PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Effective connectivity profile: A structural representation that evidences the relationship between protein structures and sequences

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

[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

The complexity of protein structures calls for simplified representations of their topology. The simplest possible mathematical description of a protein structure is a one-dimensional profile representing, for instance, buriedness or secondary structure. This kind of representation has been introduced for studying the sequence to structure relationship, with applications to fold recognition. Here we define the effective connectivity profile (EC), a network theoretical profile that self-consistently represents the network structure of the protein contact matrix. The EC profile makes mathematically explicit the relationship between protein structure and protein sequence, because it allows predicting the average hydrophobicity profile (HP) and the distributions of amino acids at each site for families of homologous proteins sharing the same structure. In this sense, the EC provides an analytic solution to the statistical inverse folding problem, which consists in finding the statistical properties of the set of sequences compatible with a given structure. We tested these predictions with simulations of the structurally constrained neutral (SCN) model of protein evolution with structure conservation, for single- and multi-domain proteins, and for a wide range of mutation processes, the latter producing sequences with very different hydrophobicity profiles, finding that the EC-based predictions are accurate even when only one sequence of the family is known. The EC profile is very significantly correlated with the HP for sequence-structure pairs in the PDB as well. The EC profile generalizes the properties of previously introduced structural profiles to modular proteins such as multidomain chains, and its correlation with the sequence profile is substantially improved with respect to the previously defined profiles, particularly for long proteins. Furthermore, the EC profile has a dynamic interpretation, since the EC components are strongly inversely related with the temperature factors measured in X-ray experiments, meaning that positions with large EC component are more strongly constrained in their equilibrium dynamics. Last, the EC profile allows to define a natural measure of modularity that correlates with the number of domains composing the protein, suggesting its application for domain decomposition. Finally, we show that structurally similar proteins have similar EC profiles, so that the similarity between aligned EC profiles can be used as a structure similarity measure, a property that we have recently applied for protein structure alignment. The code for computing the EC profile is available upon request writing to ubastolla@cbm.uam.es, and the structural profiles discussed in this article can be downloaded from the SLOTH webserver http://www.fkp.tu-darmstadt.de/SLOTH/.

Proteins 2008; 73:872–888.

Key words: structural bioinformatics; protein structure representation; hydrophobicity profile; effective connectivity; modularity.

INTRODUCTION

Protein structures can be described at many levels, ranging from individual atoms to amino acids, secondary structure, supersecondary structure, domain, and quaternary assembly. This intriguing complexity encompasses several orders of magnitude both in length and in time, and it is probably one of the reasons that make the modeling and the detailed quantitative understanding of protein folding still a difficult problem, despite the huge

Grant sponsor: Spanish Education and Science Ministry; Grant numbers: FIS2004-05073-C04-04, BIO2005-05786; Grant sponsor: Spanish Education and Science Ministry (through Ramón y Cajal program); Grant sponsor: Deutsche Forschungsgemeinschaft; Grant numbers: PO 1025/1; Grant sponsor: Acciones Integradas España Alemania program; Grant number: HA2006-0044; Grant sponsor: Deutscher Akademischer Austauschdienst program; Grant number: D/06/12848.
*Correspondence to: Ugo Bastolla, Centro de Biologia Molecular 'Severo Ochoa', (CSIC-UAM), Cantoblanco, 28049 Madrid, Spain. E-mail: ubastolla@cbm.uam.es
Received 12 September 2007; Revised 13 March 2008; Accepted 15 April 2008
Published online 5 June 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22113

attention that it has attracted both from the point of view of biophysics and from the point of view of evolutionary biology.

Consequently, big efforts have been dedicated to develop convenient simplified representations of protein sequences and structures. The simplest possible are probably vectorial representations, or profiles. It has been realized very early that protein sequences can be represented by various profiles that encode the physical and chemical characteristics of the amino acids, the most prominent one being the hydrophobicity.[1,2] Secondary structure propensity, size, and charge have been used as well. On the other hand, three-dimensional structures can also be reduced to profiles by describing the structural properties of each amino acid in the native structure,[3] first of all secondary structure and solvent accessibility.[4] It has been shown that the hydrophobicity profile of a sequence is correlated with the solvent accessibility profile of the native structure,[5] indicating that such sequence and structure profiles are interrelated.[6,7]

We define here the EC profile, a new structural profile that represents in a self-consistent way the network structure of a protein contact matrix. This structural profile is deeply related with the hydrophobicity profile (HP) derived from protein sequence, thereby representing in a very explicit way the relationship between protein sequence and protein structure. The EC allows to predict the average HP and, through it, the site-specific amino acid distributions for protein families folding into the same native structure. Such predictions exploit a maximum likelihood framework that makes them accurate even if only one sequence of the family is known. We show here that the EC-based predictions agree very well with simulations of protein evolution with stability constraints.

The EC profile is closely related with two previously defined structural profiles, the principal eigenvector of the contact matrix (PE)[8] and the revised PE.[9] The new profile improves the correlation between sequence and structure with respect to the previous defined profiles.

Besides its applications to the study of the sequence to structure relationship, the EC is also useful for structural analysis of proteins. First, its components are strongly inversely related with the temperature factors obtained in X-ray experiments, meaning that sites with large EC are more constrained in their equilibrium dynamics. Second, the EC profile allows to define in a natural way a measure of the modularity of a protein structure that correlates with the number of domains composing the structure. We are currently investigating a domain parsing method based on this idea (Teichert, Bastolla, Porto, in preparation). Finally, we show that the EC profiles of structurally similar proteins are conserved, as assessed through their normalized correlation coefficient. This suggests that the EC can be used for developing a vectorial alignment algorithm that allows to accurately align protein structures. We have developed such an algorithm, with good results.[10]

DEFINITION OF THE EFFECTIVE CONNECTIVITY PROFILE

To predict the statistical properties of protein sequences folding into a given structure, and to represent protein structures in a condensed way, we have recently introduced a family of structural profiles, the generalized effective connectivity of the contact matrix (GEC).[8,11] Interestingly, the GEC family associated to a protein structure contains as a member the hydrophobicity profile (HP) of the optimally stable sequence for that structure, computed in the framework of a model of contact interactions. This establishes a strong and explicit sequence to structure relationship. For completeness, we review here this previous work, starting from the definition of the optimal HP.

Optimal hydrophobicity profile

One of the simplest models of folding thermodynamics consists of contact interactions. In this model, protein structures are represented as a contact matrix C_{ij}, whose elements are one if two residues are in contact in the 3D structure (see Methods) and zero otherwise.

The effective contact free energy of a protein with sequence **A**, folding into a structure with contact matrix **C**, is given by $E(\mathbf{C}, \mathbf{A}) = \sum_{i<j} C_{ij}\, U(A_i,A_j)$. To get analytic insight, we consider the hydrophobic approximation in which the contact interaction matrix $U(a,b)$, expressing the effective free energy of a contact between amino acids a and b, is approximated through the main component of its spectral decomposition, $U(a,b) \approx \varepsilon_H\, h(a)h(b)$, with $\varepsilon_H < 0$. The twenty parameters $h(a)$ are proportional to the principal eigenvector of the matrix $U(a,b)$, and they are related to hydrophobicity.[12,13] They define a hydropathy scale that we call interactivity.[8] The vector $h_i \equiv h(A_i)$ is called the hydrophobicity profile (HP) of sequence **A**.

The optimally stable sequence of this simple hydrophobic model can be defined as the sequence with the lowest hydrophobic energy, which enforces high stability against unfolding. However, we have also to impose stability against alternative misfolded structures. This can be achieved by constraining the mean and mean square hydrophobicity, $\langle h \rangle$ and $\langle h^2 \rangle$ (see Methods). Angular brackets denote here average over protein sites, $\langle h \rangle = \sum_i h_i / L$.

To compute the optimal HP, we make a further approximation and neglect that the $h(A_i)$ can take only twenty discrete values. Therefore, the optimal HP is defined as the real-valued vector h_i that minimizes the quadratic form $E(\mathbf{C}, \mathbf{A}) \approx \varepsilon_H \sum_{i<j} C_{ij} h_i h_j$ for fixed $\sum_i h_i$ and $\sum_i h_i^2$.

We have adopted here two approximations: (1) Reducing the matricial contact interactions $U(a,b)$ to a quadratic form depending only on 20 parameters $h(a)$; (2) Minimizing this quadratic form as a function of continuous hydrophobicity values h_i instead of the 20 possible values of $h(A_i)$. These approximations are tested in this article by comparing the predictions based on them with the simulations of the full contact interaction model. The results presented in this article and our previous ones show that these approximations are rather accurate.

Generalized effective connectivity

We want to define a connectivity profile that depends on the global graph structure of the protein, not just on the local neighbors of each residue. This generalized effective connectivity (GEC) profile c_i can be defined self-consistently requiring that sites with large c_i components are in contact with as many as possible sites with large c_j. This condition implies that c_i should maximize the quadratic form $Q = \sum_{ij} C_{ij} c_i c_j$.

Moreover, we have to constrain the value of the mean $\langle c \rangle$ and mean square $\langle c^2 \rangle$ EC component. Fixing the mean is equivalent to choosing the normalization of the GEC components, which is arbitrary. Therefore, we can choose $\langle c \rangle = 1$, which is convenient because in this way the scale of the GEC components do not depend on the protein size.

If we interpret $p_i = c_i/L$ as the probabilistic weight of site i, the quantity $-\sum_i p_i^2$ is the exponential of the Rényi entropy of order $q = 2$ of the distribution p_i (see Methods). Therefore, constraining $\sum_i c_i^2$ is equivalent to constraining the Rényi entropy. The need to do so becomes clearer if we consider a multidomain protein. In the absence of any condition on $\sum_i c_i^2$, the solution to the above maximization problem is a vector c_i proportional to the principal eigenvector (PE) of the contact matrix C_{ij}. We now assume for simplicity that there are no interactions between different domains. In this case, the principal eigenvector vanishes outside the main domain, and therefore the GEC would vanish as well, thereby reducing the entropy of the distribution p_i, and it would not give any information about the remaining domains. In contrast, if we impose that $\sum_i p_i^2$ must be large also for multidomain proteins, we force the GEC to be a combination of several eigenvectors equally well describing all domains. The main issue of this article is how to properly set this constraint.

