys generated through gapless alignment. (2) The alignment length is not constant, since the trial structure may be aligned to a fragment of the query protein, which may be either a complete domain or a supersecondary structure. (3) Instead of the native structure, the energy function must recognize a similar structure that may have quite significantly diverged, thus being less fit to the query sequence than the native one.

Because of point (2), it is necessary to normalize the stability score derived from structural alignments in such a way that it depends at little as possible on the length and amino acid composition of the aligned fragment of the query protein, so that a fragment of the query sequence can be matched to a structural domain. In contrast, in gapless threading trial structures are aligned to the entire sequence, and therefore the effective free energy itself is a convenient stability measure.

We tried different normalizations for comparing alignments of different length. The best results were obtained with a stability score which estimates the Z score of the effective energy. The Z score is calculated by substracting from the energy of the alignment the average energy of compact structures with arbitrary contact matrix and secondary structure. This reduces the influence of the amino acid composition on the stability score, by taking into account that more hydrophobic sequences tend to have lower energy both for the native structure and for generic compact structures. The resulting difference is divided by the standard deviation of the energy of unrelated structures, in order to make the stability score of unrelated structures as little as possible dependent on the size of the alignment.

The use of the Z score as a stability score was proposed several years ago by Goldstein *et al.*,[48] inspired to the random energy model (REM)[49] applied to heteropolymers.[44,50] In this approach, the mean and mean square energy is computed by sampling generic compact structures from the PDB, for any given fragment of a given length, as Govindarajan and Goldstein did for structures on the lattice.[51] Nevertheless, to simplify the calculations, we neglected the correlations between the energies of different contacts and local energies of different residues, and focused on the mean energy of an individual contact interaction and an individual local interaction. This provides an analytic estimate of the Z score as

$$S(\mathbf{A},\mathbf{C},\mathbf{s},\mathbf{a}) = -\frac{\sum_{i<j} C_{a_i a_j} U(A_i,A_j) - N_C\langle\langle U\rangle\rangle + \varepsilon\sum_i (L(s_{a_i},A_i) - \langle\langle L(s,A_i)\rangle\rangle)}{N \times N^{-\gamma}\sqrt{\sigma_U^2 + \frac{\varepsilon^2}{N_C}\sum_i \sigma_L^2(A_i)}}. \qquad (17)$$

In the above equation, $\mathbf{a} = \{a_{i_1}\cdots a_{i_N}\}$ denotes the alignment of a fragment of the query sequence **A** with the trial structure. The length of the alignment N is calculated from the first i_1 to the last i_N residue in the query sequence aligned to some residue of the template structure. The energy of the alignment is calculated in the assumption that contacts and secondary structure are perfectly conserved from the template (t) to the query (q), that is $C_{ij}^{(q)} = C_{a_i a_j}^{(t)}$ and $s_i^{(q)} = s_{a_i}^{(t)}$. $N_C = \sum_{i<j} C_{a_i a_j}^{(t)}$ indicates the number of contacts in the template structure, which is proportional to the number of residues N, so that N_C/N tends to a finite limit as N grows. Residues in the sequences aligned to gaps are assigned zero contacts and coil secondary structure. The results improve if short range contacts are not taken into account in the calculation of the energy. Specifically, we omit all con-

tacts between amino acids separated by less than four amino acids along the sequence, imposing $C_{ij} = 0$ if $|i - j| < l_{\min} = 4$.

Double angular brackets $\langle\langle\cdot\rangle\rangle$ indicate the average over alternative compact structures. In particular, $\langle\langle U\rangle\rangle$ indicates the mean energy of an individual contact for the set of unrelated structures, estimated as explained in Materials and Methods, Stability score section and $\langle\langle L(s,a)\rangle\rangle \approx \sum_s P(s)L(s,a)$ is the mean local energy of amino acid a over a generic secondary structure type. σ_U^2 and σ_L^2 are the variance of the energy of an individual contact and of an individual local interaction, respectively (see Database section for further explanations).

The term $N^{-\gamma}$, with $\gamma = 0.6 - 0.8$ yielding the best recognition, takes into account that the standard deviation of the energy per residue, E/N, decreases with the alignment length, since the correlation between contacts and between secondary structure types vanishes at large sequence distance. To test this, we measured the energy per contact, $\sum_{i<j} C_{ij}\ U(A_i,A_j)/N_C$, obtained aligning a query protein with unrelated structures. As expected, we found that the mean value of this quantity is not correlated with the alignment length, and its standard deviation decreases with the alignment length or the number of contacts, compatible with $N^{-0.5}$, as one would predict neglecting the correlations between contacts separated by a large sequence distance. Therefore, the Z score of the contact energy per contact can be estimated as $Z_{E_C/N_C} \approx \sqrt{N_C}(\sum_{i<j} C_{ij}U(A_i,A_j)/N_C - \langle\langle U\rangle\rangle)/\sigma_U$. This quantity, multiplied times $\sqrt{N_C/N}$, is equal to Eq. (17) with $\varepsilon = 0$ and $\gamma = 0.5$, and it is expected to be uncorrelated with the length of the alignment, so that the stability score of unrelated structures does not have any trivial scaling with the alignment length.

