BMC Evolutionary Biology

BioMed Central

Research article

Open Access

A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank

Ugo Bastolla*[1], Markus Porto[2], H Eduardo Roman[3] and Michele Vendruscolo[4]

Address: [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, [3]Dipartimento di Fisica, Università di Milano Bicocca, Piazza della Scienza 3, 20126 Milano, Italy and [4]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

Email: Ugo Bastolla* - ubastolla@cbm.uam.es; Markus Porto - porto@fkp.tu-darmstadt.de; H Eduardo Roman - roman@mib.infn.it; Michele Vendruscolo - mv245@cam.ac.uk

* Corresponding author

Published: 31 May 2006

BMC Evolutionary Biology 2006, **6**:43 doi:10.1186/1471-2148-6-43

Received: 17 November 2005

Accepted: 31 May 2006

This article is available from: http://www.biomedcentral.com/1471-2148/6/43

Abstract

Background: Since thermodynamic stability is a global property of proteins that has to be conserved during evolution, the selective pressure at a given site of a protein sequence depends on the amino acids present at other sites. However, models of molecular evolution that aim at reconstructing the evolutionary history of macromolecules become computationally intractable if such correlations between sites are explicitly taken into account.

Results: We introduce an evolutionary model with sites evolving independently under a global constraint on the conservation of structural stability. This model consists of a selection process, which depends on two hydrophobicity parameters that can be computed from protein sequences without any fit, and a mutation process for which we consider various models. It reproduces quantitatively the results of Structurally Constrained Neutral (SCN) simulations of protein evolution in which the stability of the native state is explicitly computed and conserved. We then compare the predicted site-specific amino acid distributions with those sampled from the Protein Data Bank (PDB). The parameters of the mutation model, whose number varies between zero and five, are fitted from the data. The mean correlation coefficient between predicted and observed site-specific amino acid distributions is larger than $\langle r \rangle = 0.70$ for a mutation model with no free parameters and no genetic code. In contrast, considering only the mutation process with no selection yields a mean correlation coefficient of $\langle r \rangle = 0.56$ with three fitted parameters. The mutation model that best fits the data takes into account increased mutation rate at CpG dinucleotides, yielding $\langle r \rangle = 0.90$ with five parameters.

Conclusion: The effective selection process that we propose reproduces well amino acid distributions as observed in the protein sequences in the PDB. Its simplicity makes it very promising for likelihood calculations in phylogenetic studies. Interestingly, in this approach the mutation process influences the effective selection process, i.e. selection and mutation must be entangled in order to obtain effectively independent sites. This interdependence between mutation and selection reflects the deep influence that mutation has on the evolutionary process: The bias in the mutation influences the thermodynamic properties of the evolving proteins, in agreement with comparative studies of bacterial proteomes, and it also influences the rate of accepted mutations.

Background

The evolutionary information embedded in the sequences of extant biological macromolecules can be used to reconstruct their evolutionary history (see for instance Ref. [1,2]). Methods based on the Maximum Likelihood (ML) principle are quite successful in reconstructing the past of molecules and species [3], but they rely on models of the evolutionary process at the molecular level. ML computations are usually carried out assuming that the sites of a protein evolve independently, a feature that is rather unrealistic, since selection for thermodynamic stability of the native state acts on all sites at the same time, introducing correlations between sites [4,5].

Incorporating selection for thermodynamic stability in a ML framework has been recently the focus of very interesting studies [6-10]. However, for large data sets ML computations become unfeasible without the assumption of independent sites. It has been also shown that modeling site-specific residue frequencies significantly improves methods for evolutionary reconstructions [11].

The simplest models of molecular evolution are based only on the mutation process and do not attempt to evaluate its effect on fitness. Kimura's neutral model [12,13] uses a binary fitness function to represent purifying selection. Protein sequences are considered either unviable or equivalent (neutral). In this model a fraction x of the mutations gives rise to neutral mutants and the remaining fraction $1 - x$ is eliminated by purifying selection.

In this paper, we use a binary fitness function based on the evaluation of the thermodynamic stability of the native state, in the same spirit of models first introduced in the context of RNA evolution [14-16] and subsequently extended to protein evolution [6,17-28]. The model that we study is a neutral evolution model with explicit stability requirements, the Structurally Constrained Neutral (SCN) model of protein evolution [29-32]. We study the model in the limit in which the product of the population size M times the mutation rate μ is small, which means that the population is very narrowly distributed in genotype space. This limit is appropriate for animal populations. In the regime of frequent mutation, the evolutionary dynamics is different from the rare mutation regime considered here in that there is a trend towards increased mutational robustness for increasing $M\mu$ [33-35].

Stability requirements induce correlations between the sites of the macromolecule, as the free energy is a property of the system as a whole. In the present paper, we present an evolutionary model in which the sites evolve independently of each other and, in addition, are subject to structural stability. We show that the selection rules can be chosen in such a way to reproduce the evolutionary process simulated through the SCN model of protein evolution, in which the stability is explicitly evaluated. In order to eliminate the correlations induced by stability requirements there is, however, a price to pay: In the mean-field model, the effective selection depends on the mutation process. As a result, sites evolve independently but selection and mutation become interrelated, whereas in real evolution sites evolve in a correlated fashion and mutation and selection are independent processes.

We simulated the SCN model with different mutation schemes. The results of these simulations agree very well with the mean-field model, and show that both mutation and selection have influence on protein folding thermodynamics.