In formulas, the members of the GEC family of profiles c_i are described through the equations

$$\{c_i\}: \quad Q \equiv \sum_{ij} C_{ij} c_i c_j \ \text{ is maximal}, \tag{1}$$

$$\langle c \rangle \equiv \frac{1}{N} \sum_i c_i = 1 \ , \tag{2}$$

$$\langle c^2 \rangle \equiv \frac{1}{N} \sum_i c_i^2 = B \ . \tag{3}$$

Since the GEC only depends on the value of the parameter $B > 1$, it is a one parameter family of structural profiles.

The above equations are identical to those defining the optimal HP. Therefore, the normalized optimal HP $h_i^{\rm opt}/\langle h \rangle$ belongs to the GEC family, with parameter $B_{\rm HP} = \langle h^2 \rangle / \langle h \rangle^2$, where $\langle h \rangle$ and $\langle h^2 \rangle$ are the mean and mean square hydrophobicity, respectively. These quantities depend on the protein sequence, and in particular they may depend on the mutation process through which the protein evolved. The relationship between the GEC and the optimal HP determines a strong mathematical relationship between protein structure and protein sequence. We have shown in previous articles, and we will give further evidence here, that this relationship allows the quantitative prediction of the average properties of protein sequences at structurally equivalent sites.

Computation of the effective connectivity

The GEC profile can be computed by introducing two Lagrange multipliers: Λ, which imposes the condition $\langle c^2 \rangle = B$, and ϕ, which imposes the condition $\langle c \rangle = 1$. It holds

$$\sum_j C_{ij} c_j - \Lambda c_i - \phi \ = 0. \tag{4}$$

The solution of the above equations is

$$c_i(\Lambda) = \frac{\sum_j (C - \Lambda I)^{-1}_{ij}}{\sum_{kj} (C - \Lambda I)^{-1}_{kj}} \ , \tag{5}$$

where I is the identity matrix, M^{-1} represents the inverse of matrix M, and the Lagrange multiplier Λ has to be determined imposing the constraint Eq. (3).

An equivalent and more explicit representation of the GEC profile can be obtained through the spectral analysis of the contact matrix C_{ij}. This is a symmetric matrix, and therefore it possesses a complete system of L real orthonormal eigenvectors $v_i^{(\alpha)}$ associated to the real eigenvalues λ_α. Here i labels the site and α labels the eigenvector in decreasing order of λ_α, with $\alpha = 1$ denoting the principal eigenvector (PE). We denote by $\langle v^{(\alpha)} \rangle \equiv \sum_i v_i^{(\alpha)}/L$ the mean component of the eigenvector corresponding to the eigenvalue λ_α. They satisfy the normalization condition $\sum_\alpha L \langle v^{(\alpha)} \rangle^2 = 1$. We choose the direction of the eigenvector in such a way that the mean component $\langle v^{(\alpha)} \rangle$ is positive. Since all C_{ij} are positive, the PE has only positive components $v_i^{(1)} \geq 0$ and its mean value $\langle v^{(1)} \rangle$ is maximum among all eigenvectors.

With this notation, the GEC profile can be explicitly expressed as

$$c_i(\Lambda) = \sum_\alpha w_\alpha(\Lambda)\left(\frac{v_i^{(\alpha)}}{\langle v^{(\alpha)}\rangle}\right), \quad (6)$$

with the coefficients $w_\alpha(\Lambda)$ given by

$$w_\alpha(\Lambda) = \frac{\frac{L\langle v^{(\alpha)}\rangle^2}{\Lambda-\lambda_\alpha}}{\sum_\gamma \frac{L\langle v^{(\gamma)}\rangle^2}{\Lambda-\lambda_\gamma}}. \quad (7)$$

The EC of sites i that do not make any contact, $\sum_j C_{ij} = 0$, is determined only by the normalization condition, $c_i = -\phi/\Lambda$. We decided to set these EC components to zero.

It is interesting to note that, if we interpret $S[p] = -\sum_i p_i^2$ as a measure of entropy of the distribution $p_i = c_i/L$ (in fact, $\exp(-S[p])$ is the Rényi entropy, as mentioned above), and $-\sum_{ij} C_{ij}p_ip_j$ as a kind of internal energy (negative flux) of the network C_{ij}, averaged over the distribution p_i, then the EC minimizes the network free energy $F[p] = -\sum_{ij} C_{ij}p_ip_j - \Lambda S[p]$, and we have to interpret the Lagrange multiplier Λ as a network temperature.

Choosing the variance of the EC

The GEC is a one parameter family of structural profiles depending on the parameter $B > 1$ measuring the ratio between the mean square and the square of the mean components, Eq. (3). We present in this article an ansatz to choose the parameter B depending only on the protein structure. We show that this ansatz yields almost optimal structure to sequence correlation, with respect to other choices of the parameter B, between the EC and the HPs realized in protein evolution, for a wide range of amino acid compositions, yielding different values of the mean and mean square hydrophobicity.

Since exposed residues are characterized by fewer contacts and lower connectivity than buried residues, we expect that the parameter B should depend on the surface to volume ratio of the native structure. The contact matrix naturally defines a local connectivity profile, the number of contacts of each residue $\mathrm{cont}_i = \sum_j C_{ij}$, which is also dependent on the surface to volume ratio. We assume here that a suitable value of B should be the same for the optimal GEC profile as for this local connectivity profile. In formulas, we set B as

$$B_{\mathrm{cont}} = \frac{\langle \mathrm{cont}_i^2\rangle}{\langle \mathrm{cont}_i\rangle^2}. \quad (8)$$

We will call the GEC profile corresponding to this choice of B effective connectivity (EC) profile. In the present work, we investigate the properties of the EC.

Previous definitions

In previous work, we studied the structural profile corresponding to the principal eigenvector of the contact matrix, $c_i = v_i^{(1)}/\langle v^{(1)}\rangle$.[8] This is a profile of the GEC family corresponding to the special choice of parameter $B_{\mathrm{PE}} = \langle (v^{(1)})^2\rangle/\langle v^{(1)}\rangle^2$, which yields $\Lambda = \lambda_1$. The PE maximizes the normalized flux $\sum_{ij} C_{ij}c_ic_j/\sum_i c_i^2$.

For short single-domain proteins without internal modularity, all eigenvectors of the contact matrix other than the principal eigenvector have very small weights $w_\alpha = L\langle v^\alpha\rangle^2$, and the EC defined in this article is very similar to the PE. This justifies the adoption of the PE as a structural profile for small single-domain proteins in our previous work.[8,11]

For proteins with modular structure, either multidomain proteins or large single-domain proteins with internal modularity, the PE is almost zero outside the main module, and it is not sufficient to describe the entire protein structure. Therefore, in order to generalize the properties of the PE to multidomain proteins, two of us introduced the so-called revised PE (revPE),[9] which approximately coincides with the PE for small single-domain proteins. For computing the revPE, the contact matrix is modified by setting $C'_{ij} = \varepsilon(L) \ll 1$ for residues i and j that are not in contact (see Methods). In such a way, even modules that are not interacting become connected through these pseudo-contacts. The principal eigenvector of the modified contact matrix C'_{ij} is then computed and its components are subject to a nonlinear transformation to restore their original distribution.

It is easy to see that the principal eigenvector of the modified contact matrix C'_{ij} is a member of the GEC family, since it satisfies the equation $\sum_j C_{ij}c_j - \Lambda c_i - \phi = 0$, with $\phi = L\varepsilon\langle c\rangle(1\text{-}\varepsilon)$ and Λ equal to the principal eigenvalue of the modified contact matrix. This equation is equivalent to Eq. (4) defining the GEC family. If the parameter $B = \langle c^2\rangle/\langle c\rangle^2$ is the same as for the EC, this eigenvector and the EC are exactly parallel, and the EC and the revised PE are related through a nonlinear transformation (see Methods).

We verified that the revPE and the EC defined in this article are very strongly correlated, on the average $r = 0.966$ for a representative set of protein structures, as described in the Methods section. The revised PE appears to be a sigmoidal function of the EC, see Figure 1, probably due to the nonlinear transformation through which the revised PE components are obtained.

Despite this very strong correlation, we think that the EC defined in this article should be preferred, since the definition of the revised PE was more ad-hoc. We will show here that the EC profile is very strongly correlated

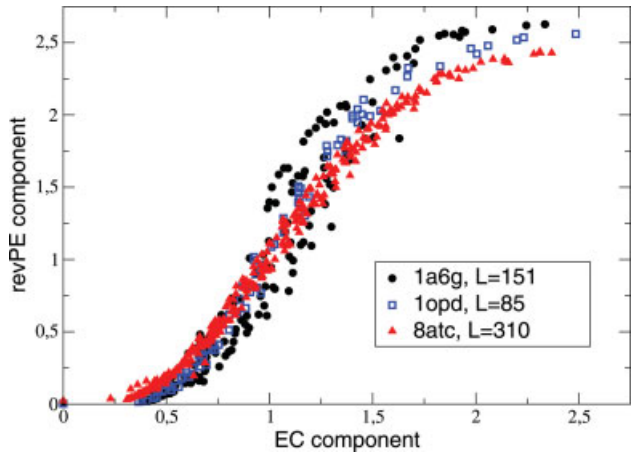

Figure 1

Revised PE components versus EC components for the proteins with PDB code 1opd ($L = 85$), 1a6g (myoglobin, $L = 151$) and 8atc, chain A ($L = 310$). The linear correlation coefficients are $r = 0.965$, $r = 0.951$, and $r = 0.984$, respectively. The sigmoidal shape and the strong correlation are typical for all proteins we studied. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

with the average HP of families of evolutionarily related proteins sharing the same fold, obtained through simulations of the SCN model, even in the case of multidomain proteins and for a wide range of mutation processes, the latter generating a wide diversity of amino acid composition. Among the profiles of the GEC family, the EC is close to optimal from this point of view, and it appears to be slightly but significantly superior to the revPE.

COMPARISON WITH SIMULATIONS

Correlation between the EC and the evolutionary averaged HP

We have seen in the previous section that the optimally stable HP belongs to the GEC family. The particular GEC profile that coincides with the optimally stable HP depends on the mean and mean square hydrophobicity, which in turn depend on the mutation process through which the protein sequence evolved.[11,14,15] This is important because the AT content of native DNA varies considerably between different organisms and environments, and it influences folding properties of the coded proteins[14,15] by favoring more hydrophobic protein sequences at the amino acid level through a bias to Adenine and Thymine at the nucleotide level.[11] However, for a wide range of possible mutation processes, hence, for a wide range of hydrophobicity, we postulate that the optimal HP will be close to the EC defined in this article, which depends only on the protein structure and which is assumed to be conserved in evolution, as we will verify below.