The local energy per residue, $\sum_i L(s_i,A_i)/N$, has also a distribution that becomes narrower with the length. Its standard deviation decreases as $N^{-0.70\pm0.03}$, that is faster than $1/\sqrt{N}$. The results presented were obtained with the exponent $\gamma = 0.75$, which provides a good recognition for the complete free energy function.

We used as a test set a randomly chosen subset of the ASTRAL40 database, containing 480 domains with less than 40% sequence identity. This threshold on sequence similarity guarantees that pairs of homologous proteins are significantly diverged both in sequence and in structure. All pairs of proteins were structurally aligned with the program Mammoth,[46,47] and for each pair the stability score was calculated using one and the other protein alternatively as query sequence and trial structure. Three types of protein pairs were distinguished:

1. *Native*, if sequence and structure belong to the same protein.
2. *Similar*, if there is a significant structural similarity between the two proteins or they have the same SCOP classification[2] at the fold level (in most cases the two conditions coincide). We use the Mammoth Z score[46,47] and the contact overlap to quantify structural similarity.
3. *Unrelated*, otherwise.

There are 250 proteins of 480 (52%) with at least one match with Mammoth Z score larger than 8 and contact overlap multiplied times $\sqrt{N}$ larger than 2.5. We measured the percentage of these 250 proteins for which at least one similar structure is present among the m most stable alignments. The recognition was considered successful if the trial structure and the query protein either are classified in the same SCOP fold, or they have Mammoth Z score larger than six, a very significant similarity which in many cases guarantees that the proteins share the same fold; otherwise, it was considered failed.

We first present results for the case where we require at least one structure with Z score of Mammoth larger than eight to be present in the set of similar structures. The dependence on the similarity threshold will be discussed later.

Without local interactions, the fraction of proteins for which similar structures appear within the m most stable ones is 50.4%, 76.0%, and 83.2% for m equal to 1, 5, and 10, respectively, and the recognition of the native structure is 98.3%.

The recognized fractions significantly increase when local interactions are included into the model. With the local interaction parameters derived from propensities, the maximum recognized fractions are 60.0%, 80.8%, and 87.2% for m equal to 1, 5, and 10, respectively, and 99.8% for the native structure. We find the best results for ε in the range [0.4,0.6], depending on the value of m. Giving more importance to smaller m, a good choice appears to be $\varepsilon \approx 0.5$, but there is not a well defined optimum value.

The local interactions derived in this work yield best results for ε in the range [0.4,0.6]. A good choice is $\varepsilon = 0.45$, a value close to the one obtained with gapless threading, but also in this case there is not any well defined optimum value. The maximum recognized fractions are 60.8%, 82.0%, and 88.0% for m equal to 1, 5, and 10, respectively, and 100% for the native structure. For all values of m, the recognition of similar but not identical structures improves at least 5% considering local interactions besides contact interactions. In contrast, the recognition of the native structure improves only by 1%.

Figure 7 shows the performances of the energy function for recognizing structures similar to the native. Local interaction parameters derived from inverse folding and derived from propensity are compared. In both figures, the similarity threshold is set at Z score of Mammoth larger than eight. In the left plot, recognition at rank 1 is shown as a function of ε. In the right plot, the best recognition for any ε is shown versus the rank m (the recognition is considered successful if a similar structure has

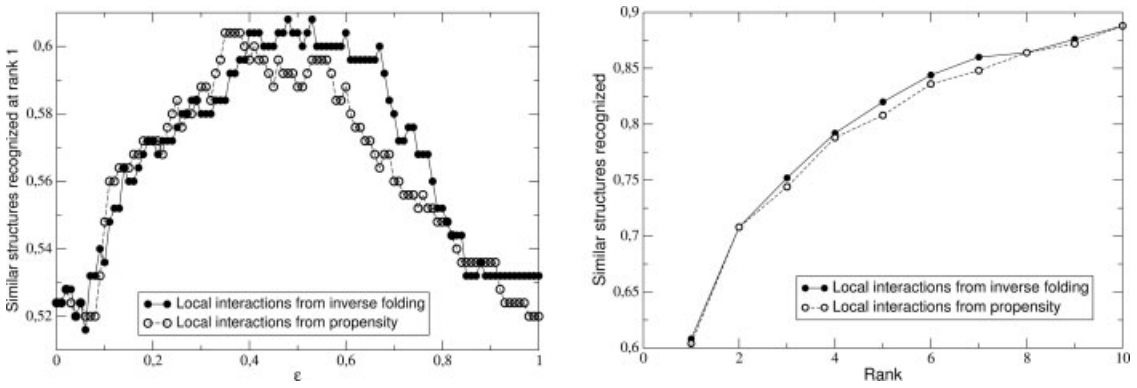

Figure 7

Left: Fraction of proteins for which at least one structure similar to the native has stability score within rank 1, as a function of the weight of local interactions, ε. Local interaction parameters derived from inverse folding are compared to parameters derived from secondary structure propensity. Right: Best fraction of recognized proteins for any ε as a function of the rank of the recognition m. The recognition is considered successful if at least one similar structure is found within the m most stable ones. For both plots, test proteins must have at least one structure with Mammoth Z score larger than eight.

one of the m highest stability scores). For most values of m, the local interaction parameters derived from inverse folding perform equally to or slightly better than the parameters derived from amino acid propensities. The difference, however, is small and it is difficult to assess its significance.

