Prediction of Site-Specific Amino Acid Distributions and Limits of Divergent Evolutionary Changes in Protein Sequences

Markus Porto,* H. Eduardo Roman,† Michele Vendruscolo,‡ and Ugo Bastolla§

*Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstr. 8, 64289 Darmstadt, Germany; † Dipartimento di Fisica and INFN, Università di Milano, Via Celoria 16, 20133 Milano, Italy; ‡ Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK; § Centro de Astrobiología (INTA-CSIC), 28850 Torrejón de Ardoz, Spain

We derive an analytic expression for site-specific stationary distributions of amino acids from the structurally constrained neutral (SCN) model of protein evolution with conservation of folding stability. The stationary distributions that we obtain have a Boltzmann-like shape, and their effective temperature parameter, measuring the limit of divergent evolutionary changes at a given site, can be predicted from a site-specific topological property, the principal eigenvector of the contact matrix of the native conformation of the protein. These analytic results, obtained without free parameters, are compared with simulations of the SCN model and with the site-specific amino acid distributions obtained from the Protein Data Bank. These results also provide new insights into how the topology of a protein fold influences its designability, i.e., the number of sequences compatible with that fold. The dependence of the effective temperature on the principal eigenvector decreases for longer proteins, as a possible consequence of the fact that selection for thermodynamic stability becomes weaker in this case.

Introduction

The reconstruction of phylogenetic distances from sequence alignments requires the use of a model of protein evolution. Evolutionary models are also needed for reconstructing phylogenetic trees in the framework of maximum likelihood methods (Felsenstein 1981). In this context, both the mutational process and the selection on protein folding and function must be taken into account. It is well known that the local environment of a protein site within the native structure influences the probability of acceptance of a mutation at that site (Overington et al. 1990). Nevertheless, structural biology still has a very limited impact in studies of phylogenetic reconstruction, and the models commonly used in these studies rely on substitution matrices that do not consider the structural specificity of different sites. The most used substitution matrices, such as JTT (Jones, Taylor, and Thornton 1992), are obtained extrapolating substitution patterns observed for closely related sequences, and they have low performances when distant homologs are concerned (Henikoff and Henikoff 1993).

To take into account the protein-level selection, it is necessary to consider site-specific amino acid distributions within a protein family (Halpern and Bruno 1998). Similarly, the use of site-specific substitution matrices improves substantially maximum likelihood methods for reconstructing phylogenetic trees (Liò and Goldman 1998; Koshi and Goldstein 1998; Koshi et al. 1999; Thorne 2000; Fornasari et al. 2002).

In the studies cited above, site-specific constraints are obtained either through simulations of a model of protein evolution or by fitting parameters in a maximum likelihood framework. In this work we derive an analytic expression, without adjustable parameters, for site-specific amino acid distributions and show that they are in very good agreement with simulations of a model of protein evolution and with site-specific amino acid distributions obtained from the Protein Data Bank (PDB).

The site-specific evolutionary patterns that we predict are associated with a key topological indicator of native state topology, namely the principal eigenvector of the contact matrix (Bastolla et al. 2004; Porto et al. 2004). Our approach is based on the structurally constrained neutral model (SCN model; Bastolla, Vendruscolo, and Roman 1999, 2000a; Bastolla et al. 2002, 2003a, 2003b), where possible mutations are tested for conservation of structural stability using a computational model of protein folding (Bastolla et al. 1998; Bastolla, Vendruscolo, and Knapp 2000b; Bastolla, Knapp, and Vendruscolo 2001). This approach is similar in spirit to the structurally constrained protein evolution model (SCPE; Parisi and Echave 2001), which uses a different criterion to assess the thermodynamic stability. Notice that the SCN model does not assume independence at different sites; in fact, we have recently shown that site dependence is linked to a rather debated feature of neutral evolution, the overdispersion of neutral substitutions (Bastolla et al. 2003b).

Site-specific amino acid distributions with Boltzmann form, similar to those that we derive here, have already been used in other studies of protein evolution—for instance by Koshi and Goldstein (1998), Koshi, Mindell, and Goldstein (1999), and Dokholyan et al. (2001, 2002). As we will discuss, the main difference between our approach and these previous models consists in the fact that we compute the Boltzmann parameters analytically, without the need for a database of sequence families.

As we have previously shown (Bastolla et al. 2003a), the site-specific conservation patterns provided by the SCN model, and therefore also by the present analytic formulation, match qualitatively structural conservation patterns found in bioinformatics studies (Ptitsyn 1998; Ptitsyn and Ting 1999). Further, the site-specific amino acid distribution that we present here also enables us to derive an estimate of the sequence entropy compatible...
with a given fold, which is termed the “designability” of the fold (Li, Tang, and Wingreen 1998; Helling et al. 2001). These results are in agreement with recent studies suggesting that the designability can be inferred from the protein topology alone (Koehl and Levitt 2002; England and Shakhnovich 2003).