Because the folded state must be thermodynamically stable, we expect that the HP of a protein sequence **A** has a positive overlap with the optimal HP of its native structure, which in turn is closely related with the EC and is almost fixed during evolution. These assumptions imply that sequences of proteins sharing the same fold, although they may have only very slight sequence similarity,[16] must have a HP that is positively correlated with the EC of the structure, which constitutes their hydrophobic fingerprint.[8]

Therefore, we decompose the HP into one vector parallel to the EC and one perpendicular to it. Selection for folding stability acts to maintain the first vector large, producing a non-zero average of the parallel vectors over an evolutionary trajectory. A simple ansatz on the average of the perpendicular vectors (see Methods) predicts that the evolutionary average of the HP, indicated here with the symbol $[h(A_i)]$, is perfectly correlated with the EC, i.e.,

$$[h(A_i)] = \big\langle [h(A_i)] \big\rangle + H\big(c_i - \langle c \rangle\big) \; . \qquad (9)$$

Here H is the ratio between the standard deviation of the average HP components and the standard deviation of the EC components. In previous work, this parameter was obtained from simulated data. However, to do so, it is necessary to estimate the quantity $\langle [h(A_i)]^2 \rangle$. Such an estimate is biased if the sampling of the evolutionary trajectory is not sufficiently large. This is not the case for the other parameter $\langle [h(A_i)] \rangle$, for which we can obtain an unbiased estimate even from just one sequence. We will use in the next section a maximum likelihood approach to estimate the parameter H that avoids the problem of the bias for reduced samples. In this section, we test the prediction that the correlation coefficient between EC and average HP is one, which does not depend on the value of the parameter H.

To test this prediction, we simulated sequence evolution with stability constraints using the SCN model[11,17,18] for eight proteins, including single and multi-domain structures of different length. Notice that in these simulations we used the full contact interaction matrix $U(a,b)$, without adopting the approximations that allowed us to analytically predict the optimal HP.

To make sure that the overlap between HP and EC is practically independent of the amino acid composition, we performed the simulations for all structures with a mutation bias at the DNA level producing sequences with an AT content (content of Adenine and Thymine) varying from 5% to 95%. In this way, the AT content influences the value of the mean and mean squared hydrophobicity, and consequently the parameter $B_{\rm HP} = \langle h^2 \rangle / \langle h \rangle^2$ that enters the definition of the optimal HP

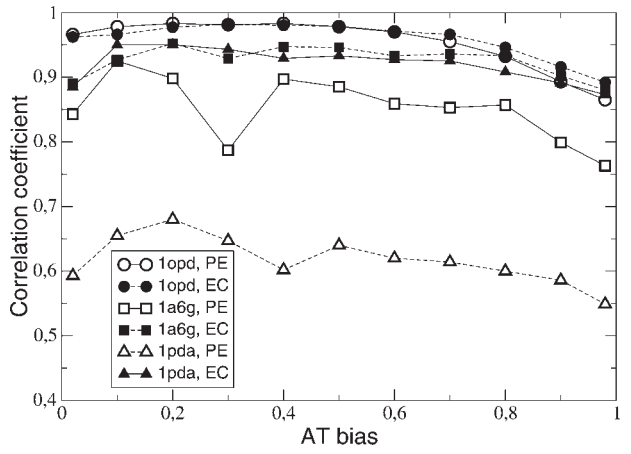

Figure 2

Correlation coefficient between the evolutionary averaged HP obtained through SCN simulations and structural profiles versus the AT mutation bias, for three representative proteins: a two-domain protein (1pda), a single-domain protein with internal modularity (1a6g, myoglobin), and a single-domain protein without internal modularity (1opd). Open symbols represent correlation with the PE profile, full symbols represent correlation with the EC profile. One can see that, except for the small single-domain protein, the correlation is significantly better for the EC than for the PE.

profile, so that different optimal HPs are generated for each value of the mutation parameters.

We measured the correlation between the average HP profile and the EC for a wide range of AT bias. This correlation is reported in Figure 2 for three representative proteins, a two-domain protein (1pda), a single-domain protein with internal modularity (1a6g, myoglobin) and a small single-domain protein without internal modularity (1opd). One can see that, in the whole range of mutation bias and for all three proteins (in fact, for all of the proteins that we simulated) the correlation coefficient is always larger than 0.87 and it depends only weakly on the AT bias, reaching the maximum values in the absence of bias (AT $\simeq$ 0.5).

We plot in the same figure the correlation coefficient between the evolutionary averaged HP and the principal eigenvector of the contact matrix (PE). For single-domain proteins such as 1opd, the PE and the EC almost coincide, and the average HP correlates almost equally well with the EC and with the PE (the correlation is typically slightly stronger with the PE). However, for single-domain proteins with internal modularity the EC and the PE are significantly different, and the HP correlates much better with the EC than with the PE. This difference increases for multidomain proteins.

We conclude from these results that the average HP correlates very well with the EC, even for extreme mutation bias that encompass a wide range of amino acid compositions. The correlation with the PE is very good for proteins without internal modularity, but it is significantly worse than with the EC for multidomain or single-domain proteins with internal modularity, for which the EC and the PE differ substantially.

Site-specific amino acid distributions

Using the results presented above, we can predict the evolutionary average hydrophobicity $[h(A_i)]$ at each site i from the EC component c_i and the hydrophobicity parameters H and $\langle[h_i]\rangle$. Furthermore, assuming maximum entropy, we can predict the site-specific amino acid distributions $\pi(a|c_i)$ only from the knowledge of the average hydrophobicity, as shown in[11,19,20]

$$\pi(a|c_i) = \frac{f(a)\exp(-\beta(c_i)h(a))}{Z(c_i)}\ , \qquad (10)$$

where the selection coefficient $\beta(c_i)$ has to be determined in such a way that the average hydrophobicity of the distribution Eq. (10) coincides with the predicted average HP, Eq. (9),

$$[h_i]_{\text{pred}} \equiv \sum_a \frac{h(a)f(a)\exp(-\beta(c_i)h(a))}{Z(c_i)} = \langle[h_i]\rangle + H\left(c_i - \langle c\rangle\right)\ . \qquad (11)$$

Here, $f(a)$ represents the expected frequency of amino acid a under mutation alone, which can be exactly computed (with the present mutation model, $f(a)$ only depends on the stationary frequencies of the four nucleotides and on the genetic code), $h(a)$ is the hydrophobicity of amino acid a analytically derived from the contact interaction energy, and $Z = \sum_a f(a)\exp(-\beta(c_i)h(a))$ is a normalization factor.

To determine $\beta(c_i)$ from Eq. (11), we still have to determine the hydrophobicity parameter H. In our previous work,[8,11] we obtained H from simulated hydrophobicity profiles as $H^2 = (\langle[h]^2\rangle - \langle[h]\rangle^2)/(\langle c^2\rangle - \langle c\rangle^2)$. However, if only a limited number of sequences is known, the estimate of $\langle[h]^2\rangle$ will be biased, since it involves the square of the evolutionary average HP at each position.

We propose here a different strategy, based on maximum likelihood. Using a multinomial model for the conditional probability of the observed amino acid distributions, $n_i(a)$, given the predicted ones, $\pi(a|c_i)$ (see Methods), we can analytically compute the derivative of the sum of log-likelihoods as

$$\frac{\partial\left(\sum_i \mathcal{L}_i\right)}{\partial H} = \langle[h]_{\text{obs}}\rangle \sum_i N_i \frac{\left([h_i]_{\text{obs}} - [h_i]_{\text{pred}}\right)\left(c_i - \langle c\rangle\right)}{\left([h_i^2]_{\text{pred}} - [h_i]^2_{\text{pred}}\right)}\ , \qquad (12)$$

where $N_i = \sum_a n_i(a)$ is the number of aligned residues at position i, $[h_i]_{\text{obs}}$ is the observed site-specific mean

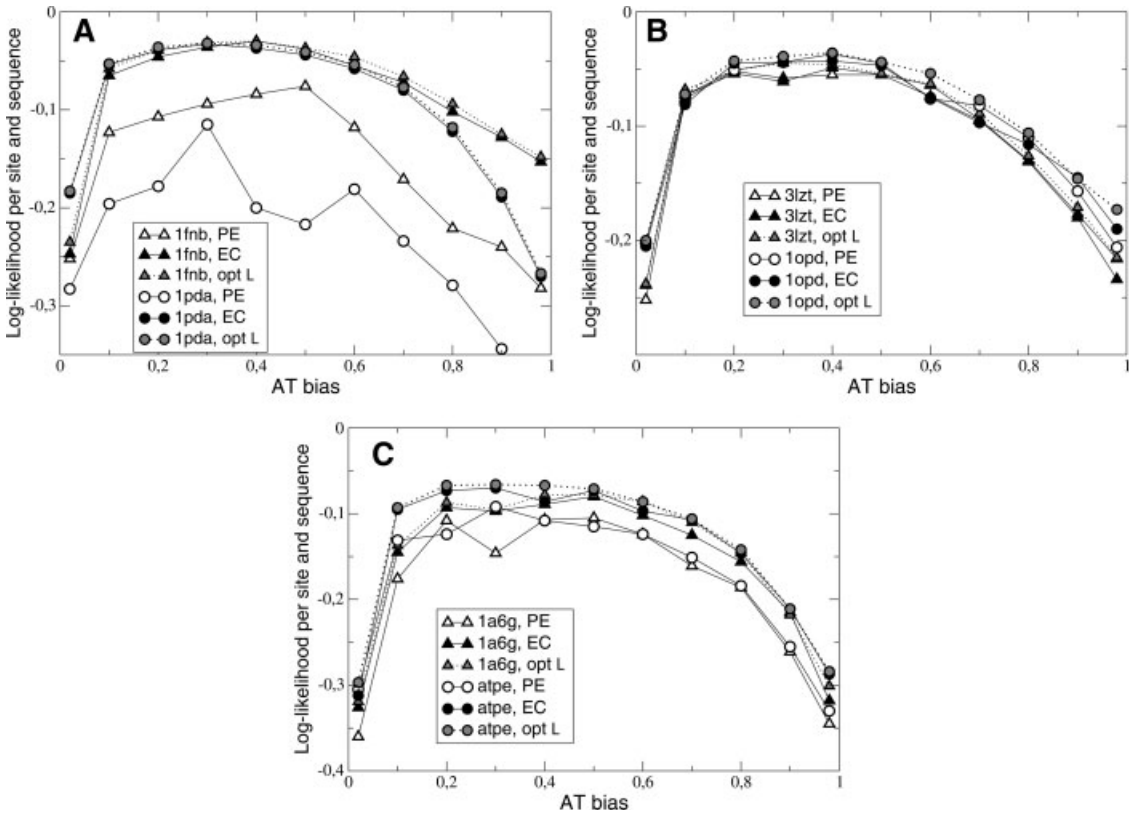

Figure 3

Log-likelihood per site and sequence of the observed given the predicted site-specific amino acid distributions. The predicted distributions are calculated using the optimal hydrophobicity parameter H in Eq. (9), with the structural profile c_i equal to the PE profile (open symbols), the EC profile (full black symbols) or the GEC Eq. (6) with Λ chosen to maximize the likelihood (full gray symbols with dotted line). (**A**) Multidomain proteins. (**B**) Single-domain proteins with internal modularity. (**C**) Small single-domain proteins lacking internal modularity. Except for the last case, the optimal Λ yields likelihood values almost indistinguishable from the EC.