The recognition improves if the structural similarity increases. Figure 8 shows, for the rank $m = 1$, the best recognized fraction for any ε as a function of the mimimum Z score of Mammoth. Only proteins for which there is at least one match above this similarity threshold are considered in the test set. The number of tested proteins decreases from 274 for $Z > 6$ to 232 for $Z > 10$. Correspondingly, the fraction of tested proteins for which a similar structure is recognized at rank 1 increases approximately linearly from 55.8% to 64.2%.

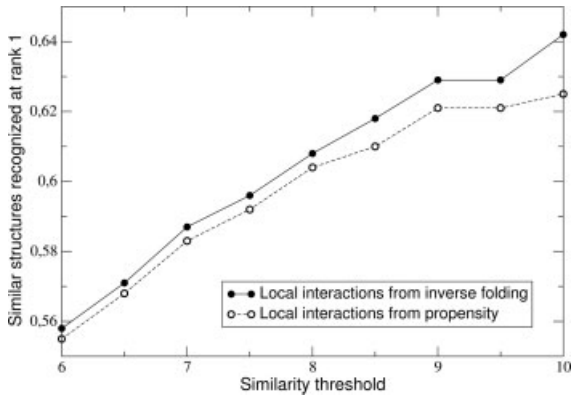

Figure 8

Best recognition for any ε and for rank m = 1 as a function of the similarity threshold, the minimum Z score of Mammoth above which two structures are considered very similar. Local interaction parameters derived from inverse folding are compared to parameters derived from secondary structure propensity.

CONCLUSIONS

In this article, we have extended an analytically solvable model of inverse folding based on mutations and on hydrophobicity,[6–8] by including in the folding model the local interactions that favor distinct secondary structures.

The local interaction parameters were obtained by fitting the complete model to site-specific amino acid distributions sampled from structurally equivalent positions in the PDB. This allows to disentangle the influence of local and nonlocal interactions on the propenstity of amino acids for secondary structures,[13] which in most approaches is used as the basis to derive such local interaction parameters through a statistical analysis.[14–18] The energetic interpretation of helical propensities is based on side chain entropy loss upon helix formation[33] and enthalpy of the helix-coil transition,[34] which are interactions that are local along the protein chain.

We have compared our local interaction parameters for the interior of helices, H, with previously determined scales computational and experimental scales of alpha helix propensities, finding high correlation coefficients $r = 0.79$ with the Chou and Fasman scale,[13] $r = 0.87$ with the Luque *et al.* scale,[52] $r = 0.89$ with the Williams scale,[53] $r = 0.92$ with the scale by Rohl *et al.*,[54] $r = 0.93$ with the scale by Pace and Scholtz,[55] and $r = 0.94$ with the scale by Muñoz and Serrano, used in the program AGADIR.[14] The two last scales were determined experimentally, considering exposed positions in the interior of the helix.

The inverse folding model with maximum likelihood parameters provides an excellent fit of site-specific amino

acid distributions sampled from the PDB, considerably improving the previous model that lacked local interaction parameters. A BIC test demonstrates that the numerous parameters of the model are very significantly determined.

The local interactions determined in this work have been combined to a contact free energy function, and the resulting effective potential was tested for its ability to recognize the native structure and structures similar to the native on the basis of the effective energy alone. Notice that, as most effective free energy functions used in threading, our function can be applied only on protein-like structures derived from the PDB, since it does not take into account steric interactions, peptidic geometry, and hydrogen bonding, which are assumed to be already protein-like in the tested structures.

The local interactions have only a minor influence for the recognition of the native structure, a task for which the contact free energy is already quite good, with a success rate of roughly 98% that becomes 99.8% with the addition of local interactions. These results do not depend significantly on whether alternative structures are generated through gapless threading or through structural alignment with gaps, the latter creating more challenging decoys, and on whether the stability is evaluated through the effective energy itself or from an estimate of its Z score, Eq. (17), which reduces the effect of the sequence length and composition on the estimated stability.

In contrast, the task of recognizing structures similar to the native but evolutionarily diverged from it is much more difficult. In this case, local interactions improve significantly the recognition, from 50 to 60%. Results improve considerably if we are less strict and consider all structures within a rank m of stability, from 60% for rank $m = 1$ to 89% for rank $m = 10$. Therefore, the local interactions have a small influence on the recognition of the native structure, whereas they improve quite significantly the recognition of similar structures.

Similar structures in the ASTRAL40 database, having less than 40% sequence identity, are characterized by contact overlap with the native structure typically smaller than 0.5. This implies that the contact energy deviates considerably with respect to the energy of the native structure. It is unlikely that a structural similarity at this level implies a strong similarity of contact energies.[56] In contrast, the secondary structure is much more conserved between similar structures, which explains why the local interaction energy plays an important role in the recognition of similar structures, despite its role in the recognition of the native structure is almost negligible.

Consistently, the recognition improves considerably if one imposes more strict structural similarity, measured by the Z score of the program Mammoth. For pairs with Mammoth's Z score larger than eight the recognition at rank 1 is 60.4%, and for the native structure itself the recognition reaches 99%. Therefore, a very high level of similarity, much above statistical significance, is needed for discriminating similar structures from unrelated structures of high stability. Hence, despite the significant improvement presented in this article, further progress is needed in order to achieve the important goal of fold recognition solely through folding stability.