Furthermore, we applied the effective evolutionary model to a non-redundant set of globular protein structures contained in the Protein Data Bank (PDB), using several mutation schemes of increasing complexity. The mutation scheme that best reproduces the observed amino acid distributions at all sites is one that takes into account the increase of the mutation rate at CpG dinucleotides. The site-specific amino acid distributions obtained through this model reproduce quite well observed amino acid distributions.

The SCN model

The SCN model evaluates the "fitness" associated with a protein sequence through a model of protein folding based on an effective free energy function (see Eq. (13) in Methods). We adopt two measures of protein folding stability: (1) with respect to the unfolded state (unfolding stability), estimated through the effective native energy, and (2) with respect to misfolded states (misfolding stabililty), estimated through the normalized energy gap (see Methods). We use a binary fitness function that assigns fitness one if both stabilities are above predefined thresholds and zero otherwise. Thus our model is neutral, since there is no fitness difference between viable proteins.

The free energy calculations introduce correlations between the sites of the protein. As a result of these dependencies, the fraction of neutral mutations x is no longer constant, as in Kimura's model, but fluctuates broadly from one sequence to another. This implies that the distribution of the number of substitutions is broader than the Poissonian distribution arising from the standard neutral model by Kimura, i.e. the process of protein evolution is overdispersed [30-32], in better agreement with empirical observations [36,37].

Results

Optimal sequence for a protein structure

We evaluate the stability of folded states through the contact free energy function

$$E(\mathbf{C},\mathbf{A}) = \sum_{ij} C_{ij} U(A_i, A_j), \qquad (1)$$

where **C** represents the binary contact matrix derived from the protein structure, **A** represents the protein sequence and $U(a, b)$ is the effective contact interaction strength between amino acids a and b, belonging to the set of twenty standard amino acids. This model of protein stability, despite its simplicity, captures several relevant features of protein folding, in particular those related with hydrophobicity. In particular, it allows to estimate the stability against unfolding and misfolding for sequence-structure pairs, in such a way that the native structure is more stable than all alternative structures for almost all protein chains in the PDB [38]. Difference in stability between homologous proteins can be related to evolutionary and ecological variables [39], and estimates of unfolding free energy are correlated with experimental measures (UB, unpublished result).

A further approximation of this model allows to design analytically the optimally stable sequence for a given fold. This approximation consists in truncating the spectral decomposition of the contact interaction matrix at the first spectral component, namely $U(a, b) \approx -h(a)h(b)$, where $h(a)$ indicates the component of the main eigenvector of $U(a, b)$ corresponding to amino acid a, which we call the *interactivity* of amino acid a. It is well known that the main eigenvector of contact interaction matrices is related to hydrophobicity [40,41]. Consistently, the interactivity scale is strongly correlated with empirical hydropathy scales as for instance the octanol scale derived by Fauchere and Pliska [42].

We define the hydrophobicity profile (HP) of a protein sequence associating at each site the interactivity value $h(A_i)$ corresponding to its amino acid. Using only the first spectral component, the interaction matrix can be reconstructed to good accuracy (the correlation coefficient between $U(a, b)$ and $-h(a)h(b)$ is $r = 0.80$). Under this approximation, the energy can be written as a quadratic form of the hydrophobicity profile

$$E(\mathbf{C},\mathbf{A}) \approx -\sum_{ij} C_{ij} h(A_i) h(A_j). \qquad (2)$$

In the following, we will use the complete contact interaction matrix $U(a, b)$, Eq. (1), for simulations of the SCN model, and the hydrophobic energy Eq. (2) for analytic computations. We further neglect the discretization of the HP into twenty values and consider the h_i as real valued independent variables. This setting allows to solve analytically the sequence design problem of determining the optimally stable sequence for a target structure. The HP with minimal energy for a given contact matrix **C** and for fixed mean square[1] $\sum_i h_i^2 / N \equiv \langle h^2 \rangle$ is parallel to the principal eigenvector (PE) of the contact matrix, whose components are denoted here by c_i, i.e. $h_i^{\mathrm{opt}} = \sqrt{N \langle h^2 \rangle c_i}$.

However, selection in the SCN model is applied not only to the native energy, which estimates stability with respect to unfolding, but also to the stability against misfolded compact structures. The relevance of molecular diseases related to misfolding and aggregation [43], the importance for cell physiology and evolution of molecular chaperones preventing misfolding [44-46], and comparative analysis of thermodynamic properties of homologous proteins [39] all indicate that selection for stability against misfolding is an important selective force. In our model, this is achieved imposing a minimal allowed value for the normalized energy gap (see Methods).

Therefore, we set out to minimize the native energy with a large value of the normalized energy gap. The computation is reported in the Methods section. The implicit analytic solution, Eq. (17), receives its main contribution from the PE. This contribution is overwhelming in case of structures without internal modularity. Therefore, we further approximate the optimal HP as the sum of the PE profile plus a term which is constant at all sites. This is equivalent to stating that the correlation coefficient between PE and HP is one, and yields

$$h_i^{\mathrm{opt}} \approx \sqrt{\frac{\langle h^2 \rangle - \langle h^2 \rangle}{\left(\langle c^2 \rangle - \langle c \rangle^2 \right)}} (c_i - \langle c \rangle) + \langle h \rangle. \qquad (3)$$

Before ending the section, we list the approximations involved in the above equation and its limits of validity. For comparison with simulation results, see Fig. 1 below.