hydrophobicity (see Methods), which we can estimate without bias even for just one protein sequence, $[h_i]_{\rm pred}$ is a linear function of H given by Eq. (9), and $[h_i^2]_{\rm pred} = \sum_a \pi(a|c_i)\ [h(a)]^2$ depends on H through the coefficients $\beta(c_i)$. We can then obtain the hydrophobicity parameter H by maximizing the sum of the log-likelihood numerically through gradient methods.

We show in Figure 3 the maximum likelihood per site and per sequence calculated by estimating $[h_i]_{\rm pred}$ through three structural vectors belonging to the GEC family, described by Eq. (6): (a) The PE (empty symbols); (b) The EC (black symbols); (c) The GEC profile corresponding to the optimal Λ yielding maximum likelihood at each value of the mutation bias (gray symbols). We distinguish three kinds of protein structures,

1. Multidomain proteins (1fnb and 1pda).
2. Small single-domain proteins without internal modularity (3lzt and 1opd).
3. Single-domain proteins with internal modularity (1a6g and 1aqt).

For all of these proteins, and all of the proteins that we simulated, the predictions based on the EC get very close to those obtained by optimizing Λ independently for each value of the AT bias. These results support the choice of the parameter B through Eq. (8), since in this way the site-specific amino acid distributions predicted through the structural profile are very close to optimal within the GEC family, and the correlation coefficient between observed and predicted average hydrophobicity is very close to one. Notice also that extreme mutation bias (AT smaller than 0.1 or larger than 0.8) are much worse fit both through the EC and through the GEC profile with optimum Λ, as seen from their low likelihood values, showing that for the corresponding extreme hydrophobicity values our analytic model becomes inaccurate. However, such extreme nucleotide bias are usually not encountered in nature.

In contrast, the predictions based on the PE profile are almost as good as those based on the EC for small single domain proteins without internal modularity, for which the EC and the PE almost coincide, see Figure 3(B), but

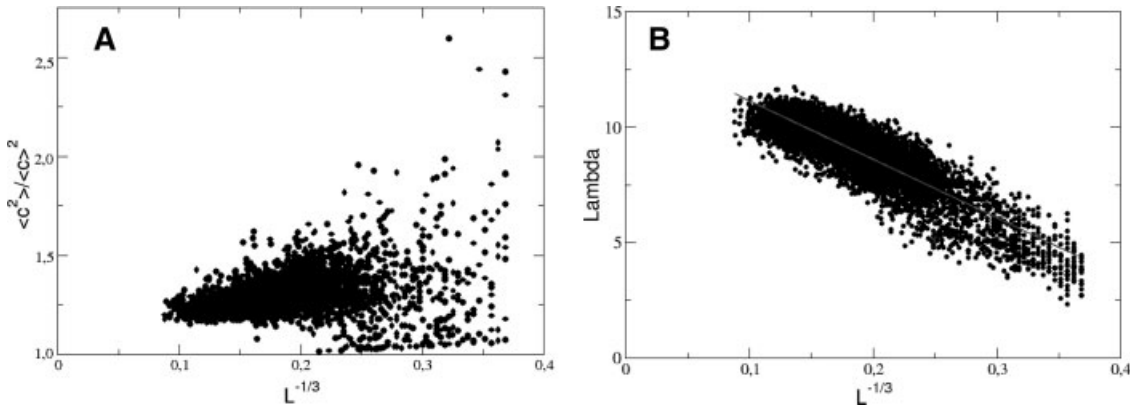

Figure 4

The value of $\langle cont^2\rangle/\langle cont\rangle^2$ used to determine the parameter $\mathbf{\Lambda}$ through constraint $\langle cont^2\rangle/\langle cont\rangle^2 = \langle EC^2\rangle/\langle EC\rangle^2$ (**A**) and the corresponding parameter $\mathbf{\Lambda}$ (**B**) versus chain length L for 7465 nonredundant protein structures. It can be seen that both quantities scale with L as the surface to volume ratio, $L^{-1/3}$.

they become much worse, as witnessed by their much lower likelihood, for multidomain proteins, Figure 3(A), and also for single-domain proteins with internal modularity, Figure 3(C). This comparison shows that using the EC instead of the PE yields a considerable improvement in the quality of the predictions.

ANALYSIS OF PDB STRUCTURES

Statistical analyses of EC properties

We computed the EC profile for a nonredundant set of protein structures having less than 50% sequence identity with each other.[21] From this set, we eliminated nonglobular structures and structures determined by NMR spectroscopy (see Methods).

First, we examine the quantity B_{cont}, Eq. (8), which represents the ratio between mean of the square and squared mean of the local connectivity profile. As expected, this variable scales as the surface to volume ratio $L^{-1/3}$ (correlation coefficient $r = 0.41$), with large corrections to scaling for small proteins with large surface to volume ratio (see Fig. 4(A)).

We expect that the principal eigenvalue of the contact matrix, λ_1, is related to the number of contacts per residues. In fact, it can be shown that $\lambda_1 > N_{cont}/L$ and that they are equal if each residue has exactly the same number of contacts. Similarly, the Lagrange multiplier Λ, determined by imposing the condition Eq. (8), is expected to be related to the number of contacts per residue N_{cont}/L, which scales as the surface to volume ratio, $L^{-1/3}$. Figure 4(B) confirms this expectation.

Next, we examine the correlation between the structural EC profile introduced in this article and the HP profile representing the protein sequence. For comparison, we also measured the correlation between the HP and the PE profile. They are shown in Figure 5. One can see that the two correlations are very similar for short proteins, for which the EC and the PE almost coincide, but for large proteins the correlations are significantly stronger for the EC profile. The mean correlation coefficient is $\overline{r}$(HP, PE) = 0.352 for the PE, and $\overline{r}$(HP, EC) = 0.434 for the EC.

The correlation coefficient between the EC and the HP decreases very significantly with chain length ($r = -0.23$, student $t = -19$, corresponding to vanishingly small probability). This behavior is expected on the ground

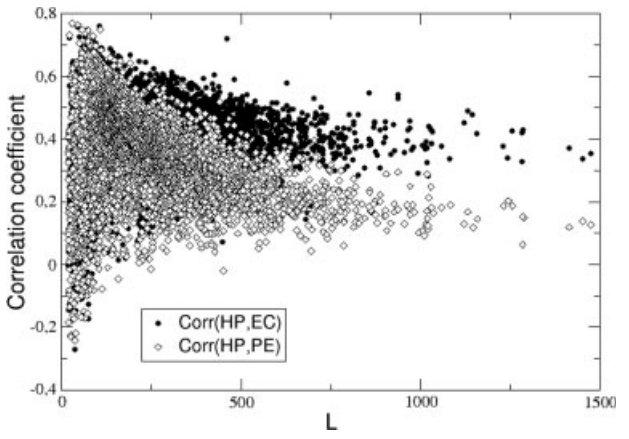

Figure 5

Correlation coefficient between the hydrophobicity profile HP and the EC and PE profiles for 7465 protein structures versus chain length. The correlation coefficients are almost equal for small proteins, for which PE and EC almost coincide. However, the correlation between EC and HP does not depend on chain length, whereas the correlation between PE and HP decreases with chain length for long proteins, where PE and EC are different.

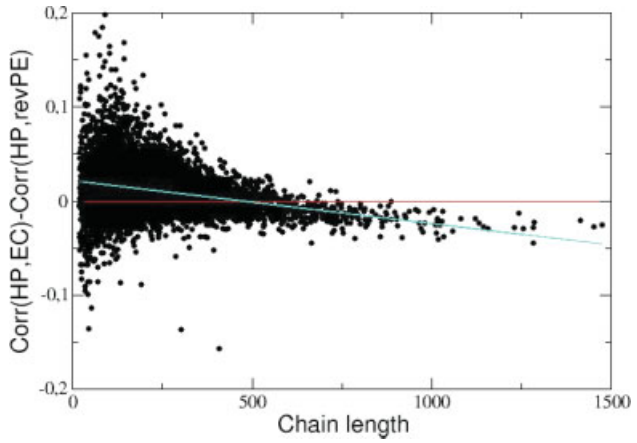

Figure 6

Difference between the correlation coefficient of the HP and the EC and that of the HP and the revised PE versus chain length. The difference is on the average positive for short proteins and always negative for long proteins, meaning that the HP is on the average more correlated with the EC for short proteins and it is more correlated with the revised PE for long proteins. Globally, the HP is better correlated with the EC. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

that long proteins have more contacts per residue and they are easier to stabilize, so that native interactions are less optimized in long proteins.[22]

We also measured the correlation coefficients between the HP profiles and the revised PE previously defined. The mean correlation is $r = 0.424$ for the revised PE and $r = 0.434$ for the EC, a small but significant difference. Notice that both profiles perform much better than the PE. When looking in more detail at the correlation coefficient versus chain length, we can see that the EC performs significantly better than the revised PE for short proteins, whereas the opposite holds for long proteins, see Figure 6.