MATERIALS AND METHODS

Effective connectivity

We represent the global characteristics of a protein structure with N residues through the contact matrix C_{ij}, a binary matrix indicating whether sites i and j are in contact ($C_{ij} = 1$) or not ($C_{ij} = 0$). A contact is defined by at least one pair of heavy atoms belonging to two different amino acids closer than 4.5 Å. The EC is the vector c_i that maximizes the quadratic form $Q = \sum_{ij} C_{ij} c_i c_j$ with the constraints $\langle c_i \rangle = 1$ and $\langle c_i^2 \rangle = \langle n_c^2 \rangle / \langle n_c \rangle$, where angular brackets denote here the average over protein sites and $\langle n_c \rangle$ and $\langle n_c^2 \rangle$ are the mean and mean squared number of contacts.

The EC can be computed from the eigenvectors of the contact matrix through the formula[8]

$$c_i = \frac{\sum_\alpha \frac{v_i^{(\alpha)}}{\Lambda - \lambda_\alpha}}{\sum_\alpha \frac{\langle v^{(\alpha)} \rangle}{\Lambda - \lambda_\alpha}}, \qquad (18)$$

where $v_i^{(\alpha)}$ denotes the i-th component of the α-th eigenvector of the contact matrix, λ_α is the corresponding eigenvalue, and the parameter Λ is implicitly defined by the constraint $\langle c_i^2 \rangle = \langle n_c^2 \rangle / \langle n_c \rangle$ For single domain globular proteins, the EC is parallel with the principal eigenvector (PE), corresponding to the largest eigenvalue, but whereas the PE components strongly depend on the protein size N (they scale as $1/\sqrt{N}$), c_i is almost independent of protein size.

Secondary structures

We represent local protein characteristics through the secondary structure, determined through the DSSP program.[31] We consider four discrete secondary structure states distinguished by DSSP, E (extended, that is beta sheet), H (helix), T (turn) and S (bend). Beta bridges are here classified as strands. Positions not labeled by DSSP are considered coils (C). We also distinguish the first four helical position H1 (the position where N-caps are found), H2, H3, H4, and the last position of the helix, HY. Positions after the fourth and before the C-terminus are clustered together (H). We also treat separately the C positions preceding and following the helix, which we call H0 and HZ, respectively. Therefore, an helix is coded as H0-(H1, H2, H3, H4, H, . . . , HY)-HZ. For strands, we distinguish the first position (E1), the last one (EY), and positions preceding and following the strand (E0 and EZ, respectively), describing each strand as E0-(E1, E , . . . , EY)-EZ.

Inverse folding model without local interactions

We review here for completeness the prediction of site-specific amino acid distributions for a model in which the folding free energy is represented as a contact free energy function, Eq. (5), without local interactions, that is $\varepsilon = 0$. This analytic prediction is in good agreement with simulations of the SCN model (see Ref. 8). We indicate site-specific distributions as conditional distributions of amino acids given the structure at site i, using the notation $P(A|B)$.

Our starting point is Eq. (2), expressing the predicted site-specific average hydrophobicity. We assume that this is the only condition representing selection for folding stability. We first consider equiprobable mutations from any amino acid to any other one, and assume that the resulting site-specific amino acid distributions have maximal entropy for average hydrophobicity given by Eq. (2). This implies that they are exponential, or Boltzmann, distributions of the form $P(a|c_i) \propto \exp(-\beta(c_i)h(a))$.[23] This distribution can be interpreted as the stationary distribution of a site-specific substitution processes, consisting of (1) equiprobable mutations; and (2) a site-specific selection process, with acceptance probability of a mutation from amino acid a to b given by

$$P_{\mathrm{sel}}(a, b|c_i) = \min(1,\ \exp[-\beta_i(h(b) - h(a))]). \qquad (19)$$

This selection process fulfils detailed balance (also called reversibility in molecular biology), that is $P(a|c_i)P_{\mathrm{sel}}(a,b|c_i) = P(b|c_i)P_{\mathrm{sel}}(b,a|c_i)$

We then consider a mutation process at the DNA level, which consists of identically distributed and independent mutations at each nucleotide site. We assume that the mutation process fulfils detailed balance. Therefore, its stationary properties depend only on the stationary frequencies of the four nucleotides, $f(n)$, with $n \in \{\mathrm{A,T,G,C}\}$. The stationary frequency of amino acid a under mutation alone is the sum of the frequencies of each codon $\mathbf{n}$ coding for a, which in turn are obtained as the product of the frequencies of their three nucleotides n_1, n_2, n_3. Combining the mutation and the selection process, we obtain the following amino acid distribution

$$P(a|c_i) \propto w_{\mathrm{mut}}(a)\ \exp(-\beta(c_i)h(a)) = \sum_{\mathbf{n}:\mathcal{A}[\mathbf{n}]=a} f(n_1)f(n_2)f(n_3)\ \exp(-\beta(c_i)h(a)), \qquad (20)$$

where $\mathcal{A}[\mathbf{n}]$ indicates the amino acid coded by the codon $\mathbf{n}$ It is easy to see that this stationary distribution fulfils detailed balance with respect to both the mutation and selection process.