1. We neglected the lower eigenvectors in the spectral expansion of the contact interaction matrix $U(a, b)$. Simulations of the SCN model with equiprobable mutations show that this approximation is rather good, since one can get correlation coefficients larger than 0.95 between the simulated optimal HP and the PE of single-domain proteins, which would not be the case if the other components of the contact interaction matrix would pose significant constraints.

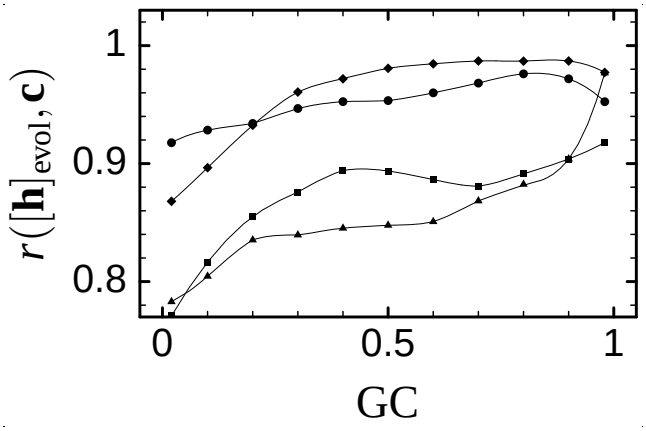

Figure 1
Correlation coefficient between the average HP and the PE for SCN simulations with various mutation models yielding different GC biases, for three single-domain proteins, lysozyme (PDB code 31zt, circles), phosphocarrier protein Hpr (PDB code 1opd, diamonds), and myoglobin (PDB code 1a6g, squares), and for the small two-domains protein ATP synthase ε unit (ATPE, PDB code 1aqt, triangles).

2. We neglected the discretization of the hydrophobicities in twenty values corresponding to the amino acids. This approximation is usually good for values of $\langle h \rangle$ and $\langle h^2 \rangle$ as observed in real proteins, which are much larger than the minimum and much smaller than the maximum of the twenty hydrophobicity values, but it is violated for extreme mutation bias.

3. We used the REM approximation for the normalized energy gap [47], which has been calculated through threading in the simulations. This is not a problem, since there is very good correlation between the REM estimate and the threading result at fixed sequence length. Moreover, since the threading calculation overestimates the normalized energy gap for long proteins, the REM approximation may even yield a more appropriate estimate [47].

4. We assumed that the lower eigenvectors contribute negligibly to the optimal hydrophobicity. This approximation is violated for multi-domain proteins, where other eigenvectors have non-negligible contributions. For single-domain proteins the corrections are usually small in the cases that we simulated.

Evolutionary average of the hydrophobicity profile

In the SCN model, following Kimura's neutral model, all sequences having stability properties above some predetermined threshold are selectively equivalent. Therefore, the optimal sequence is very unlikely to be realized during evolution. However, all viable sequences must be sufficiently stable, which implies that they must have large correlation coefficient with the optimal HP. Thus, in SCN simulations protein sequences are expected to move around the optimal sequence, so that the evolutionary average of the HP almost coincides with the optimal HP [48]. This condition can be written formally as

$$[h(A_i)]_{evol} \equiv \sum_{\{a\}} \pi_i(a) h(a) = h_i^{opt}. \qquad (4)$$

The evolutionary average of the HP is indicated as $[\mathbf{h}]_{evol}$ and it is expressed as a sum over all amino acids $\{a\}$ of the site specific amino acid distribution at site i resulting from the evolutionary process, $\pi_i(a)$. The average sequence so defined is closely related to the prototype sequence defined by Bornberg-Bauer [21,22], which is maximally stable both thermodynamically and against mutations.

Combining Eq. (3) and (4), we obtain an analytic prediction of the average value of the site-specific amino acid distributions,

$$\sum_{\{a\}} \pi_i(a) h(a) = \sqrt{\frac{\langle h^2 \rangle - \langle h^2 \rangle}{\left(\langle c^2 \rangle - \langle c \rangle^2 \right)}} (c_i - \langle c \rangle) + \langle h \rangle. \qquad (5)$$

The quantities $\langle h \rangle$ and $\langle h^2 \rangle$ are the only parameters that we need in order to compute the average HP at all sites of the protein as a function of the PE. In the following, these parameters will be measured in the simulations.

If the correlation coefficient between the PE and the average HP would be one, then the prediction Eq. (5) would be exact. We verified that the correlation coefficient is large (see below) and the analytic prediction yields a good fit of the average HP. The slope of the linear relationship between average HP and PE has a typical relative root mean square error of 9% with respect to the predicted value. The intercept has a typical relative error of 34%, but the mean absolute error is very small, 0.02, which is less than 5% of the mean HP. In the following, we further discuss the quality of the prediction reporting the correlation coefficient, which gives a strong indication of the relative error.

We plot in Fig. 1 the correlation coefficient between the average HP and the PE for SCN simulations with various mutation models characterized by different bias towards the nucleotides C+G, for lysozyme (PDB code 31zt), phosphocarrier protein Hpr (PDB code 1opd) and myoglobin (PDB code 1a6g), which are three single-domain proteins, and ATP synthase ε subunit (ATPE, PDB code 1aqt), which, despite its small size of 135 residues, has two domains, a small beta barrel and a two-helix bundle.