Last, we examined the correlation between the temperature factors measured in X-ray experiments and the two structural profiles EC and PE. We expect to find a negative correlation, since sites with large EC are structurally more constrained and they are expected to undergo smaller fluctuations in the native state. A more formal argument can be based on the relationship between the EC and the normal modes of the protein. The vibrational dynamics in the folded state can be described through the normal modes of the Hessian matrix of the protein in the native state. In the framework of elastic network models,[23] the Hessian matrix can be estimated through the Laplacian matrix associated to the contact matrix of the native structure, $L_{ij} = \delta_{ij} \sum_k C_{ik} - C_{ij}$. The main eigenvectors of this Laplacian matrix correspond to the lowest frequency normal modes. This formal analogy also leads to expect that the lowest frequency normal modes are inversely related with the EC.

We find significant correlations between the experimental temperature profiles and the structural profiles. The correlation is $\overline{r}(b_{factor}, PE) = 0.429$ with the PE profile and $\overline{r}(b_{factor}, EC) = 0.479$ with the EC profile, see Figure 7. The correlations are slightly but significantly better for the EC profile than for the PE profile.

Modularity

The spectral properties of the contact matrix reflect its modular organization. This modularity can be made rather explicit using the expression of the EC as a sum of eigenvectors of the contact matrix, with coefficients given in Eq. (7).

In particular, for a compact and homogeneous structure in which almost all buried sites have the same number of contacts, the PE can be shown to be almost constant at buried sites. The coefficient of the PE in the expression of the EC, Eq. (7), is proportional to $L\langle v^{(1)}\rangle^2$, which is almost one for a homogeneous contact matrix (minus a correction proportional to the surface to volume ratio). Therefore, the contribution of all other eigenvectors to the EC vanish, and for such a homogeneous compact structure the PE and the EC are expected to be very similar. This is indeed observed for short, single-domain proteins.

In contrast, if a protein structure consists of distinct modules such that there are many intramodular contacts but the modules are only sparesly connected, which is particularly true for multidomain proteins, the PE is almost zero outside the main domain, and the contribution of other eigenvectors to the EC is not negligible, so that the EC and the PE differ substantially. It is therefore natural

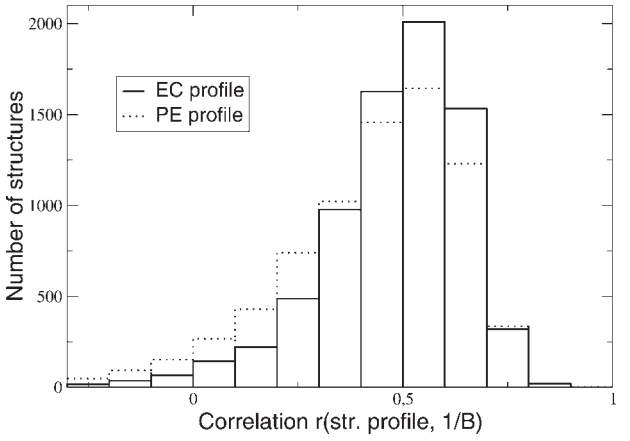

Figure 7

Histogram of the correlation coefficients between the inverse of the temperature profile measured in X-ray experiments and the EC and PE profiles for 7465 protein structures. The correlations are significant for most proteins, and the average quality is similar for the two structural profiles, with the EC profile yielding on the average a slightly better correlation.

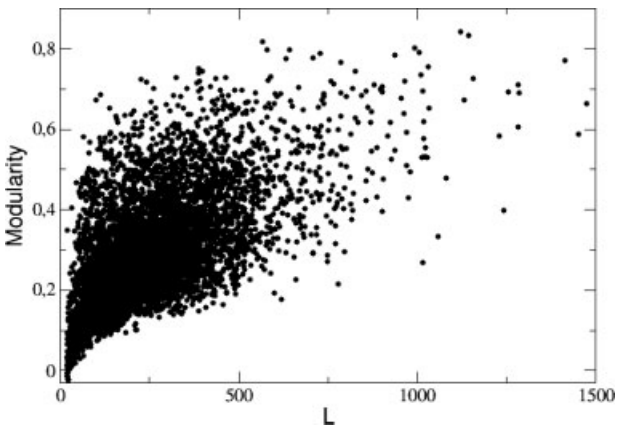

Figure 8

Modularity, defined as the contribution to the EC of eigenvectors other than the PE, Eq. (13), versus chain length.

to define the modularity of a contact matrix as the contribution to the EC of eigenvectors other than the PE,

$$\text{modularity} = 1 - \frac{\sum_i v_i^{(1)} \text{EC}_i}{\sqrt{\sum_i (\text{EC}_i)^2}} = 1 - w_1(\Lambda) \,, \qquad (13)$$

where we have used Eq. (6). If modularity is small, the PE and the EC almost coincide. Modularity is represented in Figure 8 versus chain length. There is a strong correlation between these two variables, Corr(L, modularity) = 0.63.

We expect that modularity as defined above reflects the internal structure of the protein, in particular its domain composition. To test this relationship, we have decomposed proteins into domains using the algorithm Protein Domain Parser (PDP).[24] As expected, there is positive correlation between the number of domains and modularity, $r(N_{\text{domains}}, \text{modularity}) = 0.623$, see Figure 9. Nevertheless, chain length correlates with the number of domains stronger than modularity, $r(N_{\text{domains}}, L) = 0.686$, see Figure 9(B), and it also correlates very strongly with modularity. To test how much of the correlation between modularity and number of domains is due to the effect of chain length, we measured the partial correlation between modularity and the number of domains at fixed length, finding $r(N_{\text{domains}}, \text{modularity} \cdot L) = 0.341$, which is highly significant for our large data set. This suggests that our definition of modularity captures the modular structure of multidomain proteins.

From Eq. (7) one sees that the weights $w_\alpha(\Lambda)$ become more homogeneous as Λ increases, whereas when Λ is close to λ_1 most of the weight is concentrated at the principal eigenvector $\alpha = 1$. Therefore, for a given structure, modularity $1 - w_1(\Lambda)$ increases with the Lagrange multiplier Λ. For our representative set of structures, we observe that modularity and Λ are positively correlated. Nevertheless, this correlation is essentially due to chain length, which is correlated both with modularity (see Fig. 8) and with Λ (see Fig. 4(B)). The partial correlation between modularity and Λ at constant length is not significant, $r = 0.026$.

This lack of correlation between modularity and Λ at constant length supports our choice of the parameter Λ through Eq. (8). In fact, if Λ is chosen randomly, we would expect to find a trivial correlation between Λ and modularity, due to the fact that modularity increases with Λ for a given contact matrix. However, the observed correlation is entirely due to the effect of chain length, meaning that this trivial effect is removed through our choice of Λ.

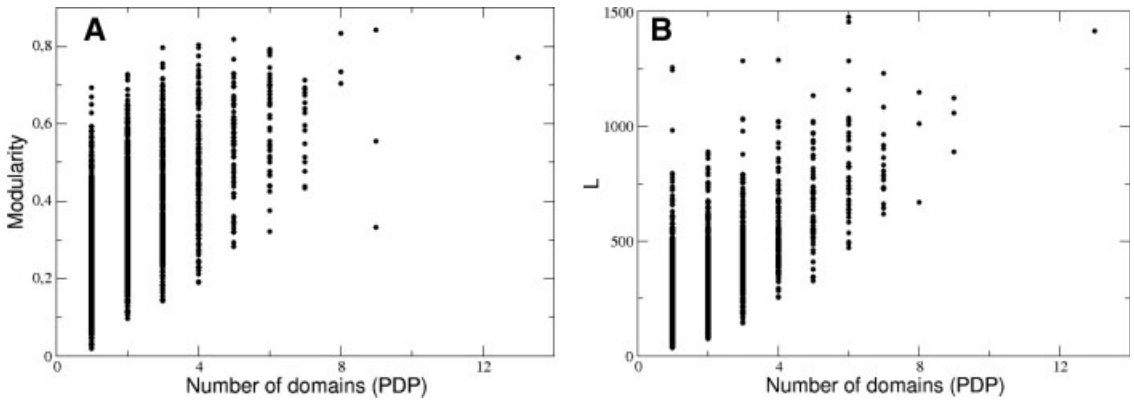

Figure 9

Modularity (**A**) and chain length (**B**) versus the number of domains according to the algorithm PDP.

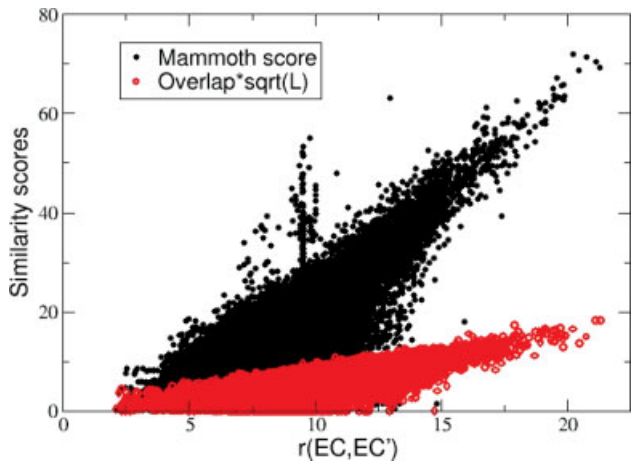

Figure 10

Similarity scores between 56,450 pairs of evolutionary related proteins in the same SCOP superfamily, with less than 40% sequence identity. Mammoth similarity score and overlap similarity score are represented versus the EC similarity score.

Evolutionary conservation of the EC profile

Because the EC of a protein structure defines the hydrophobic fingerprint of the corresponding sequence family, the similarity between two EC profiles can be used to assess the similarity between the two protein structures. To test this idea, we defined an EC similarity score and compared it with two other similarity scores, the contact overlap and the Mammoth score.

These similarity scores are meaningful as a measure of structural similarity only after optimal superimposition of the two structures, i.e., a structural alignment. The properties of the EC directly suggest to align two EC profiles maximizing their EC similarity scores as a way to align two protein structures. If the similarity score is local, the optimal alignment can be calculated very efficiently, for instance through dynamic programming. We have recently developed such an algorithm.[10] In the present article, however, we have used the structural alignment algorithm Mammoth[25] and we have calculated similarity scores based on the alignment given by Mammoth.