**Background**

In the SCN model (Bastolla et al. 2002, 2003a), starting from the protein sequence in the PDB, amino acid mutations are randomly proposed and accepted according to a stability criterion based on an effective model of protein folding (Bastolla et al. 2000b, 2001). We use a coarse-grained representation of protein structures as a binary contact matrix $C_{ij}$ with elements equal to one if residues $i$ and $j$ are in contact (at least one pair of heavy atoms, one for each amino acid, are less than 4.5 Å apart), and zero otherwise. The effective free energy for a sequence $A$ in configuration $C$ is approximated by an effective contact free energy function $E(A, C)$, 

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

where $U$ is a $20 \times 20$ symmetric matrix with $U(a, b)$ representing the effective interaction, in units of $k_BT$, of amino acids $a$ and $b$ when they are in contact; we use the interaction matrix derived by Bastolla, Knapp, and Vendruscolo (2001). The free energy function given by equation (1) assigns lowest free energy to the experimental native structure, provided that all contacts are present.

For testing the stability of a protein conformation, we use two computational parameters: (1) The effective energy per residue, $E(A, C)/N$, where $N$ is the protein length, which correlates strongly with the folding free energy per residue for a set of 18 small proteins that are folding with two-states thermodynamics (correlation coefficient 0.91; U. Bastolla, unpublished result); (2) The normalized energy gap $\alpha$, which characterizes fast folding model sequences (Bastolla et al. 1998) with respect to the normalized energy landscapes (Bryngelson and Wolynes 1987; Goldstein, Luthey-Schulten, and Wolynes 1992; Abkevich, Gutin, and Shakhnovich 1994; Gutin, Abkevich, and Shakhnovich 1995; Klimov and Tirumalai 1996). A mutated sequence is considered thermodynamically stable if both computational parameters are above predetermined thresholds (Bastolla et al. 2003a).

The interaction matrix $U$ can be written in spectral form as $U(a, b) = \sum_{\alpha=1}^{20} e_\alpha u^{(\alpha)}(a) u^{(\alpha)}(b)$, where $e_\alpha$ are the eigenvalues, ranked by their absolute value, and $u^{(\alpha)}$ are the corresponding eigenvectors. The main contribution to the interaction energy comes from the principal eigenvector $u^{(1)}$, which has a negative eigenvalue $e_1 < 0$, as $e_1 u^{(1)}(a) u^{(1)}(b)$ has correlation coefficient 0.81 with the elements $U(a, b)$ of the full matrix. It is well known that hydrophobic interactions determine the most significant contribution to pairwise interactions in proteins, so that the components of the main eigenvector are strongly correlated with experimental hydrophathy scales (Casari and Sippl 1992; Li, Tang, and Wingreen 1997). Considering only this main component, we define an approximate effective energy function $H(A, C)$, yielding a good approximation to the full contact energy given by equation (1),

$$\frac{H(A, C)}{k_BT} = e_1 \sum_{i<j} C_{ij} h(A_i) h(A_j).$$

We call $h(A) = u^{(1)}(A)$ the hydrophobicity profile (HP) of sequence $A$ (Bastolla et al. 2004). This is an $N$-dimensional vector whose $i$th component is given by $h(A_i) = u^{(1)}(A_i)$. The 20 parameters $h(a) = u^{(1)}(a)$ are obtained from the principal eigenvector of the interaction matrix, and we call them *interactivity* parameters.

The HP provides a vectorial representation of protein sequences. In turn, a convenient vectorial representation of protein structures may be obtained by the principal eigenvector of the contact matrix $C$, which we denote by PE and we indicate by $c$. This vector maximizes the quadratic form $\sum_i c_i c_j$ with the constraint $\sum_i c_i^2 = 1$. In this sense, $c_i$ can be interpreted as the effective connectivity of position $i$, since positions with large $c_i$ are in contact with as many as possible positions $j$ with large $c_j$. All its components have the same sign, which we choose by convention to be positive. Moreover, if the contact matrix represents a single connected graph (as it does for single-domain globular proteins), the information contained in the principal eigenvector is sufficient to reconstruct the whole contact matrix (Porto et al. 2004).

For any given protein fold, identified through its PE, we define the optimal HP, $h_{\text{opt}}$, as the HP that minimizes the approximate effective free energy, given by equation (2), for fixed first and second moment of the hydrophobicity vector, $(h) = N^{-1} \sum h(A_i)$ and $(h^2) = N^{-1} \sum h(A_i)^2$. We impose a condition on the mean hydrophobicity, $(h)$, because, if a sequence is highly hydrophobic, not only the native structure but also alternative compact structures will have favorable hydrophobic energies. The selection to maintain a large normalized energy gap is therefore expected to place constraints on the value of $(h)$.

The above property of the PE, $c$, suggests that the optimal HP, $h_{\text{opt}}$, should be strongly correlated with the PE (Bastolla et al. 2004). In this formulation, $(h)$ and $(h^2)$ are not determined by the native structure but depend on the mutation and selection process (see below).