Except for extreme mutational bias the correlation coefficient between average HP and PE is always larger than 0.8. Lysozyme and Hpr have a similar behavior, characterized by high correlation coefficient between the average HP and the PE. The decrease of the correlation at extreme mutation bias is mainly due to the finite values of the minimal and maximal hydrophobicity. The same effect is present for myoglobin and ATPE, but for these proteins the correlation is lower because the contribution of other eigenvectors to the optimal HP is not negligible. Nevertheless, the correlation $r([\mathbf{h}]_{evol}, \mathbf{c})$ is close to one for all four proteins for almost all the mutation models that we simulated. We found that a good predictor of the relevance of other eigenvectors is the quantity η_2 defined in Eq. (20). This quantity is small (≈ 0.04) for the first two proteins and larger (≈ 0.4) for the other two and correlates strongly (r = -0.935) with $r([\mathbf{h}]_{evol}, \mathbf{c})$ at zero GC bias (GC = 0.5)[2]. Since simulations of ATPE represents the worst case for our analytic theory, we will focus on them as an example in the rest of the paper.

Effective selection model based on the principal eigenvector of the contact matrix

Our aim here is to exploit the results on the average HP in order to define an effective selection model that reproduces the SCN results based on explicit protein thermodynamics.

Let us first consider a mutation model where mutations from any amino acid to any other one are equiprobable. This is the mutation model that we simulated in our previous studies [29-31]. In this case, we assume that Eq. (5) is the only condition acting on the stationary amino acid distributions $\pi_i(a)$. This assumption is translated into the requirement that each $\pi_i(a)$ is the distribution of maximum entropy with mean values given by Eq. (5). As it is well known, this is an exponential, or Boltzmann, distribution of the form [49]

$$\pi_i(a) \propto \exp[-\beta_i h(a)], \quad (6)$$

The site-specific Boltzmann parameters β_i can be computed analytically imposing that the average values $\sum_a \pi_i(a)h(a)$ are given by Eq. (5), which depends only on the PE components c_i and on the two parameters $\langle h \rangle$ and $\langle h^2 \rangle$.

Details on the calculation of β_i are given in the Methods section. β_i takes both positive and negative values. Through the theorem of the implicit function it is easy to see that β_i is a decreasing function of c_i. This has a very simple interpretation: Positions with large c_i are buried in the core of the protein, so they tend to have larger mean hydrophobicity, Eq. (5), and with higher probability they are occupied by hydrophobic amino acids, thus having more negative β_i. Positions with small c_i are exposed, tend to be hydrophilic, and have large and positive β_i. This result is not surprising qualitatively, but it is remarkable that it allows to compute quantitatively the probability to observe a hydrophobic amino acid as a function of a structural indicator, the component c_i of the principal eigenvector of the contact matrix, and on the two parameters $\langle h \rangle$ and $\langle h^2 \rangle$.

Modelling amino acid distributions using Boltzmann distributions of physico-chemical properties had been previously proposed by Goldstein and coworkers [50,51] and by Shaknovich and coworkers [27,28]. Our approach differs from previous ones in the sense that we compute the Boltzmann parameter explicitly as a function of a structural indicator, the principal eigenvector of the contact matrix.

Several structural properties of proteins are found to follow Boltzmann distributions as well, i.e. the frequency of a structural motif is exponentially depending on its energy. For a review and a theoretical explanation based on the Random Energy Model, see [52].

The above result allows us to define a stochastic evolution process with independent sites that reproduces the site-specific distributions obtained through the SCN model where sites are interacting. The simplest site-specific transition matrices having Eq. (6) as its equilibrium distribution and satisfying detailed balance have the form

$$P_{sel}^{(i)}(a, b) = \min\{1, \exp[-\beta_i [h(b) - h(a)]]\}. \quad (7)$$

It is possible to extend this selection model to take into account functional conservation, which is not considered in this paper.

Mean-field model with selection and mutation

The condition on the site-specific average hydrophobicity, Eq. (5), is independent of the mutation model in our analytic approximation. Simulations show a weak dependence on the mutation parameters, see Fig. 1. For strong mutation bias, at very low GC, this dependence is a consequence of the fact that the average hydrophobicity approaches its maximum value. For intermediate bias, this dependence can be rationalized noting that the relevance of other eigenvectors depends on the parameter $1 - \langle h \rangle / (\sqrt{N\langle h^2 \rangle} \langle c \rangle)$, which varies with the mutation bias. In any case, Eq. (5) is a good approximation to SCN simulations, except for extreme mutation bias.

The stochastic selection process described by Eq. (7) imposes that the average hydrophobicity follows Eq. (5). This effective selection process reproduces the site-specific amino acid distributions obtained through simulations of the SCN model with equiprobable mutation. Here we simulate the SCN model with a more realistic mutation process that takes into account the genetic code and represent mutations nucleotide level. The stochastic process that corresponds to this modified SCN model is the combination of two processes [53]: (1) A mutation process identical to the one simulated in the SCN model; (2) The selection process described by Eq. (7), which imposes that the site-specific average HP is perfectly correlated with the PE.