Computation of site-specific distributions

Our algorithm for predicting site-specific amino acid distributions according to Eq. (8) proceeds in three steps. First, we obtain from the PDB the mean and mean square hydrophobicity, $\langle [h] \rangle = \sum_i w_i [h_i]$ and $\langle [h]^2 \rangle = \sum_i w_i [h_i]^2$. Here $[h_i]$ denotes the average hydrophobicity at site class i characterized by EC c_i, and w_i is the number of sites that fall in site class i. Second, using the observed parameters $\langle [h] \rangle$ and $\langle [h]^2 \rangle$ we calculate from Eq. (11) the predicted average hydrophobicities $[h_i^{\mathrm{pred}}]$ at each site class. Third, for any value of c_i we solve numerically Eq. (10), computing its left hand side using the observed conditional distribution of secondary structure type given the EC, $P(s|c)$, the local interaction parameters, and trial values of β. The value of β for which this l.h.s. coincides with the previously computed value of $[h_i^{\mathrm{pred}}]$ up to numerical accuracy yields the predicted value of $\beta(c)$ at that site. To speed up this computation, we take advantage from the fact that Eq. (11) is a monotonic function of $\beta(c)$.

Parameters of the model

The parameters of the inverse folding model are:

1. *Hydrophobicity scale.* We tested 11 hydrophobicity scales from the literature, listed below: (1) The KD82 hydropathy scale, derived to identify trans-membrane helices using diverse experimental data[37]; (2) The L76 hydropathy scale, which was derived by using experimental data and theoretical calculations[57]; (3) The R88 hydropathy scale, which is based on the transfer of solutes from water to alkane solvents[36]; (4) The augmented Whilmey-White (WW01) hydropathy scale, derived to improve recognition of trans-membrane helices[58]; (5) The G98 classification of amino acids into polar, hydrophobic, and amphiphylic classes, adopted by Gu *et al.*[59] to investigate the relationship between the hydrophobicity of a protein and the nucleotide composition of the corresponding gene; (6) The MP78 hydropathy scale, derived from statistical properties of globular proteins[60]; (7) The AV hydropathy scale, derived by averaging 127 normalized hydropathy scales published in the literature[39]; (8) The FP83 hydropathy scale, derived from the experimental measurement of octanol/water partition coefficients[38]; (9) The ZZ04 scale, also called buriability, proposed by Zhou and Zhou[35]; (10) The interaction scale IH, obtained from the main eigenvector of the interaction matrix $U(a,b)$ used in this work[23]; (11) The optimized interactivity scale, or connectivity scale CH, which maximizes the correlation with the principal eigenvector of protein contact matrices for a non-redundant set of PDB structures.[23]

   In this work, we derived a new hydrophobicity scale, Eq. (14), indicated here by the symbol ZC.

   The hydrophobicity scales can be normalized in such a way that they have mean value zero and variance equal to one without changing the distribution.

Therefore, the number of independent parameters is 18 for each scale. Note that they are not free parameters.

2. *Mean hydrophobicities.* The model needs two parameters for mean and mean square hydrophobicity, $\langle [h_i] \rangle$ and $\langle [h_i]^2 \rangle$. Here $[h_i]$ is the average hydrophobicity of all sites having EC in a narrow range around c_i. These parameters were calculated from the sampled data as

$$\langle [h_i] \rangle = \sum_c P(c) \left(\sum_a h(a) P(a|c) \right), \tag{21}$$

$$\langle [h_i]^2 \rangle = \sum_c P(c) \left(\sum_a h(a) P(a|c) \right)^2. \tag{22}$$

3. *Local interaction parameters.* They are $20 \times S$ parameters, where S is the number of secondary structure types, but only $19 \times (S - 1)$ are independent, since we impose the $20 + S$ constraints

$$\sum_a P(a) \ \exp[-L(s,a)] = 1, \tag{23}$$

$$\sum_s P(s) \ \exp[-L(s,a)] = 1. \tag{24}$$

These conditions are automatically fulfilled by the local parameters obtained through the amino acid propensity of secondary structure, Eq. (1), since it holds $\sum_s P(s)[P(a|s)/P(a)] = 1$ and $\sum_a P(a)[P(a|s)/P(a)] = 1$.

4. *Nucleotide frequencies.* The mutation model depends on four nucleotide frequencies, of which only three are independent, since their sum must be one.
5. *Amino acid frequencies.* In alternative to the nucleotide frequencies, we considered here the amino acid frequencies $f(a)$ as free parameters. Imposing the condition $\sum_a f(a) = 1$ leaves us with 19 additional parameters.

Thus, for $S = 16$, we get a total of $(S - 1) \times 19 + 3 = 288$ free parameters if we use nucleotide frequencies, and $S \times 19 = 304$ parameters if we use amino acid frequencies, plus 20 hydrophobicity parameters obtained from the literature or calculated.

Database

We considered proteins of known structures with less than 50% sequence identity. Nonglobular proteins, as judged on the basis of a cut-off on the number of contacts C, $C/N < 3.5 + 7.8N^{-1/3}$, were excluded. The $N^{-1/3}$ term comes from surface to volume scaling in globular proteins, see for instance Ref. 12. We filtered out membrane proteins imposing a cut-off value of the average hydrophobicity $\langle h \rangle < 0.165$ (with the IH hydropathy scale). Sites with very large or very small values of the EC were also excluded.

Data were divided into 36 bins of EC values c_i, ranging from -1.2 to 2.4, and 16 secondary structure types. This amounts to $36 \times 16 = 576$ structural classes, to which correspond 10,944 independent probability values. The total number of residues constituting our data set was 932,006, on the average more than 90 per each bin, which allows to estimate reasonably well the probabilities, at least for the most populated sets of parameters.