The optimal HP constitutes an analytic solution to the sequence design problem with energy function given by equation (2), and an approximate solution to sequence design with the full contact energy function given by equation (1). In the SCN evolutionary model, attempted mutations are accepted whenever the effective free energy and the normalized energy gap overcome predefined thresholds. Therefore, we do not expect that the optimal HP is ever observed during evolution, but we do expect thermodynamically stable sequences compatible with a given fold to have HPs distributed around the optimal one. We have verified this prediction, finding that the HPs of individual selected sequences are correlated with the principal eigenvector of the protein fold with correlation coefficient of 0.45 (averaged over seven protein folds and...
over hundred thousands of simulated sequences per fold), whereas the HP averaged over all sequences compatible with a given fold, $[h]_{evol}$, correlates much more strongly with the principal eigenvector of that fold, with mean correlation coefficient 0.96 averaged over the same seven folds (Bastolla et al. 2004). These results show that one can recover the optimal HP through an evolutionary average of the HPs compatible with the protein fold.

We found a similar result for the protein families represented in the PFAM database (Bateman et al. 2000) and in the families of structurally similar proteins (FSSP) (Holm and Sander 1996) database. In this case, however, the correlation between the principal eigenvector and the evolutionary average of the HP is weaker: The average correlation coefficient is 0.57 for PFAM families and 0.58 for FSSP families (Bastolla et al. 2004). This weaker correlation is not unexpected, since functional conservation, which is not accounted for in the SCN model, plays an important role in protein evolution, and the effective energy function that we use to test thermodynamic stability is only an approximation. In addition, the number of sequences used to calculate the average HP in the PFAM and FSSP databases is much smaller than the number of sequences obtained by the SCN model. When the average HP is computed using only a few hundred of SCN sequences, the same order of magnitude as in PFAM or FSSP families, the correlation with the principal eigenvector also drops to values comparable to those observed for the PFAM and the FSSP sequence databases.

**Results**

**Derivation of $\pi_a$**

The results presented above suggest that the SCN model of protein evolution can be represented as a trajectory in sequence space where the HP moves in the vicinity of the optimal HP, which is strongly correlated with the principal eigenvector of the protein fold’s contact matrix (Bastolla et al. 2004). In this work, we use this analogy to compute analytically the site-specific distribution of amino acid occurrences $\pi_a$, where $i$ indicates a protein position and $a$ indicates one of the 20 amino acid types.

Our previous results indicate that the evolutionary average of the hydrophobicity vector, $[h]_{evol}$, is correlated with the principal eigenvector $c$ of the native contact matrix. To derive an analytical expression for $\pi_a$, we now assume that the correlation coefficient between the principal eigenvector and the average HP is 1, i.e.,

$$
\langle \langle h \rangle_{evol} \cdot c \rangle = \langle \langle h \rangle_{evol} \rangle \cdot \langle c \rangle = \langle c \rangle \cdot \langle \langle h \rangle_{evol} \rangle = 1,
$$

and we obtain

$$
[h]_{evol} = \sum_a \pi_a \langle h \rangle_a = A\langle c \rangle / \langle c \rangle - 1 + B,
$$

where the sum over $\{a\}$ is over all amino acids, and

$$
A = \frac{\langle [h]_{evol}^2 \rangle - \langle [h]_{evol} \rangle^2}{\langle c^2 \rangle - \langle c \rangle^2}, \quad B = \langle [h]_{evol} \rangle.
$$

We are representing here two kinds of averages: angle brackets denote the average over the $N$ positions of the protein, $\langle f \rangle = N^{-1} \sum_i f_i$, and the corresponding standard deviation is denoted by $\sigma_f^2 = \langle f^2 \rangle - \langle f \rangle^2$, whereas square brackets denote position-specific evolutionary averages, $\langle [f] \rangle_{evol} = \sum_a \pi_a \langle f \rangle_a$.

Equations (3,4) are the conditions that the stationary distributions $\pi_a$ have to fulfill in order to guarantee a perfect correlation between the principal eigenvector and the average HP. We assume that these conditions are the only requirements that the $\pi_a$ have to meet, so that at stationarity the $\pi_a$ are the distributions of maximum entropy compatible with the above conditions. It is well known that the solution of this optimization problem are Boltzmann-like (exponential) distributions, characterized by an effective “temperature” $[\beta]^{-1}$ that, in this context, varies from site to site and measures the tolerance of site $i$ to accept mutations over very long evolutionary times,

$$
\pi_a = \frac{\exp[-\beta \langle h \rangle_a]}{\sum_b \exp[-\beta \langle h \rangle_b]},
$$

with the constraint given by equation (3),

$$
\sum_a \exp[-\beta \langle h \rangle_a] \langle h \rangle_a - A \langle c \rangle / \langle c \rangle - 1 - B = 0.
$$

Equation (6) states an analytical relation between the “Boltzmann parameter” $\beta$, and the principal eigenvector component $c_i$, given the two evolutionary parameters $A$ and $B$. One sees from this equation that $\beta$ equals zero if $c_i / \langle c \rangle = 1 + B^{-1}(\langle h \rangle_{evol} - \langle c \rangle)$, and $\beta$ becomes negative for larger $c_i$ and positive for smaller $c_i$. We expect the relationship between $\beta_i$ and $c_i$ is almost linear in the range $|c_i / \langle c \rangle| < 1$. In addition, $\beta_i$ tends to minus infinity when the average hydrophobicity at site $i$, $\langle h \rangle_{evol}$, tends to the maximally allowed value, and to plus infinity when the average hydrophobicity at site $i$ tends to the minimum allowed value.