For implementing the nucleotide mutation model, we define the state of each site i as a codon $\mathbf{n} = \{n_1 n_2 n_3\}$. The substitution process is then decomposed into mutation and selection processes according to

$$P^{(i)}(\mathbf{n},\mathbf{n}') \equiv P_{\mu}^{\mathrm{COD}}(\mathbf{n},\mathbf{n}')P_{\mathrm{sel}}^{(i)}(\mathcal{A}[\mathbf{n}],\mathcal{A}[\mathbf{n}']) \quad (\mathbf{n} \neq \mathbf{n}'), \tag{8}$$

where $P_{\mu}^{\mathrm{COD}}(\mathbf{n}, \mathbf{n}')$ is the codon mutation matrix arising from the mutation process at the nucleotide level (see Methods), the selection process is represented in Eq. (7), and $\mathcal{A}[\mathbf{n}]$ represents the amino acid coded by the codon $\mathbf{n}$. The diagonal elements are defined through the normalization condition

$$P^{(i)}(\mathbf{n},\mathbf{n}) = 1 - \sum_{\mathbf{n}'\neq\mathbf{n}} P_{\mu}^{\mathrm{COD}}(\mathbf{n},\mathbf{n}')P_{\mathrm{sel}}^{(i)}(\mathcal{A}[\mathbf{n}],\mathcal{A}[\mathbf{n}']). \tag{9}$$

We first assume that the mutation process at the DNA level satisfies detailed balance, also called reversibility in the molecular evolution literature. This means that the stationary nucleotide frequencies $f(n)$ satisfy the equation $f(n_1)\, P_{\mu}^{\mathrm{nuc}}(n_1, n_2) = f(n_2)\, P_{\mu}^{\mathrm{nuc}}(n_2, n_1)$, where $P_{\mu}^{\mathrm{nuc}}(n_1, n_2)$ is the mutation matrix at the nucleotide level.

Under this hypothesis, the stationary distribution for the full substitution process can be decomposed as the product of the amino acid frequency $w_{\mathrm{AA}}(a)$ expected from the mutation process without selection, which is the same for all sites, times the site-specific distributions due to the selection process, Eq. (6),

$$\pi_i(a) \propto w_{\mathrm{AA}}(a)\, \exp[-\beta_i\, h(a)], \tag{10}$$

$$w_{\mathrm{AA}}(a) = \sum_{\mathbf{n}} \delta(a,\mathcal{A}[\mathbf{n}])w_{\mathrm{COD}}(\mathbf{n}). \tag{11}$$

The factor $w_{\mathrm{AA}}(a)$ is obtained as the sum of the expected frequencies of its codons under mutation alone, $w_{\mathrm{COD}}(\mathbf{n}) = f(n_1)\, f(n_2)\, f(n_3)$, with n_1, n_2, n_3 the three nucleotides composing the codon $\mathbf{n}$, and $f(n)$ is the stationary frequency of nucleotide n under mutation alone. It is easy to see that the distribution Eq. (10) satisfies detailed balance with respect to the full substitution process, Eq. (8).

Eq. (6) is a special case of Eq. (10), with a constant mutation factor $w_{\mathrm{AA}}(a) \equiv 1$ for all amino acids, consistently with the assumption that all mutations are equiprobable. Combining Eq. (10) with Eq. (5), we obtain

$$\frac{\sum_{\{a\}} h(a) w_{\mathrm{AA}}(a) \exp[-\beta_i h(a)]}{\sum_{\{a\}} w_{\mathrm{AA}}(a) \exp[-\beta_i h(a)]} = \sqrt{\frac{\langle h^2\rangle - \langle h\rangle^2}{\left(\langle c^2\rangle - \langle c\rangle^2\right)/\langle c\rangle^2}}\left(\frac{c_i}{\langle c\rangle} - 1\right) + \langle h\rangle. \tag{12}$$

This equation is the central result of this work. It allows to compute the Boltzmann exponents β_i, and from them the site-specific amino acid distributions as a function of the normalized PE of a protein structure, $c_i/\langle c\rangle$, which depend little on the protein length, of two hydrophobicity parameters, $\langle h\rangle$ and $\langle h^2\rangle$, and of the stationary frequencies $w_{\mathrm{AA}}(a)$ of the mutation model, which must fulfill detailed balance. From this equation one sees that the Boltzmann exponents β_i, which define the selection process, depend on the mutation factors $w_{\mathrm{AA}}(a)$. This contrasts with the fact that in the SCN model mutation and selection are two independent processes. The entanglement between selection and mutation is a consequence of the mean-field approach: We reduce an evolutionary process where sites are interrelated to a process where sites evolve independently, but under the global constraint given by Eq. (5). Because of this global constraint, the evolutionary process becomes dependent on the average properties of the amino acid chain, which in turn depend on the mutation process. Hence, at the mean-field level, mutation and selection become interrelated.

SCN model with genetic code

To test the mean-field model, the SCN model was simulated using several mutation schemes. In a first set of simulations, we used mutation matrices that fulfil detailed balance and depend on the stationary frequencies $f(n)$, $n \in \{\mathrm{A, T, C, G}\}$, and on the transition-transversion ratio k (see Methods). We further imposed the conditions $f(\mathrm{C}) = f(\mathrm{G})$ and $f(\mathrm{A}) = f(\mathrm{T})$, called type 2 parity rule [54], so that there are only two parameters in the mutation matrix.