Parameters optimization

We optimized the model parameters through gradient ascent. This simple optimization scheme was fast and accurate enough, since runs starting from different initial conditions converged after some hundreds of steps to quite similar parameter sets.

The score optimized is the sum of the log-likelihood of the observed number of amino acids for each structural class given the model and its parameter,

$$\mathcal{L} = \sum_i \log\left[P\left(n_{i,a}|\pi_{i,a}\right)\right], \tag{25}$$

where $n_{i,a}$ is the observed number of residues of type a at site class i and $\pi_{i,a}$ is the predicted amino acid frequency. The likelihood $\mathcal{L}$ is calculated through a multinomial model as

$$\begin{aligned} \mathcal{L} &= \sum_i \log\left[P\left(n_{i,a}|\pi_{i,a}\right)\right] \\ &= \sum_i \left[\sum_a n_{i,a} \log\left(\pi_{i,a}\right) - \sum_a \log\left(n_{i,a}!\right) + \log(N_i!) \right], \end{aligned} \tag{26}$$

where $N_i = \sum_a n_{i,a}$ is the total number of residues at site class i and the symbol $n!$ denotes the factorial of n. The log-likelihood depends on the number of elements in each site class, N_i. To compare data sets with different distributions of residues per site class, we normalized the score by dividing it by the absolute value of the maximum log-likelihood, which is attained when $\pi_{i,a} = n_{i,a}/N_i$. When N_i is large, Stirling's formula shows that the maximum log-likelihood for amino acid distributions scales approximately as $-(19/2) \times \log(N_i)$. The normalized score, which is reported in the text, is

$$l = \frac{\sum_i \mathcal{L}_i - \mathcal{L}_{i,\max}}{\sum_i \left|\mathcal{L}_{i,\max}\right|} \tag{27}$$

$$\begin{aligned} \mathcal{L}_{i,\max} = &\sum_a \left(n_{i,a} \ \log\left(n_{i,a}\right) - \log\left(n_{i,a}!\right) \right) + \log(N_i!) \\ &- N_i \ \log(N_i) \end{aligned} \tag{28}$$

We also calculated the weighted mean correlation coefficient $r[n_{i,a},\pi_{i,a}]$ of the observed and the predicted distribution at each site class, weighted with the square root of the number of residues in the structural class,

$$R = \sum_i \frac{\sqrt{N_i}}{\sum_k \sqrt{N_k}} r\left[n_{i,a}, \pi_{i,a}\right]. \tag{29}$$

Bayes information criterion

To compare the log-likelihood of models with a different number of free parameters N_{PAR}, we use the BIC. The BIC penalizes models with too many free parameters with respect to the data (overfitting), correcting the log-likelihood as

$$\mathrm{BIC} = -2\mathcal{L} + N_{\mathrm{PAR}} \times \ln(N_{\mathrm{DATA}}), \tag{30}$$

where N_{DATA} is the number of data fitted. A model is considered better the smaller the BIC.

Our data set consists of 576 (36 × 16) amino acid distributions, each giving 19 numbers, which results in $N_{\mathrm{DATA}} = 576 \times 19 = 10{,}944$, so that $\mathrm{BIC} = -2\mathcal{L} + 9.3N_{\mathrm{PAR}}$.

Stability score

The effective energy that we use in this work is the sum of contact and local interactions,

$$E(\mathbf{C},\mathbf{S},\mathbf{A}) = \sum_{i<j} C_{ij}U(A_i,A_j) + \varepsilon \sum_{i,s} \delta(s,s_i)L(s,A_i). \tag{31}$$

The recognition improves if we exclude from this calculation very short range contacts, considering only those with $|i-j| \geq l_{\min} = 4$.

The stability score defined in Eq. (17) estimates the Z score[48] of the effective energy of the tested structure, defined as

$$Z = \frac{E(\mathbf{C}^{\mathrm{test}},\mathbf{A}) - \langle\langle E(\mathbf{C},\mathbf{A})\rangle\rangle}{\langle\langle E(\mathbf{C},\mathbf{A})^2\rangle\rangle - \langle\langle E(\mathbf{C},\mathbf{A})\rangle\rangle^2} \tag{32}$$

where the double angular brackets $\langle\langle\cdot\rangle\rangle$ denote average over the set of alternative compact configurations for sequence **A**. We estimate the average energy as

$$\begin{aligned}\langle\langle E_{\mathrm{C}}(\mathbf{C},\mathbf{A})\rangle\rangle &= \sum_{i<j}\langle\langle C_{ij}\rangle\rangle U(A_i,A_j)\\ &\approx N_{\mathrm{C}}^{\mathrm{test}} \sum_{i<j} w(|i-j|)U(A_i,A_j) = N_{\mathrm{C}}^{\mathrm{test}}\langle\langle U\rangle\rangle,\end{aligned} \tag{33}$$

$$\begin{aligned}\langle\langle E_{\mathrm{L}}(\mathbf{S},\mathbf{A})\rangle\rangle &= \varepsilon\sum_{i,s}\langle\langle\delta(s,s_i)\rangle\rangle L(s,A_i) \approx \varepsilon\sum_{i,s}P(s)L(s,A_i)\\ &= \varepsilon\sum_i\langle\langle L(s,A_i)\rangle\rangle.\end{aligned} \tag{34}$$

In Eq. (33), we make two simplifying assumptions: (1) Alternative compact structures have the same number of contacts $N_{\mathrm{C}}^{\mathrm{test}}$ as the test structure; (2) The probability $P(C_{ij})$ to observe contact C_{ij} in the ensemble of alternative structures depends only on $|i-j|$, that is $P(C_{ij}) = w(|i-j|)$. In Eq. (34), we assume that the probability to observe secondary structure type s at site i, $\langle\langle\delta(s,s_i)\rangle\rangle$, does not depend on i. These assumptions are necessary to overcome the limitations in sampling alternative conformation.