The Mammoth similarity score, based on the percentage of structurally superimposed residues, the contact overlap similarity score, and the EC similarity score are defined in the Methods section. They are normalized by chain length in such a way that the similarity score of unrelated proteins depends as little as possible on their length.

We plot in Figure 10 the Mammoth and the overlap similarity scores versus the EC similarity score for 56450 pairs of evolutionarily related proteins belonging to the same superfamily according to the SCOP classification of proteins[26] and having less than 40% pairwise sequence similarity. For most but not all of the pairs of related proteins the similarity scores are significant with respect to pairs of unrelated proteins, which means that the features on which they are based (spatially superimposed residues, residues in contacts, and EC) are conserved in evolution. The three similarity scores are strongly correlated with each other, with correlation coefficients $r = 0.85$ (Mammoth score and contact score), $r = 0.74$ (Mammoth score and EC score) and $r = 0.69$ (contact score and EC score). It is somewhat surprising that, despite being derived from the contact matrix, the EC score is more strongly correlated with the Mammoth score than with the contact score.

DISCUSSION AND CONCLUSIONS

The relationship between protein structures and sequences is key to understanding protein evolution, and for the important goals of predicting protein structure and function from sequence. Here, we define a structural profile based on network theory that allows to analytically address the statistical inverse folding problem: predicting the statistical properties of protein sequences sharing a given native structure.

In a previous article, we and others had defined a one-parameter family of structural profiles for proteins, the generalized effective connectivity of the contact matrix (GEC).[11] The GEC establishes a strong mathematical relationship between protein sequence and protein structure, based on the evolutionary conservation of folding stability. In fact, in the framework of models of folding stability based on contact interactions, the optimally stable hydrophobicity profile (optimal HP) and the evolutionary average HP associated to a given protein structure belong to the GEC family of this structure. However, the average HP depends on the protein sequence through the mean and mean square hydrophobicity, which in turn may depend on the mutation process through which the protein evolved.[11,14,15] In this article, we propose a simple and structurally motivated definition of a particular profile of the GEC family that (a) depends only on the protein structure, and (b) correlates very strongly with the evolutionary HP for a wide range of mean hydrophobicities, corresponding to different amino acid compositions and mutation frequencies. We call this structural profile the effective connectivity profile (EC).

Compared with the entire GEC family, the EC has almost optimal correlation with the evolutionary average HP generated through simulated protein evolution with stability constraints for a wide range of mutation processes. We have studied previously two other structural profiles for proteins. We show here that they are also members of the GEC family. The first one is the principal eigenvector of the contact matrix (PE), which is very

similar to the EC for small single-domain proteins without internal modularity. However, if the contact matrix has a modular structure, such as for multidomain proteins, the PE is not a good description of the entire protein. Accordingly, we find in such cases that the structure to sequence relationship is much better expressed through the EC than through the PE. The revised PE was later introduced in order to generalize the properties of the PE to multidomain proteins.[9] This profile depends on two parameters that were empirically determined so that PE and revised PE approximately coincide for small proteins. The EC defined here is very strongly correlated, through a sigmoidal relationship, to the revised PE, but it has a more intrinsic and less ad-hoc definition, and it correlates more strongly with observed HPs.

The structural profile defined here has several applications in bioinformatics, involving the sequence to structure relationship and the analysis of protein structures.

The first application consists in predicting sequence profiles based on structure, with potential applications for the task of fold recognition. In fact, the EC allows to predict the evolutionary average HP and, through it, the site-specific amino acid distributions for a protein family of known structure. We have proposed here a new method to this purpose, which is accurate even when only one protein sequence is known. The method was tested against sequence families generated through simulations of the SCN model of protein evolution with stability constraints, yielding very good results. Therefore, the EC is very promising for predicting sequence profiles from protein structures.

Preliminary results show that the likelihood of observed given predicted distributions is high also for real protein families, suggesting a possible application to fold recognition. In such case, the background amino acid distributions in the absence of selection, $f(a)$, should be obtained through the statistics of the PDB instead of being computed from the simulated mutation process as we do in our simulations.

We have also found a strong correlation between the HP and the EC in a large representative set of proteins in the PDB. This correlation between the structural profiles and the sequence profile decreases for long proteins, as it is expected from the finding that native interactions are less optimized in long proteins.[22]

Furthermore, the EC also allows to represent a mean-field model of protein sequence evolution.[11] In such a model, the protein is subject to global constraints on folding stability, but each protein site evolves independently, thereby making it possible to use such a model in the framework of maximum likelihood studies of molecular evolution. Such studies very rarely take into account structural constraints, due to the computational difficulty to deal with correlated evolution.

The second field of applications in bioinformatics concerns the equilibrium dynamics of proteins. Residues with large EC component are expected to be more constrained in the native state, and therefore they are expected to have smaller dynamical amplitudes, which are quantitatively expressed through the b factors measured in X-ray experiments. We have found indeed a strongly significant inverse correlation between the b factors and the EC or the PE. This relationship is stronger for the EC than for the PE, although the difference is only marginal. In this way, the EC allows to easily estimate the magnitude of the displacements in the native state, in a way that is closely related to the elastic network model.[23]

Third, the EC can be applied to studies of structure evolution. We have shown that the EC is evolutionary conserved and that it can be used to compare related protein structures, suggesting its use for protein structure alignment. We exploited this interesting possibility in the new structure alignment program SABERTOOTH.[10]

Fourth and last, the EC has interesting applications for analyzing the modular structure of a protein. For homogeneous structures with a similar number of contacts at all buried positions, we expect that the PE is almost constant, so that $L\langle v^{(1)}\rangle^2 \approx 1$ minus a term proportional to the surface to volume ratio $L^{-1/3}$, and the contribution of the PE to the EC profile, $w_1(\Lambda) \propto L\langle v^{(1)}\rangle^2/(\Lambda-\Lambda_1)$, is close to one even for large α. In contrast, for modular structures the PE is almost zero outside the main domain, so that $L\langle v^{(1)}\rangle^2$ is small. Therefore, the contribution to the EC profile of eigenvectors other than the PE, $1 - w_1(\Lambda)$, is a useful measure to identify proteins that contain internal modules such as domains. We have called this quantity *modularity*. This quantity was tested using the algorithm PDP to define the domains. We found a strong correlation, $r = 0.62$, between modularity and number of domains defined through PDP. Only part of this correlation can be attributed to the effect of chain length, which correlates both with modularity and the number of domains. The partial correlation between modularity and the number of domains at fixed chain length is in fact strongly significant, $r = 0.34$, which means that our modularity measure indeed captures some properties of the internal modularity of proteins. We are currently developing a domain parsing algorithm based on this information, plus the requirement that residues belonging to the same domain should be as much as possible continuous along the sequence.

Summarizing, we have introduced here a network theoretical profile that allows to represent a protein structure as a vector. We would like to emphasize that this profile is not restricted to contact matrices of protein structures, but it can be associated to any graph. In the context of proteins, the EC profile expresses in a very simple mathematical form the structure to sequence relationship, it allows to predict the distribution of amino acids at each site in the family of proteins sharing the same native structure, it allows to represent a mean-field

model of protein evolution, it is related to the dynamic amplitude in the native state ensemble, and it can represent in a very natural way the modular structure of the protein, suggesting several interesting bioinformatics applications.

METHODS

Contact model of folding stability

Here we adopt a simple model of folding free energy based on contact interactions,

$$E(\mathbf{C},\mathbf{A}) = \sum_{i<j} C_{ij} U(A_i, A_j) \ . \qquad (14)$$

Here C_{ij} is the contact matrix, whose elements are one if two heavy atoms belonging to residues i and j are closer than 4.5 Å in the native structure and zero otherwise. We also set $C_{ij} = 0$ for pairs $|i - j| < 4$ to avoid trivial short range contacts. The contact interaction matrix $U(a,b)$ is a twenty times twenty symmetric matrix. We use the parameters determined in Ref. 27.

The folding free energy allows to estimate the stability with respect to the unfolded state. Stability of the native state with respect to compact, wrongly folded configurations can be estimated through the normalized energy gap α, defined as the minimum ratio between the free energy difference between the native state and any other state and their structural divergence,

$$\alpha(\mathbf{C}_{\rm nat},\mathbf{A}) = \min_{\mathbf{C}} \frac{E(\mathbf{C},\mathbf{A}) - E(\mathbf{C}_{\rm nat},\mathbf{A})}{|E(\mathbf{C}_{\rm nat},\mathbf{A})|(1 - q(\mathbf{C},\mathbf{C}_{\rm nat}))} \ , \qquad (15)$$

where the contact overlap $q(\mathbf{C}, \mathbf{C}')$ is a normalized similarity measure between two aligned contact matrices, expressing their fraction of shared contacts

$$q(\mathbf{C},\mathbf{C}') = \frac{\sum_{ij} C_{ij} C'_{ij}}{\sqrt{\sum_{ij} C_{ij} \sum_{ij} C'_{ij}}} \ . \qquad (16)$$

A large value of α guarantees that the native structure has lower energy than all other alternative structures, and that the low energy structures are related with the native, in the sense that they have a large overlap $q(\mathbf{C}, \mathbf{C}_{\rm nat})$.

There are two ways to compute the normalized energy gap. The first one consists in sampling protein-like alternative structures, for instance by threading sequence $\mathbf{A}$ without gaps (gapless threading) against a nonredundant structure database from the Protein Data Bank. This method is easy to implement but it has the drawback to overestimate the energy gap for large proteins, for which the sampling is very poor. The second method is based on estimating the normalized energy gap through the Random Energy Model,[28,29] which yields

$$\alpha_{\rm REM}(\mathbf{C}_{\rm nat},\mathbf{A}) = \frac{N_{\rm C}\langle U\rangle_{\mathbf{A}} - \sigma_{U,\mathbf{A}}\sqrt{2N_{\rm C}\log(m_L)} - E(\mathbf{C}_{\rm nat},\mathbf{A})}{|E(\mathbf{C}_{\rm nat},\mathbf{A})|(1-q_0)} \ , \qquad (17)$$

where $N_{\rm C}$ is the number of native contacts, $\langle {\rm U}\rangle_{\mathbf{A}}$ and $\sigma_{U,\mathbf{A}}$ are the mean and standard deviation of the interaction energy for all possible contacts, native and nonnative, within sequence $\mathbf{A}$, and m_L is the number of independent contact matrices for a protein of length L, satisfying physical constraints of hard core repulsion, hydrogen bonding and compactness, which we estimated as $\log(m_L) \approx 0.1 \times L + 4$.[22]

This equation implies that constraining the normalized energy gap at fixed native free energy is equivalent to constraining the mean and mean square contact interactions.