Fig. 2 shows the site-specific amino acid distributions obtained through SCN simulations for the protein ATPE (PDB code 1aqt), divided by the frequencies expected under mutation alone, $w_{\mathrm{AA}}(a)$. The distributions plotted in Fig. 2(a) and 2(b) refer to a site with small PE component, $c_i/\langle c\rangle = 0.43$, favoring amino acids with low hydrophobicity. We show data for two different mutational

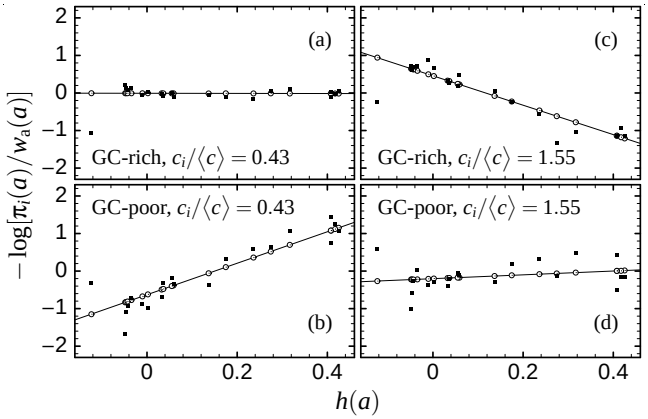

Figure 2
Comparison of the site-specific amino acid distribution $\pi_i(a)$ obtained from simulations of the SCN model for ATPE (PDB code 1aqt, full symbols) and from the mean-field model (lines and open symbols) at site i = 128 with $c_i/\langle c\rangle$ = 0.43 [(a) and (b)] and at site i = 82 with $c_i/\langle c\rangle$ = 1.55 [(c) and (d)]. The upper panels (a) and (c) show the case of high GC mutational bias, whereas the lower ones (b) and (d) show low GC mutational bias.

biases, one favoring GC-rich codons (f(C) + f(G) = 0.8), which tend to be more hydrophylic, and one favoring GC-poor codons (f(C) + f(G) = 0.2), which tend to be more hydrophobic. The exponent β_i changes with the mutational bias: For large GC bias the codons coding for hydrophilic amino acids are favored at the mutation level, and the β_i is almost zero, meaning that almost no purifying selection is needed at this site. The contrary holds when the mutation bias favors GC-poor codons: In this case β_i is large, and the site experiences strong purifying selection.

The opposite situation is observed in Fig. 2(c) and 2(d), obtained for a site with $c_i/\langle c\rangle$ = 1.55. In this case amino acids with high hydrophobicity are preferred, and β_i is almost zero when the mutational bias favors GC-poor codons, whereas it is negative and large when GC-rich codons are favored at the mutational level. In both cases, the mean-field model (straight lines and open symbols) fits very well the results of the SCN simulations (full symbols). The average over all sites of the correlation coefficients between predicted and observed amino acid distributions, for all mutational biases simulated, lies in the range $\langle r\rangle$ = 0.83 and $\langle r\rangle$ = 0.92 and increases as a function of the number of sequences examined (we simulated about 10^6 sequences for each mutational bias).

In all cases studied, the stationary amino acid distributions only depended on the stationary nucleotide frequencies $f(n)$ and did not depend on the transition-transversion ratio, in agreement with Eq. (10). We also simulated a mutation process that does not fulfil detailed balance. In this case, Eq. (10) does not represent the stationary amino acid distribution of the process, which must be explicitly computed from the full transition matrix. Also in this case, Eq. (10), calculated with the stationary frequencies $f(n)$ of the pure mutation process, is a very good approximation of the stationary distribution for high GC bias, but not for low GC, because of the large frequency of stop codons, which are strongly negatively selected (data not shown).

As a result of the change of the selection parameters β_i with the mutation bias, the probability that a mutation is accepted depends on the mutation bias. This probability is also influenced by the transition-tranversion ratio k. The probability that a mutation is accepted, obtained from SCN simulations, and analytically computed from the effective stochastic process Eq. (8) (see Methods), is plotted in Fig. 3. The analytical calculation is in very good agreement with simulation results. Interestingly, the acceptance probability has a maximum as a function of the GC frequency, meaning that there is an optimal mutation bias that maximizes the rate of accepted mutations. This maximum is achieved at f(C) + f(G) > 0.5, since stop codons, which are the most deleterious mutations, are rich in AT. The acceptance probability increases with the transition to transversion ratio k, which favors more conservative mutations.

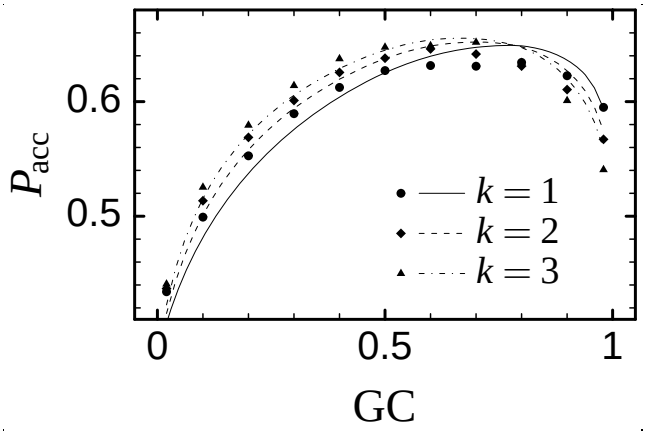

Figure 3
Acceptance probability for a mutation P_{acc}, calculated in SCN simulations for ATPE (PDB code 1aqt, symbols) and in the mean-field model (lines) for three different values of the transition to transversion ratio k as a function of the GC content, f(C) + f(G). The mutation model is such that P(C) = P(G) and P(T) = P(A), assuming type 2 parity rule [54].

Influence of the mutation bias on protein folding thermodynamics

In the SCN model, mutation and selection parameters influence the properties of protein folding thermodynamics arising from simulated evolution.