We calculated $w(|i-j|)$ from the PDB, averaging it over proteins of different length. Since each $w(l)$ scales as the protein length to the power minus two, we defined a new weight $w'(l) = cw(l)$ normalized in such a way that the mean value over all possible contacts is one, and averaged this quantity over different proteins p of length L_{p},

$$w(l) = \sum_{\mathrm{p}} \frac{(L_{\mathrm{p}}-l_{\min})(L_{\mathrm{p}}-l_{\min}-1)}{2} \frac{\sum_i C^{(\mathrm{p})}_{i,i+l}}{N_{\mathrm{C}}^{(\mathrm{p})}(L_{\mathrm{p}}-l)} \tag{35}$$

where $N_{\mathrm{C}}^{(\mathrm{p})}$ is the number of contacts of protein p and $C^{(\mathrm{p})}$ is its contact matrix. However, we got better results when, instead of the $w(l)$ calculated through this formula, we used the guess

$$w(l) \propto l^{-1}. \tag{36}$$

These weights are normalized for a fragment of length L such that the sum of the weight over all possible contacts is one,

$$\langle\langle U\rangle\rangle = \frac{\sum_{l\geq l_{\min}} w(l)\sum_i U(A_i,A_{i+l})}{\sum_{l\geq l_{\min}} w(l)(L-l-1)}. \tag{37}$$

Similarly, we define the average local interaction energy of amino acid a as

$$\langle\langle L(s,a)\rangle\rangle = \sum_s P(s)L(s,a). \tag{38}$$

The calculation of the standard deviation of the energy would require to compute the correlations between different contacts, between contacts and secondary structures, and between secondary structures. Nevertheless, neglecting these correlations we obtained better results than when we computed the correlations $\langle\langle C_{ij}C_{kl}\rangle\rangle$, $\langle\langle C_{ij}\delta(s,s_k)\rangle\rangle$, and $\langle\langle\delta(s,s_i)\delta(s',s_k)\rangle\rangle$ averaging over proteins of different length, in order to improve the sampling. We therefore estimate the variance of the contact energy E_{C} and the local energy E_{L} in a simpler way as

$$\begin{aligned}\sigma^2_{E_{\mathrm{C}}} &\approx N_{\mathrm{C}}\sigma^2_{\mathrm{U}}\\ &= N_{\mathrm{C}}\sum_l \left(w(l) - N_{\mathrm{C}}w(l)^2\right)\sum_i U(A_i,A_{i+l})^2,\end{aligned} \tag{39}$$

$$\begin{aligned}\sigma^2_{E_{\mathrm{L}}} &\approx \sum_i \sigma^2_{\mathrm{L}}(A_i)\\ &= \sum_i\left[\sum_s P(s)L(s,a)^2 - \left(\sum_s P(s)L(s,a)\right)^2\right].\end{aligned} \tag{40}$$

Using these equations, we evaluate the Z score of the energy as

$$Z_E = N^{\gamma} \frac{\sum_{i<j} C_{ij} U(a_i, A_j) - N_C \langle\langle U \rangle\rangle + \varepsilon \sum_i [L(s_i, A_i) - \langle\langle L(s, A_i) \rangle\rangle]}{N \sqrt{\frac{N_C}{N} \sigma_U^2 + \frac{\varepsilon^2}{N} \sum_i \sigma_L^2(A_i)}}. \quad (41)$$