Equation (6) can be interpreted as follows: (1) Positions with a large eigenvector component $c_i$ are buried in the core of the protein and are therefore, with high probability, occupied by hydrophobic amino acids (positive $h(a)$) and have large and negative $\beta_i$; (2) surface positions with small $c_i$ are more likely occupied by polar amino acids (negative $h(a)$) and have large and positive $\beta_i$; and (3) intermediate positions are the most tolerant, having a small negative or small positive $\beta_i$.

Similar forms of site-specific amino acid distributions have been previously proposed by other authors; see for instance Koshi and Goldstein (1998), Koshi, Mindell, and Goldstein (1999), and Dokholyan et al. (2001, 2002). We will discuss analogies and differences between these previous approaches and the present one in the last section.

**Validation: SCN Model**

We first verified these predictions using the simulated trajectories obtained through the SCN model. We found that the site-specific stationary distributions of amino acids, $\pi_a$, are well fitted by an exponential function of hydrophobicity, $\pi_a \propto \exp[-\beta_i \langle h \rangle_a]$, where we use the interaction scale given by the main eigenvector of our interaction matrix.
Fig. 1.—“Boltzmann parameter” $\beta_i$ as a function of the scaled principal eigenvector component $c_i/\langle c \rangle$ as obtained by the SCN model for ribonuclease (PDB id. 7rsa). The line shows the analytical prediction given by equation (7), obtained using the mean hydrophobicity $h_{\text{evol}}$ and the variance $\langle h_{\text{evol}}^2 \rangle - \langle h_{\text{evol}} \rangle^2 = 0.0077$ as observed in the simulations of the SCN model. The dashed part of the curve indicates the forbidden area $c_i < 0$. The inset exemplifies the numerical obtained $-\ln(c_i/\langle c \rangle)$ vs. hydrophobicity $h(a)$ of amino acid $a$, as obtained for ribonuclease at protein position $i = 50$ (with $c_{50}/\langle c \rangle = 0.124$), yielding $\beta_{50} = 4.92$.

Analytically, the expected site-dependent Boltzmann parameter $\beta_i$ can be calculated in an implicit form by rewriting equation (6) as

$$c_i/\langle c \rangle = 1 + A^{-1} \left[ \sum_i h(a) \exp(-\beta_i h(a)) - B \right],$$

(7)

giving $c_i$ as a function of $\beta_i$. To obtain $\beta_i$ as a function of $c_i$, Equation (7) has to be inverted numerically, and the parameters $A$ and $B$ are obtained from the simulation data by Equation (4).

The predictions derived in this way, for the ribonuclease fold (PDB id. 7rsa), are compared in figure 1 to the $\beta_i$ obtained by fitting the distributions $\pi_i(a)$ simulated through the SCN model. The prediction does not involve any adjustable parameter, since $A$ and $B$ are calculated from $\langle h_{\text{evol}} \rangle$ and $\langle h_{\text{evol}}^2 \rangle$, as determined by the results of the SCN model. The latter values do not depend on the native structure, but they depend generally on the mutation and selection process, and specifically on its realization as simulated through the SCN model. The other proteins studied show a similar behavior, and the mean correlation coefficient between predicted and observed $\beta_i$ is 0.935 (cf. table 1).

One notices from figure 1 that $\beta_i$ depends almost linearly on the scaled principal eigenvector component $c_i/\langle c \rangle$ over a wide range of values. The slope of this linear part, the derivative $-\partial \beta_i/\partial c$ taken at $\dot{c} = c/\langle c \rangle = 1$, scales with length $N$, and it can be well fitted by the expression $A_0 + A_1/\langle N \rangle$, with $A_0 = 2.50 \pm 0.21$ and $A_1 = 104.8 \pm 9.2$ for the proteins listed in table 1 (see figure 2a). One sees from this expression that, as the protein length increases, the Boltzmann parameter $\beta_i$ becomes less dependent on $c_i/\langle c \rangle$ and therefore more homogeneous. This result, at first sight unexpected, can be better understood considering equation (7), which implies that the derivative of $\beta_i$ with respect to $c_i = c_i/\langle c \rangle$ is proportional to the standard deviation of the hydrophobicity,

$$\partial \beta_i/\partial c \vert_{c_i = 0} = \sqrt{\langle h_{\text{evol}}^2 \rangle - \langle h_{\text{evol}} \rangle^2} / \langle h_{\text{evol}} \rangle^2.$$  

The standard deviation of the hydrophobicity, measured from simulations of the SCN model, decreases as a function of chain length as $B_0 + B_1/N$ with $B_0 = 0.068 \pm 0.016$ and $B_1 = 2.55 \pm 0.17$ (see figure 2b). Thus, site-specificity decreases for longer proteins because the evolutionary average of the HP becomes more homogeneous across different positions. This result can be explained by noticing that not only the standard deviation but also the mean hydrophobicity, $\langle h_{\text{evol}} \rangle$, decreases as a function of chain length (notice again the two kinds of averages). So, the result that site-dependent Boltzmann parameters become more homogeneous for longer proteins ultimately depends on the fact that longer proteins tend to be less hydrophobic.