In both the SCN and the mean-field model, the nucleotide content at different codon positions responds differently to the mutation bias. The second position is the most difficult to mutate, whereas the third one is almost completely neutral, with $GC_3 \approx GC_{mut}$, since most transitions at third codon position are synonymous and they are always accepted in the SCN model. Thymine content at second codon position is least dependent on the mutation bias. Almost all codons containing T in the second position code for hydrophobic amino acids, so that the T content at second position is strongly correlated with the hydrophobicity of the coded protein and it is strongly constrained by selection for thermodynamic stability. A mutation bias towards T at the nucleotide level corresponds to a mutation bias towards more hydrophobic amino acids.

Also the selection thresholds on the native energy and the normalized energy gap influence the hydrophobicity. In fact, the more hydrophobic a protein is, the more stable it is against unfolding (it has lower effective energy), and the less stable it is against misfolding (it has smaller normalized energy gap). Stronger selection for a large normalized energy gap has thus the effect to reduce the hydrophobicity, and stronger selection for a low energy has the effect to increase it. The results presented here are obtained varying the mutation bias at constant selection thresholds, so that the differences between different simulations are due to mutation alone. We choose the selection thresholds equal to 98% of the values of the stability parameters in the PDB sequence. In this way, the PDB sequence is always selected, and we get stringent stability criteria that are as uniform as possible between different proteins.

Therefore, mutation and selection parameters modify the trade-off between the two kinds of stability and influence the mean and mean square hydrophobicity, $\langle h \rangle$ and $\langle h^2 \rangle$. These quantities are used as parameters to calculate the site-specific distributions in the mean-field model. Unfortunately, we could not predict them analytically as a function of the mutation and selection parameters. The parameters used in the mean-field model were therefore derived from the sequences generated through SCN simulations. Nevertheless, using parameters $\langle h \rangle$ and $\langle h^2 \rangle$ that do not depend on the mutation bias in Eq. (5) still produces a good agreement between the mean-field and simulation results.

The dependence of hydrophobicity on the mutation bias has a deep influence on protein folding thermodynamics. For all the proteins that we simulated, the mean square hydrophobicity $\langle h^2 \rangle$ is a decreasing function of the G+C content (or, equivalently, an increasing function of the A+T content), causing the normalized energy gap to increase and the unfolding free energy per residue to decrease for increasing GC. We observed this effect in our SCN simulations. The same qualitative influence of the mutation bias on protein folding thermodynamics was inferred through a statistical analysis of the properties of orthologous proteins in different bacterial species evolving with different mutation biases [39]. Fig. 4 shows the mean hydrophobicity $\langle h \rangle$, the effective energy per residue $-E/N$ and the normalized energy gap α as a function of the GC content. Full symbols and lines are derived from simulations of the SCN model and open symbols are derived from the computational study of bacterial proteomes mentioned above [39]. Both sets of points show a similar trend, but the dispersion of thermodynamic properties is much larger in the bacterial proteomes than in SCN simulations.

There is however a qualitative difference between real and simulated data concerning mean hydrophobicity. In real sequences there is a negative correlation between the GC content of the gene and the mean hydrophobicity ($r = -0.57$, $P < 10^{-4}$), as expected from the fact that T in second position codes for hydrophobic residues. In simulated sequences the correlation between GC content and mean hydrophobicity is negative in the whole GC range, but it is positive in the range of biologically observed GC values for two of the proteins simulated, ATPE and myoglobin. For the other two proteins, lysozyme and Hpr, the correlation is always negative. In contrast, the root mean square hydrophobicity $\langle h^2 \rangle$ is a monotonically decreasing function of the GC for all four proteins (data not shown). Interestingly, the proteins showing a positive correlation between GC content and mean hydrophobicity are those for which other eigenvectors besides the PE are relevant. This suggests that the hydrophobicity has a uneven distribution between the different modules that correspond to the relevant eigenvectors, so that, at decreasing GC mutation pressure, the mean hydrophobicity increases in one module and decreases in the other one responding to selection for stability against misfolding. The net effect is the lowering of the mean hydrophobicity while increasing its within sequence variation. It is possible that the resulting positive correlation between GC and hydrophobicity is an artifact of the SCN model, but it is also possible that it is a property of modular proteins that was not detected in the study of bacterial proteomes, since in that study data from unimodular and from multi-modular proteins were averaged together. This point deserves therefore further investigation.