A further simplification of the above model is given by the hydrophobic approximation $U(a,b) \approx \varepsilon_H h(a) h(b)$, where $h(a)$ is the principal eigenvector of the contact interaction matrix. Within this approximation, the contact free energy can be approximated as

$$E(\mathbf{C},\mathbf{A}) \approx \varepsilon_H \sum_{i<j} C_{ij} h(A_i) h(A_j), \qquad (18)$$

We used this approximation only for performing analytical calculations, whereas all simulations were performed with the full contact interaction model. In the framework of the hydrophobic approximation, the mean and mean square contact interactions are given by $\langle U\rangle_{\mathbf{A}} = \sum_{ij} U(A_i,A_j)/L^2 \approx \varepsilon_H \ \langle h(A_i)\rangle^2$ and $\langle U^2\rangle_{\mathbf{A}} = \sum_{ij} U(A_i,A_j)^2/L^2 \approx \varepsilon_H \ \langle h(A_i)^2\rangle^2$. Therefore, within the hydrophobic approximation, constraining the normalized energy gap for fixed native free energy can be achieved by constraining the mean and mean square hydrophobicities $\langle h\rangle$ and $\langle h^2\rangle$.

SCN model

The structurally constrained neutral (SCN) model of protein evolution[17,18] simulates the evolution of a protein coding gene in the limit of small mutation frequency, $M\mu \ll 1$, where μ is the mutation rate per gene and generation and M is the population size. In this limit, at most one mutant allele arises at each generation, and we can simulate just one protein at each step of the model. In the spirit of Kimura's neutral evolution theory,[30,31] we consider a binary fitness function, $F = 1$ (viable) if the folding properties of the mutant protein are above predefined thresholds and $F = 0$ (lethal) if they are below threshold. Viable mutations are fixed in the population (substitution), whereas lethal mutations are eliminated.

Synonymous mutations are always considered viable and mutations to stop codons are considered lethal. To

compute the fitness of missense mutations, we estimate the folding free energy and the normalized energy gap for the mutated amino acid sequence (see above). The threshold values are chosen equal to 98% of the value of the energy and the normalized energy gap for the sequence of the target structure in the PDB, $\mathbf{A}_0$, which is the starting point of the simulation: $E_{\rm thr} = 0.98\ E(\mathbf{C}_{\rm nat}, \mathbf{A}_0) < 0$, $\alpha_{\rm thr} = 0.98\ \alpha(\mathbf{C}_{\rm nat}, \mathbf{A}_0) > 0$. A viable sequence must have $E(\mathbf{C}_{\rm nat}, \mathbf{A}) < E_{\rm thr}$ and $\alpha(\mathbf{C}_{\rm nat}, \mathbf{A}) > \alpha_{\rm thr}$, so that they may be only marginally less stable than the PDB sequence.

Mutation is simulated at the nucleotide level. At each step of the simulated evolutionary process, one nucleotide is mutated at random with probabilities given by the HKY mutation matrix $P_\mu^{\rm nuc}(n,n')$.[32,33] This matrix obeys detailed balance with respect to the stationary frequencies $\phi(n)$, where n is one of the four nucleotides, $\sum_n \phi(n) P_\mu^{\rm nuc}(n,n') = \phi(n')$. The four stationary frequencies $\phi(n)$ are the only parameters of the mutation matrix that influence the stationary amino acid frequencies in the absence of selection, $f(a)$, i.e., the amino acid frequencies under mutation alone. They can be expressed as the sum of the stationary frequencies of all the codons coding for amino acid a,

$$f(a) = \sum_{n_1 n_2 n_3} \delta(a, \mathcal{A}[n_1 n_2 n_3]) \phi(n_1) \phi(n_2) \phi(n_3)\ , \qquad (19)$$

where $n_1 n_2 n_3$ are the three nucleotides forming the codon, $\mathcal{A}[n_1 n_2 n_3]$ is the amino acid coded by codon $n_1 n_2 n_3$, the Kronecker's delta selects codons coding for amino acid a, and $\phi(n)$ is the stationary frequency of nucleotide n in the HKY model.

Furthermore, we only simulated mutation matrices satisfying the condition $\phi(A) = \phi(T)$ and $\phi(C) = \phi(G)$, which is called Chargaff 2nd parity rule.[34] Therefore, the mutation matrices that we used can be represented using just one parameter, the combined A plus T content in the absence of selection, $\phi(A) + \phi(T)$.

For consistency with our previous work, we simulated the mutation process using HKY mutation matrices that take into account that transitions (A to G or T to C and reverse) are more likely than transversions (all other kinds of nucleotide mutations), adopting rates $P_\mu^{\rm nuc}(n,n') = \mu k \phi(n')$ if the mutation from n to n' is a transition and $P_\mu^{\rm nuc}(n,n') = \mu \phi(n')$ if it is a transversion, with $k = 2$ representing the transition to transversion ratio. We verified in our previous work that this parameter does not modify significantly the stationary amino acid distributions observed in simulated protein evolution, which agree very well with our analytic formula, Eq. (21).

Evolutionary average HP

We analytically compute the evolutionary average HP in two steps:

1. We compute the HP corresponding to the optimally stable sequence. For doing this, we adopt the hydrophobic approximation, Eq. (18), and the continuous approximation, neglecting that $h(A_i)$ can assume only twenty values, and we maximize the quadratic form $\sum_{ij} C_{ij} h_i h_j$ subject to two constraints on mean and mean squared hydrophobicity, $\sum_i h_i/L = \langle h \rangle$ and $\sum_i h_i^2/L = \langle h^2 \rangle$. Within these approximations, the normalized optimal HP, $h_i^{\rm opt}/\langle h \rangle$, belongs to the GEC family defined in this article, with parameter $B_{\rm HP} = \langle h^2 \rangle / \langle h \rangle^2$. We further assume that, among all profiles of the GEC family, the optimal HP is almost parallel to the EC for most values of $B_{\rm HP}$ realized in protein evolution, $h_i^{\rm opt}/\langle h \rangle \propto c_i$.
2. We decompose the HP of a generic protein sequence into two vectors, one parallel and one perpendicular to the optimal HP c_i: $h(A_i) = \gamma(\mathbf{A}) c_i + b_i$, with $\sum_i b_i c_i = 0$. The evolutionary average of the HP, denoted through square brackets, is $[h(A_i)] = [\gamma(\mathbf{A})] c_i + [b_i]$. We may assume that the evolutionary average of the perpendicular vector, $[b_i]$, does not depend on the site. However, to impose the condition that $[b_i]$ is perpendicular to c_i, we have to subtract from this constant vector a vector proportional to the c_i, leading to the ansatz $[b_i] \propto (1 - \langle c \rangle c_i / \langle c^2 \rangle)$. Summing the vector produced through this ansatz with the parallel vector $[\gamma(\mathbf{A})]\ c_i$, we see that the evolutionary average of the HP has correlation coefficient equal to one with the EC, or

$$[h(A_i)] = \big\langle [h(A_i)] \big\rangle + H\big(c_i - \langle c \rangle\big)\ , \qquad (20)$$

where H represents the ratio between the standard deviation of the average HP components and the standard deviation of the EC components, and it is evaluated here using a maximum likelihood approach.

Likelihood calculations

We predict the site-specific amino acid distribution at each protein site i generated through the simulation of the evolutionary process using the following system of equations

$$\pi(a|c_i) = \frac{f(a) \exp(-\beta(c_i) h(a))}{\sum_{a'} f(a') \exp(-\beta(c_i) h(a'))}\ , \qquad (21)$$

$$\sum_a h(a) \pi(a|c_i) = \big\langle [h(A_i)] \big\rangle + H\big(c_i - \langle c \rangle\big)\ . \qquad (22)$$

Here, the EC profile c_i structurally characterizes site i, the hydropathy scale $h(a)$ is obtained from the main eigenvector of the contact interaction matrix $U(a,b)$ used in the simulation (it is the eigenvector corresponding to the

most negative eigenvalue), the selection coefficient $\beta(c_i)$ has to be determined imposing that the mean value of the distribution π is equal to the prediction based on the EC profile, the parameter $\langle [h(A_i)] \rangle$ is determined by averaging the hydrophobicity profile $h(A_i)$ over the evolutionary trajectory (square brackets) and over all sites (angular brackets), and the mutation frequencies $f(a)$ can be analytically computed through Eq. (19) from the parameters of the mutation model used in the simulation.

In the above equation, the only free parameter is the hydrophobicity parameter H. We choose the value of this parameter by maximizing the sum of the log-likelihoods of the observed distributions given the predicted ones. For making the computation numerically feasible, we assume that distributions at different sites i are independent, and that the likelihood to observe $n_i(a)$ residues of type a at site i, given that the true probability is $\pi_i(a)$, is given by a multinomial model

$$P(\{n_i(a)\}) = N_i! \prod_a \frac{[\pi_i(a)]^{n_i(a)}}{n_i(a)!} \tag{23}$$

(here, $N_i \equiv \sum_a N_i(a)$ is the number of aligned sequences observed at site i). The log-likelihood at site i is then given by

$$\mathcal{L}_i = \log(N_i!) - \sum_a \log(n_i(a)!) + \sum_a n_i(a) \log(f(a)) - N_i \beta(c_i, H)[h_i]_{\text{obs}} - N_i \log(Z(c_i, H)) \,, \tag{24}$$

where $[h_i]_{\text{obs}}$ is the observed site-specific mean hydrophobicity, which can be reliably estimated even for a small number of protein sequences as

$$[h_i]_{\text{obs}} \equiv \frac{\sum_a h(a) n_i(a)}{N_i} \,. \tag{25}$$

The normalization factor $Z(c_i, H)$ is given by

$$Z(c_i, H) = \sum_a f(a) \exp(-\beta(c_i, H) h(a)) \tag{26}$$

To compute the partial derivative of the log-likelihood with respect to the hydrophobicity parameter H, we have to compute the partial derivative of the selection coefficient β with respect to H using Eq. (22) and the implicit function theorem, finding

$$\frac{\partial \beta(c_i, H)}{\partial H} = \frac{\langle [h]_{\text{obs}} \rangle (c_i - \langle c \rangle)}{\left([h_i^2]_{\text{pred}} - [h_i]^2_{\text{pred}} \right)} \,, \tag{27}$$

where $[h_i]_{\text{pred}} \equiv \langle [h]_{\text{obs}} \rangle + H (c_i - \langle c \rangle)$ is the predicted site-specific average hydrophobicity, Eq. (22), and $[h_i^2]_{\text{pred}} \equiv \sum_a [h(a)]^2 \pi_i(a)$. Inserting the above expression in the partial derivative of the log-likelihood with respect to H, we find

$$\frac{\partial \left(\sum_i \mathcal{L}_i \right)}{\partial H} = \langle [h]_{\text{obs}} \rangle \sum_i N_i \frac{\left([h_i]_{\text{obs}} - [h_i]_{\text{pred}} \right) (c_i - \langle c \rangle)}{\left([h_i^2]_{\text{pred}} - [h_i]^2_{\text{pred}} \right)} \,. \tag{28}$$

This expression can be computed numerically in order to maximize the total likelihood as a function of H through a gradient based method.