### Validation: PDB Structures

We compared our predictions to site-specific distributions obtained from a representative subset of the PDB. We considered a nonredundant subset of single-domain globular proteins in the PDB, with a sequence identity below 25% (Hobohm and Sander 1994). The condition of being globular was verified by imposing that the fraction of contacts per residue was larger than a length-dependent threshold, $N_{\text{C}}/N > 3.5 + 7.8N^{-1/3}$. This functional form represents the scaling of the number of contacts in globular proteins as a function of chain length (the factor $N^{-1/3}$ comes from the surface-to-volume ratio), and the two parameters were chosen to eliminate outliers with respect to the general trend, which are mainly nonglobular structures. The condition of being single-domain was verified by imposing that the normalized variance of the PE components was smaller than a threshold, $(1 - N(c)^2)/N(c)^2 < 1.5$. Multi-domain proteins have PE components which are large inside their main domains and small outside them. The

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB id.</th>
<th>$N$</th>
<th>$\beta$</th>
<th>$R$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rubredoxin (mesophilic)</td>
<td>1ero</td>
<td>53</td>
<td>0.98</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Rubredoxin (thermophilic)</td>
<td>1bfr_A</td>
<td>53</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>SH3 domain</td>
<td>1aey</td>
<td>58</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>451c</td>
<td>82</td>
<td>0.94</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Ribonuclease</td>
<td>7rsa</td>
<td>124</td>
<td>0.94</td>
<td>0.83</td>
<td>0.89</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>3lzt</td>
<td>129</td>
<td>0.93</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>1afg</td>
<td>151</td>
<td>0.95</td>
<td>0.94</td>
<td>0.88</td>
</tr>
<tr>
<td>Ubiquitin conjugating</td>
<td>1u9a_A</td>
<td>160</td>
<td>0.90</td>
<td>0.88</td>
<td>0.79</td>
</tr>
<tr>
<td>TIM barrel</td>
<td>7tim_A</td>
<td>247</td>
<td>0.94</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td>Kinesin</td>
<td>1bg2</td>
<td>323</td>
<td>0.87</td>
<td>0.79</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Note.—Correlation coefficients between site-specific quantities predicted using the equations derived in this work and observed in SCN simulations. Fourth column: “Boltzmann parameters” $\beta$. Fifth column: Rigidities $R$. Sixth column: Substitution rates $r$. Note that sites containing cysteine residues are not included in the calculation of the correlation coefficients, as cysteine residues form disulfide bridges (which are very poorly represented through hydrophobicity) and are strictly conserved in our evolutionary model.
PE components would be exactly zero outside the main domains when the domains do not share contacts. Therefore, multi-domain proteins are characterized by a larger normalized variance of PE components with respect to single-domain ones. We have verified that the threshold of 1.5 is able to eliminate most of the known multi-domain proteins and very few of the known single-domain proteins (data not shown).

We thus selected 774 sequences of various lengths, 404 of which were shorter than 200 amino acids. We first counted the number of each of the 20 amino acids as a whole data set, divided in bins of similar lengths. We tested whether this interesting observation also holds for site-specific distributions obtained from the PDB. To this end, we divided our set of proteins into five classes of lengths considered, the fit was reasonably good, with correlation coefficients always larger than 0.95. The obtained slopes are plotted against 1/N in figure 4. Although the errors are rather large, the data are compatible with the functional form $-\partial \beta /\partial \epsilon \approx A_0 + A_1/N$. The parameters that we obtained by fitting this relationship are also compatible, within error bars, with those obtained from

\begin{align}
\frac{\langle c \rangle}{\langle c \rangle} = 1 + \tilde{A} = 1 + \frac{\sum_a \hat{h}(a) \exp[-\hat{\beta}_c(a)/\hat{\epsilon}(a)]}{\sum_a \exp[-\hat{\beta}_c(a)/\hat{\epsilon}(a)]} - \tilde{B} \tag{9}
\end{align}

where the square brackets $[\hat{h}]_{\text{PDB}}$ now denote, instead of the evolutionary average over a protein family, the average over all positions with fixed $c_f/(c)$, even belonging to different structures, whereas angle brackets, $\langle [\hat{h}]_{\text{PDB}} \rangle_{c_f/c}$, denote the average over all values of $c_f/(c)$. The denominator $\langle (c^2 - \langle c \rangle^2) / (\langle c \rangle^2) \rangle_{\text{PDB}}$ indicates the quantity $\langle (c^2 - \langle c \rangle^2) / (\langle c \rangle^2) \rangle_{\text{PDB}}$, obtained individually for each structure, averaged over the whole set of structures.

The observed Boltzmann parameters are compared in figure 3 to the predictions of equation (9). Notice that the agreement is indeed remarkable, as the predictions do not involve any adjustable parameter, because $\hat{A}$ and $\tilde{B}$ are calculated from the PDB data.