To get to Eq. (17), we consider γ as a free parameter instead of using $\gamma = 0.5$, and multiply the Z score by a factor $\sqrt{N_C/N}$. Both these modifications improve the recognition of similar structures.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
3. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602.
4. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic clasification of protein domain structures. Structures 1997;5:1093–1108.
5. Rost B. Protein structures sustain evolutionary drift. Fold Des 1997;2:S19–S24.
6. Porto M, Roman HE, Vendruscolo M, Bastolla U. Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. Mol Biol Evol 2005;22:630–638, 1156.
7. Bastolla U, Porto M, Roman HE, Vendruscolo M. Structure, stability and evolution of proteins: principal eigenvectors of contact matrices and hydrophobicity profiles. Gene 2005;347:219–230.
8. Bastolla U, Porto M, Roman HE, Vendruscolo M. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. BMC Evol Biol 2006;6:43.
9. Bastolla U, Roman HE, Vendruscolo M. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J Theor Biol 1999;200:49–64.
10. Bastolla U, Porto M, Roman HE, Vendruscolo M. Connectivity of neutral networks, overdispersion and structural conservation in protein evolution. J Mol Evol 2003;56:243–254.
11. Bastolla U, Porto M, Roman HE, Vendruscolo M. Statistical properties of neutral evolution. J Mol Evol 2003;57:S103–S119.
12. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most protein native structures in the Protein Data Bank. Proteins 2001;44:79–96.
13. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13:211–222.
14. Muñoz 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.
15. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
16. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
17. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
18. Costantini S, Colonna G, Facchiano AM. Amino acid propensities for secondary structures are influenced by the protein structural class. Biochem Biophys Res Commun 2006;342:441–451.
19. Waterhous DV, Johnson WC, Jr. Importance of environment in determining secondary structure in proteins. Biochemistry 1994; 33:2121–2128.
20. Lawrence JR, Johnson WC. Lifson-Roig nucleation for α-helices in trifluoroethanol: context has a strong effect on the helical propensity of amino acids. Biophys Chem 2002;101/102:375–385.
21. Casari G, Sippl MJ. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. J Mol Biol 1992;224:725–732.
22. Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79:765–768.
23. Bastolla U, Porto M, Roman HE, Vendruscolo M. The principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins 2005;58:22–30.
24. Teicher F, Porto M. Vectorial representation of single- and multidomain protein folds. Eur Phys J B 2006;54:131–136.
25. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
26. Koshi JM, Goldstein RA. Models of natural mutation including site heterogeneity. Proteins 1998;32:289–295.
27. Koshi JM, Mindell DP, Goldstein RA. Using physical-chemistry based substitution models in phylogenetic analysis of HIV-1 subtypes. Mol Biol Evol 1999;16:173–179.
28. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. J Mol Biol 2001;312:289–307.
29. Dokholyan NV, Mirny LA, Shakhnovich EI. Understanding conserved amino acids in proteins. Physica A 2002;314:600–606.
30. Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N. A maximum likelihood framework for protein design. BMC Bioinformatics 2006;7:326.
31. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
32. Lobry JR. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 1997;205:309–316.
33. Creamer TP, Rose GD. Alpha-helix-forming propensities in peptides and proteins. Proteins 1994;19:85–97.
34. Richardson JM, Lopez MM, Makhatadze GI. Enthalpy of helix-coil transition: missing link in rationalizing the thermodynamics of helix-forming propensities of the amino acid residues. Proc Natl Acad Sci USA 2005;102:1413–1418.
35. Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. Proteins 2004;54:315–322.
36. Roseman MA. Hydrophobicity of polar amino-acid side chains is markedly reduced by flanking peptide bonds. J Mol Biol 1988;200:513–522.
37. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
38. Fauchere JL, Pliska V. Hydrophobic parameters of amino acid side chain from the partitioning N-acetyl amino acid amides. Eur J Med Chem 1983;18:369–375.
39. Palliser CC, Parry DA. Quantitative comparison of the ability of hydropathy scales to recognize surface β-strands in proteins. Proteins 2001;42:243–255.

40. Doig AJ, Baldwin RL. N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. Prot Sci 1995;4:1325–1336.
41. Cochran DA, Penel S, Doig AJ. Effect of the N1 residue on the stability of the alpha-helix for all 20 amino acids. Prot Sci 2001; 10:463–470.
42. Cochran DA, Doig AJ. Effect of the N2 residue on the stability of the alpha-helix for all 20 amino acids. Prot Sci 2001;10:1305–1311.
43. Monera OD, Sereda TJ, Zhou NE, Kay CM, Hodges RS. Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. J Pept Sci 1995;1:319–329.
44. Bryngelson JD, Wolynes PG. Spin-glasses and the statistical-mechanics of protein folding. Proc Natl Acad Sci. USA 1987;84: 7524–7528.
45. Go N. Theoretical studies of protein folding. Ann Rev Biophys. Bioeng 1983;12:183–210.
46. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an authomated method for protein structure comparison. Prot Sci 2002;11:2606– 2621.
47. Lupyan D, Leo-Macias A, Ortiz AR. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 2005;21: 3255–3263.
48. Goldstein R, Luthey-Schulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992; 89:4918–4922.
49. Derrida B. Random energy model: an exactly solvable model of disordered systems. Phys Rev B 1981;24:2613–2626.
50. Shakhnovich EI, Gutin AM. Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach. Biophys Chem 1989;34:187–199.
51. Govindarajan S, Goldstein RA. Optimal local propensities for model proteins. Proteins 1995;22:413–418.
52. Luque I, Mayorga OL, Freire E. Structure-based thermodynamic scale of α-helix propensities in amino acids. Biochemistry 1996;35: 13681–13688.
53. Williams RC, Chang A, Juretic D, Loughran S. Secondary structure predictions and medium range interactions. Biochim Biophys Acta 1987;916:200–204.
54. Rohl CA, Chakrabartty A, Baldwin RL. Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol. Prot Sci 1996;5:2623–2637.
55. Pace CN, Scholtz M. A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 1998;75:422–427.
56. Zhang B, Jarozszewski L, Rychlewski L, Godzik A. Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies. Fold Desi 1997;2:307–317.
57. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 1976;104:59–107.
58. Jayasinghe S, Hristova K, White SH. Energetics, stability, and prediction of transmembrane helices. J Mol Biol 2001;312:927–934.
59. Gu X, Hewett-Emmett D, Li WH. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 1998;102/103:383–391.
60. Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acid residues in globular proteins. Nature 1978;275:673–674.