Structure alignment and similarity scores

In this work, we use the Mammoth algorithm[25] in order to align protein structures. The algorithm calculates the percentage of superimposed residues, PSI, which is the fraction of aligned residues of the two structures that lie within a 4 Å threshold after optimal rigid body superimposition. Mammoth returns a measure of significance of the alignment, here called the Mammoth score, which is minus the logarithm of the probability that an alignment of unrelated proteins yields the measured PSI, given the length of the shortest protein. This amounts to normalizing the PSI in such a way that the Mammoth score for unrelated protein pairs follows an extreme value distribution almost independent of chain length.[25] Mammoth scores above 4 suggest a biologically significant structural relationship between two proteins.

The second structural similarity score is the contact overlap, which can be calculated after structural alignment as the normalized number of common contacts through Eq. (16). Unlike the Mammoth score, the contact overlap induces a distance measure $D(\mathbf{C}, \mathbf{C}') = 1 - q(\mathbf{C}, \mathbf{C}')$ that has the properties that $D(\mathbf{C}, \mathbf{C}') \geq 0$, $D(\mathbf{C}, \mathbf{C}) = 0$, $D(\mathbf{C}, \mathbf{C}') = D(\mathbf{C}', \mathbf{C})$, and it fulfills the triangular inequality. However, it is more difficult to normalize this measure in such a way that the overlap of unrelated structures does not depend on their length. Since the average overlap decreases with chain length, we define the overlap similarity score as

$$Q(\mathbf{C}', \mathbf{C}) = q(\mathbf{C}', \mathbf{C}) \sqrt{\min(L, L')} \,, \tag{29}$$

and verified that its average value for unrelated pairs is almost independent of $\min(L, L')$.

Last, we defined the EC similarity score as the correlation coefficient between the EC profiles of two aligned proteins over their aligned positions. For a large number of observations, i.e., large number of aligned sites L_{ali}, and for statistically independent variables, the correlation coefficient multiplied by $\sqrt{L_{\text{ali}}}$ tends to a normal variable, i.e., a Gaussian variable with mean zero and standard deviation one. Therefore, we define the significance of the EC score through

$$r(\mathrm{EC},\mathrm{EC}') = \sqrt{L_{\mathrm{ali}}}\,\frac{\sum_i \mathrm{EC}_{a(i)}\mathrm{EC}'_{b(i)} - \sum_i \mathrm{EC}_{a(i)}\sum_i \mathrm{EC}'_{b(i)}/L_{\mathrm{ali}}}{\sqrt{\left[\sum_i (\mathrm{EC}_{a(i)})^2 - \left(\sum_i \mathrm{EC}_{a(i)}\right)^2/L_{\mathrm{ali}}\right]\left[\sum_i (\mathrm{EC}_{b(i)})^2 - \left(\sum_i \mathrm{EC}_{b(i)}\right)^2/L_{\mathrm{ali}}\right]}} \tag{30}$$

where i labels positions in the alignment excluding the initial and final gap, whose number is L_{ali}, $a(i)$ denotes the site in protein a at the i-th position of the alignment, and the EC at gap positions is set to zero.

Protein sets

We simulated eight proteins with the SCN model. Four of these proteins are multidomain, with PDB codes 1pda (296 residues), 1fnb (296 residues), 8atc, chain A (310 residues) and 9wga, chain A (171 residues). Two of them are small single-domain proteins, with PDB codes 1opd (85 residues) and 3lzt (lysozyme, 129 residues). Two are classified as single-domain proteins, but have internal modularity, with PDB codes 1aqt (ATP synthase epsilon chain, gene atpe, which is a 135 residues beta structure with attached two-helix bundle), and 1a6g (myoglobin, 151 residues, where the cavity of the heme group, which we do not represent in the structural model used to simulate protein evolution, divides the protein in two parts).

As a large database for statistical analysis, we used the nonredundant set NR50 of PDB entries having less than 50% pairwise identity, which was downloaded from the PDB web site. From this set we excluded structures solved through NMR spectroscopy, whose contact matrices have slightly different properties from X-ray structures,[27] and structures that are not globular. This last condition was imposed through the following empirical inequalities

$$-0.7 \le \frac{N_{\mathrm{cont}}}{L} - \left(4.00 - 8.07L^{-1/3}\right) \le 0.8 \tag{31}$$

In this way, we selected 7465 protein structures for statistical analysis.

For the analysis of EC conservation, we used the database ASTRAL40,[35] containing protein domains parsed and classfied in the SCOP database and having less than 40% sequence identity. From this database, we constructed all pairs of evolutionary related proteins in the same SCOP superfamily, making a total of 56,450 protein pairs.

Revised PE

The revised PE[9] is computed starting from the contact matrix C_{ij} through the following steps:

1. The contact matrix is modified by setting $C'_{ij} = \varepsilon(L)$ for residues i and j that are not in contact. The length dependent parameter $\varepsilon(L)$ is calculated as $\varepsilon(L) = \min\{0.01, 0.02/[\log(L)-2]\}$.
2. The principal eigenvector y_i of the modified contact matrix is obtained.
3. The components y_i are transformed nonlinearly as $c'_i = \mathcal{G}^{-1}(y_i)$, see Ref. 9, and afterwards normalized so that the mean over the structure is 1.

The functional form of $\varepsilon(L)$, the transformation $\mathcal{G}^{-1}(y)$, and the values of the parameters were determined by imposing two conditions: (i) The revised PE should coincide with the original PE for small single-domain structures, and (ii) the site-specific amino acid distributions based on the revised PE should fit well the amino acid distributions obtained from a large nonredundant subset of the PDB containing both single- and multi-domain folds.[9]

REFERENCES

1. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
2. Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in 3-dimensional protein-structure. J Mol Biol 1983;171:479–488.
3. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known 3-dimensional structure. Science 1991;253:164–170.
4. Rost B, Sander C. Progress of 1d protein-structure prediction at last. Proteins 1995;23:295–300.
5. Bowie JU, Clarke ND, Pabo CO, Sauer RT. Identification of protein folds – matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 1990;7:257–264.
6. Wilmanns M, Eisenberg D. 3-dimensional profiles from residue-pair preferences—identification of sequences with beta/alpha-barrel fold. Proc Natl Acad Sci USA 1993;90:1379–1383.
7. Huang ES, Subbiah S, Levitt M. Recognizing native folds by the arrangement of hydrophobic and polar residues. J Mol Biol 1995; 252:709–720.
8. Bastolla U, Porto M, Roman HE, Vendruscolo M. The principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins 2005;58:22–30.
9. Teichert F, Porto M. Vectorial representation of single- and multi-domain protein folds. Eur Phys J B 2006;54:131–136.
10. Teichert F, Bastolla U, Porto M. SABERTOOTH: Protein structural alignment based on a vectorial structure representation. BMC Bioinformatics 2007;8:425. Also Available at: http://www.fkp.tu-darmstadt.de/SABERTOOTH/.
11. 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.
12. 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.
13. 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.
14. D'Onofrio G, Jabbari K, Musto H, Bernardi G. The correlation of protein hydropathy with the base composition of coding sequences. Gene 1999;238:3–14.
15. Bastolla U, Moya A, Viguera E, van Ham RCHJ. Genomic determinants of protein folding thermodynamics. J Mol Biol 2004;343: 1451–1466.
16. Rost B. Protein structures sustain evolutionary drift. Fold & Des 1997;2:S19–S24.
17. Bastolla U, Roman HE, Vendruscolo M. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J Theor Biol 1999;200:49–64.
18. Bastolla U, Porto M, Roman HE, Vendruscolo M. Statistical properties of neutral evolution. J Mol Evol 2003;57:S103–S119.
19. 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.
20. 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.
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein Data Bank. Nucl Ac Res 2000;28:235–242.
22. Bastolla U, Demetrius L. Stability constraints and protein evolution: the role of chain length, composition, and disulphide bonds. Prot Eng Des Sel 2005;18:405–415.
23. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold & Des 1997;2:173–181.
24. Alexandrov N, Shindyalov I. PDP: protein domain parser. Bioinformatics 2003;19:429–430.
25. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for protein structure comparison. Prot Sci 2002;11:2606–2621.
26. 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.
27. 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.
28. Derrida B. Random Energy Model: an exactly solvable model of disordered systems. Phys Rev B 1981;24:2613.
29. Shakhnovich EI, Gutin AM. Biophys Chem 1989;34:187.
30. Kimura M. Evolutionary rate at the molecular level. Nature 1968;217:624–626.
31. Kimura M. The neutral theory of molecular evolution. New York: Cambridge University Press, 1983.
32. Hasegawa M, Kishino H, Yano T. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22: 160–174.
33. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein coding DNA sequences. Mol Biol Evol 1994;11: 725–736.
34. Lobry JR. Properties of a general model of DNA evolution under no-strand-bias conditions J Mol Evol 1995;40:326–330.
35. Brenner SE, Koehl P, Levitt M. The Astral compendium for protein structure and sequence analysis. Nucl Acid Res 2000;28: 254–256.