We stated in the previous section that, with SCN data, the dependence of the Boltzmann parameters on the PE components becomes weaker, and the amino acid distributions become more homogeneous, for longer chains. We tested whether this interesting observation also holds for site-specific distributions obtained from the PDB. To this end, we divided our set of proteins into five classes of proteins with less than 100, 200, 300, 400, and 500 amino acids, respectively, and we measured the dependence of the Boltzmann parameters on the PE components for each class. From these data, we obtained the derivative of the Boltzmann parameters at $\hat{\epsilon} = c_f/(c) = 1$ by fitting a straight line to $\hat{\beta}(\hat{c})$ for $\hat{c}$ in the range between 0.5 and 1.5. For all lengths considered, the fit was reasonably good, with correlation coefficients always larger than 0.95. The obtained slopes are plotted against 1/N in figure 4. Although the errors are rather large, the data are compatible with the functional form $-\partial \hat{\beta} /\partial \hat{\epsilon} \approx A_0 + A_1/N$. The parameters that we obtained by fitting this relationship are also compatible, within error bars, with those obtained from

\begin{align}
\hat{A} = \sqrt{\langle [\hat{h}]_{\text{PDB}}^2 \rangle_{c_f/c} - \langle [\hat{h}]_{\text{PDB}} \rangle_{c_f/c}^2 / \langle (c^2 - \langle c \rangle^2) / (\langle c \rangle^2) \rangle_{\text{PDB}}} \tag{10}
\end{align}

$\tilde{B} = \langle [\hat{h}]_{\text{PDB}} \rangle_{c_f/c}$,

\begin{align}
\hat{\beta}(\hat{c}) = \frac{\sum_a \hat{h}(a) \exp[-\hat{\beta}_c(a)/\hat{\epsilon}(a)]}{\sum_a \exp[-\hat{\beta}_c(a)/\hat{\epsilon}(a)]} - \hat{\beta}(1)
\end{align}
SCN data: We find, for PDB structures, $A' = 2.33 \pm 0.17$ and $A_1 = 117 \pm 25$. Although a more careful study is still necessary on this issue, the present results seem to confirm this length effect on the amino acid distributions.

Conservation and Designability

The results that we present imply that there is a direct relationship between a structural indicator, the principal eigenvector of the contact matrix, and site-specific measures of long-term evolutionary conservation, and hence of limits in divergent evolutionary changes. This relationship also provides a link between the topology of a fold and its designability.

One convenient measure of the amino acid conservation at a given position is given by the rigidity, defined as

$$ R_i = \frac{\sum_{a} \pi_i(a)^2}{\left(\sum_{a} \exp(-\beta_i h(a))\right)^{1/2}}. $$

$R_i = 1$ means that the same amino acid is present at position $i$ in all sequences, i.e., the conservation is total and $\beta_i^{-1} = 0$. In general, the rigidity decreases with increasing temperature $[\beta_i]^{-1}$. One can use equations (7,11) to compute the rigidity from the principal eigenvector. However, the dependence becomes clearer when fitting directly the rigidity as a function of the principal eigenvector component. We did this on the data generated through the SCN model, finding that the best fit is parabolic, and the three parameters depend on the protein length. Instead of the PE component $c_i$, we use the rescaled variable $x_i = \sqrt{N} c_i$, which is normalized such that its mean squared value is one; i.e., $\langle x_i^2 \rangle = N^{-1} \sum_i x_i^2 = 1$. Using this scaled variable, the rigidities of all proteins in our data set can be fitted by the expression

$$ R_i^{\text{pred}} = A + \frac{B}{N^d} \left(x_i - C(x_i) - \frac{D}{\sqrt{N}}\right)^2, $$

with $A = 0.0630 \pm 0.0017, B = 0.28 \pm 0.16, C = 1.879 \pm 0.102, D = -5.36 \pm 0.89$, and $d = 0.47 \pm 0.11$. The factor $N^{-d} \approx N^{-1/2}$ can be interpreted as a further indication that rigidities tend to become more homogeneous for larger proteins, as previously noted concerning the Boltzmann parameters. Equation (12) predicts the rigidity for all the proteins that we studied with the SCN model, with an average root mean square relative error of around 10% and an average correlation coefficient between predicted and observed rigidities of 0.88 (cf. table 1). These values change very little in leave-one-out tests, where we discard one protein for estimating the parameters and then use it as a blind test.

A standard information-theoretic measure of site-specific sequence conservation is given by the entropy of the amino acid distribution,

$$ S_i = -\sum_{a} \pi_i(a) \log[\pi_i(a)] $$

$$ = \log[Z(\beta_i)] + \beta_i [h_i]_{\text{eval}}, $$

where $Z(\beta_i) = \sum_a \exp(-\beta_i h(a))$. The entropy attains its maximum value, $S_i = \log(20)$, at $\beta_i = 0$ and it decreases with increasing $[\beta_i]$. Predictions of the entropy based on a different approach, equation (14), using aligned protein families have been obtained by Dokholyan et al. (2001, 2002).

An important property of the entropy is that its exponential, $\exp(S_i)$, provides an estimate of the average number of amino acid types acceptable at position $i$ over very long evolutionary times. The exponential of the sum of all site-specific entropies, $\exp(\sum_i S_i)$, gives an estimate of the sequence space compatible with a given fold, termed the designability of the fold, where we are assuming that the amino acid distributions at different positions are independent. Although this assumption is clearly oversimplified, the estimate of designability that can be obtained should be a valuable approximation, and our approach allows us to connect it explicitly to a topological feature of the protein native structure (Koehl and Levitt 2002; England and Shakhnovich 2003).

Discussion and Conclusions

We have predicted analytically that amino acids at specific positions are distributed according to a Boltzmann law. In the latter, the role of energy is played by the hydrophobicity, measured through an interactivity scale, and that of temperature by a quantity strongly correlated with a structural indicator, namely the corresponding component of the principal eigenvector of the native contact matrix (PE). This prediction, which does not involve any free parameter, is in very good agreement both with simulations of the SCN model of protein evolution and with site-specific amino acid distributions that we obtained from a representative subset of single-domain globular proteins in the PDB.

The relationship between a structural profile (the PE) and an evolutionary profile (the Boltzmann parameters) that we find has an interesting interpretation. Positions with a large principal eigenvector component are strongly interacting with other positions in the core of the protein. Therefore, they preferentially host strongly hydrophobic amino acids (large negative $\beta_i$). In contrast, positions with a small PE are contained in loops and, with higher probability, they host polar amino acids (large positive $\beta_i$).
Despite this general interpretation, we have verified, using the SCN model of protein evolution, that the PE has the strongest predictive power among several similar structural indicators that quantify the difference between core and surface positions of a protein. As alternative structural indicators, we considered the total number of contacts, the number of long-range contacts (separated by more than 10 residues along the sequence), and the contact order (the average loop length of each contact involving a given site), and we correlated them with measures of site-specific conservation, but we always found a correlation significantly weaker than for the PE.

The distributions derived here refer to very long evolutionary times, when memory of the starting sequence has been lost. We recall the assumptions that we made for deriving the site-specific distributions. The first assumption is that selection on folding stability can be represented effectively as a maximum correlation between the hydrophobicity profile (HP) of sequences compatible with a given fold and the optimal HP of that fold, the latter basically coinciding with the PE. This assumption follows directly from an approximation of our effective free energy function with its principal (hydrophobic) component. The second assumption is that the average of the HP of selected sequences over very long evolutionary times has a correlation coefficient of unity with the PE; i.e., all other energetic contributions average out. The third assumption is that this correlation is the only relevant property of the site-specific amino acid distributions; in other words, these distributions are the distributions of maximum entropy whose site-specific averages have correlation of one with the PE, thus fulfilling the stability requirement. From these three assumptions, the Boltzmann form of the amino acid distributions follows. To compute the site-specific Boltzmann parameters, however, we still have to know the positional mean and standard deviation of the site-specific HPs. These quantities are determined by the mutation process and by the selection parameters. They can be computed directly from the data, in such a way that the analytic prediction does not contain any free parameter.

Boltzmann distributions, such as those proposed in this work, have a long history in studies of protein structure and evolution. Structural properties of native protein structures, as for instance amino acid contacts, have been assumed to be Boltzmann-distributed (Miyazawa and Jermigan 1985), and Boltzmann statistics for structural elements were predicted in stable folds of globular proteins (Finkelstein, Badreddinov, and Gutin 1995). Our work may point to an additional explanation for such distributions.

Shakhnovich and Gutin (1993) proposed a model of sequence design through Monte Carlo optimization, which produced a Boltzmann distribution in sequence space. A mean field approximation of this model (Dokholyan et al. 2001, 2002) results in site-specific amino acid distributions of the form

\[
\pi_i(a) \propto \exp[-\beta \phi_i(a)], \tag{14}
\]

formally similar to equation (5). There are, however, three important differences between our formulation and equation (14). First, equation (14) was derived as an site-independent approximation to a Boltzmann distribution for entire sequences, whereas we derived equation (5) from the relationship between average hydrophobicity at a given site and the PE component. Second, in equation (14), the Boltzmann parameter $\beta$ is the same for all sites, whereas we obtain the $\beta_i$ profile along the protein structure from the PE. Third and most important, to compute equation (14), Dokholyan et al. (2001, 2002) used aligned families of natural proteins, whereas our computation only needs the PE and two empirical values, the average and the standard deviation along the hydrophobicity profile.

Koshi and Goldstein (1998) and Koshi, Mindell, and Goldstein (1999) assumed site-dependent Boltzmann distributions of physicochemical amino acid properties, deriving from it a protein evolution model that they use for phylogenetic reconstruction in a maximum likelihood framework. Because the properties they used are hydrophobicity and amino acid size, their proposed distributions are formally a general case of those derived in this work. There are, however, two important differences between their approach and ours: first, in our approach sites are structurally classified according to the PE component, which is a structural indicator very strongly correlated with conservation; second, whereas the Boltzmann parameters are treated as free parameters by Koshi, Mindell, and Goldstein and fitted in a maximum likelihood framework, in our approach they are computed analytically.

Kinjo and Nishikawa have very recently pointed out the existence of a strong relationship between hydrophobicity and the main eigenvector of substitution matrices derived from protein alignments with various values of the sequence identity of the aligned proteins (Kinjo and Nishikawa 2004). They looked at the eigenvector corresponding to the largest eigenvalue (in absolute value) of the substitution matrices. For high sequence identities (above 35%), this eigenvector indicates the propensity of the amino acid to mutate over short evolutionary times (mutability). For low sequence identities (below 35%), corresponding to long evolutionary times, this eigenvector is very strongly correlated with hydrophobicity. This correlation is easily understood in light of the result presented here. In fact, Kinjo and Nishikawa used the Henikoff and Henikoff’s method (Henikoff and Henikoff 1992) for deriving substitution matrices from observed frequencies of aligned amino acids at positions with various PE values. Using our notations, these substitution matrices can be indicated as $M(a, b) = \log[(\pi_i(a) \pi_i(b))/(\pi_i(a)\pi_i(b))],$ where the angle brackets denote positional average. In other words, these substitution matrices measure the tendency of two residue types $a$ and $b$ to co-occur at the same sites. The relationship between large time substitution matrices and hydrophobicity therefore gives independent support to our main result.

In dealing with protein families generated from natural evolution, our approach has two main difficulties: first, it cannot take into account functional conservation, and second, a large number of sequence pairs separated by very long evolutionary time is required. Because of these difficulties, the site-specific conservation predicted by our model agrees only qualitatively with site-specific conservation observed in aligned protein families (Bastolla et al. 2003a). For tackling these difficulties, we tested our
analytic predictions on a large set of proteins in the PDB. Positions belonging to different protein folds were counted together in the same structural class characterized by the value of the normalized PE.

The agreement between the analytic prediction and the observed distributions may be surprising in view of the fact that equation (9) takes into account neither the genetic code nor the codon usage. On the one hand, it is known that the observed frequencies of amino acids correlate strongly with their number of synonymous codons, and even more strongly with their cumulative codon frequencies expected in the simple hypothesis that the three bases constituting each codon are independently distributed (Sueoka 1961). Moreover, the frequency of nucleotides at coding positions correlate strongly with the frequency of nucleotides at noncoding positions such as the synonymous 3rd codon positions (Bernardi and Bernardi 1986), which are thought to reflect the mutational process. On the other hand, mutations are not adequately represented in the SCN model, which is a model at the amino acid level where all amino acid mutations are equiprobable.

The most straightforward way to include the biases of the mutational process is to modify equation (5) with

\[ \pi_i = w_i \exp(-\beta_i \mathit{h}_i) \]

where the weights \( w_i \) are either the number of synonymous codons or their cumulative frequency, estimated using average nucleotide compositions. We will address this issue in future work.

We also observed a length effect in the dependence of the Boltzmann parameters \( \beta_i \) on the PE. For longer proteins, this dependence becomes weaker, approximately following a law of the type \( A_0 + A_1/N \). This result holds for SCN simulations and, although less significantly, for frequency distributions obtained from the PDB. Formally, the result directly follows from the fact that the evolutionary average of the hydrophobicity, \( \mathit{h}_i \), varies less and less across different positions \( i \) for longer proteins. This length dependence of the HP is consistent with the fact that the position average of \( \mathit{h}_i \) decreases with chain length; i.e., the length of a protein influences its composition (White 1992; Bastolla and Demetrius, in preparation), such that longer proteins tend to become less hydrophobic. (We note, however, that these results were obtained using the interactivity scale, and they cannot be significantly generalized to other hydrophathy scales.) These results can be understood considering that longer proteins are stabilized by a larger number of interactions per residue. Selection for protein stability is therefore expected to be weaker on individual interactions. In support of this prediction, the average interaction energy per contact in PDB structures was found to decrease with chain length (Bastolla and Demetrius, unpublished data).

The site-specific amino acid frequencies that we derived may find applications in the estimation of site-specific substitution matrices (Bateman et al. 1998; Bateman, Birney, Durbin, Eddy, Howe, and Sonnhammer 2002), and site-specific structural conservation. Our approach may provide an analytic understanding of the limits imposed by structural and functional constraints to the divergent evolution of protein sequences and of the reliability of phylogenetic reconstructions from protein sequences for anciently diverged taxa (Meyer, Cusanovich, and Kamen 1986). A better understanding of structural conservation may also improve the prediction of functional conservation at sites that appear more conserved than expected on a structural basis alone (Casari, Sander, and Valencia 1995; Ota, Kinoshita, and Nishikawa 2003). The dependence of amino acid distributions on the PE that we presented here is also a step toward determining the structural determinants of protein “designability” (Li, Tang, and Wingreen 1998; Helling et al. 2001; England and Shakhnovich 2003), namely the volume of sequence space compatible with a given fold.

Acknowledgments

M.P. gratefully acknowledges financial support by the guest program of the Max-Planck-Institut für Physik komplexer Systeme in Dresden, Germany, during the early stages of this project. M.V. is supported by the Royal Society and by the Leverhulme Trust. U.B. is sponsored by the I3P program of the Spanish CSIC co-financed by the European Social Fund. We gratefully acknowledge discussions with Alfonso Valencia, Julian Echave, and Gustavo Parisi, and correspondence with Akira Kinjo.

Literature Cited


Peter Lockhart, Associate Editor

Accepted October 15, 2004