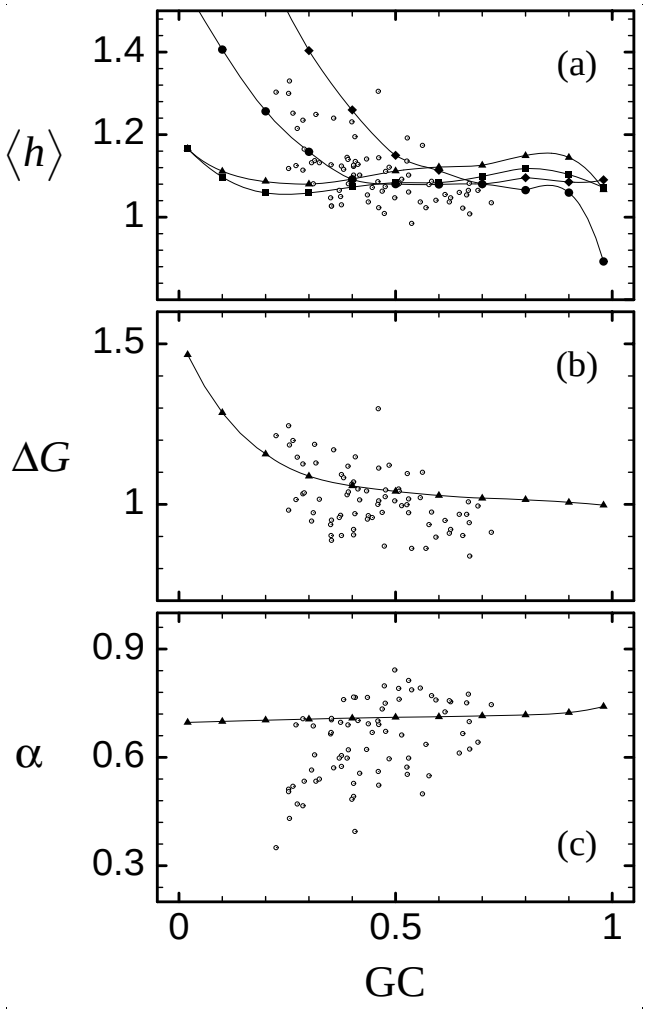

Figure 4
Full symbols and lines indicate average properties of protein folding thermodynamics in SCN simulations, open symbols indicate the same quantities in the proteomes of different bacterial species [39]. The horizontal axis represents the GC mutation bias for SCN simulations and the GC content at third codon position of the bacterial genes, (a) Mean hydrophobicity. SCN results are rescaled by a factor 8.6 and correspond to three single-domain proteins, lysozyme (PDB code 3lzt, circles), phosphocarrier protein Hpr (PDB code 1opd, diamonds), and myoglobin (PDB code 1a6g, squares), and for the small two-domain protein ATP synthase ε unit (ATPE, PDB code 1aqt, triangles), (b) Mean unfolding free energy. SCN results are rescaled by a factor 4.3. Only ATPE is represented, the other proteins being qualitatively equivalent. (c) Mean normalized energy gap. SCN results are rescaled by a factor 1.3. Only ATPE is represented, the other proteins being qualitatively equivalent.

It should also be noted that the normalized energy gap α is much smaller in bacterial proteomes corresponding to genomes with low GC content than expected from SCN simulations. As discussed in Ref. [39], these proteomes with very low GC content belong to obligatory intracellular bacteria, whose effective population size is severely reduced by the bottlenecks that they experience in their intracellular lifestyle. Because of their reduced populations, natural selection is expected to be less effective in eliminating deleterious mutations, causing their phenotypic properties, such as the folding stability of their proteins, to be less stable than for large free living populations [55-57].

Fig. 4 also suggests that natural selection acts on different protein properties for different mutation bias. When the mutation bias favors GC, the normalized energy gap tend to be higher than the threshold and the unfolding free energy tend to be small. In this case, most of the mutations that are selected against are eliminated because they yield proteins too unstable against unfolding. On the contrary, when the mutation pressure favors AT, most of the mutations that are eliminated yield proteins that are unstable against misfolding.

Last, we note that, despite simulated and observed properties have a qualitatively similar response to the mutation bias, from a quantitative point of view simulated quantities depend much more strongly on the mutation parameters. In particular, the GC content at first and second codon position, which influences the nature of the coded amino acids, depends on the mutation bias much more strongly in simulated genes than in real genes (see the Discussion).

Site-specific amino acid distributions in the PDB

The predicted optimal hydrophobicity vector depends on the two parameters $\langle h \rangle$ and $\langle h^2 \rangle$, in particular through the ratio $\langle h \rangle / \sqrt{\langle h^2 \rangle}$, in such a way that the optimal HP is almost parallel to the PE when $\langle h \rangle / \sqrt{\langle h^2 \rangle}$ is almost equal to $\sqrt{N}\langle c \rangle$ (see Methods). Before applying our model to real proteins, we measured these quantities in a non-redundant subset of the Protein Data Bank (PDB) [58], containing both single-domain and multi-domain proteins, filtered to select only globular proteins.

Protein sequences in this set have $\langle h \rangle$ and $\langle h^2 \rangle$ contained in a narrow range, with standard deviation equal to 1/10 of the average value or smaller. Correspondingly, $\tau = \langle h \rangle / \sqrt{\langle h^2 \rangle}$ lies in a narrow range between 0.4 and 0.65 (mean value 0.56), and $\tau / \sqrt{N}\langle c \rangle$ lies in a range between 0.5

and 1.1 (mean value 0.77). The largest values belong to multi-domain proteins, whose $\sqrt{N}$ <*c*> is much smaller than for single-domain proteins. The values of $\tau/\sqrt{w_1}$ are close to one, thus supporting our approximation to neglect eigenvectors other than the PE in the computation of the average HP. The distributions are plotted in Fig. 5, which also shows the distribution of $\sqrt{w_1} = \sqrt{N}$ <*c*>.

In the following, we will consider only single-domain globular proteins (see Methods) and assume that their optimal HP is well approximated by Eq. (5), with the same parameters <*h*> and <h^2> for all proteins. This is justified by the narrow distribution of these quantities. In this way, we are able to predict the average hydrophobicity for structurally equivalent positions, having the same value of $c_i/$<*c*>, in all structures in the PDB, using only two parameters.

We sampled amino acid distributions at site classes characterized by the same value (within a narrow range) of $c_i/$<*c*>). These observed site-specific distributions were then compared with the distributions arising from the mean-field model with different mutation schemes.

The agreement between predicted and observed distributions was measured through their mean correlation coefficient <*r*>. The parameters of the mutation models were fitted optimizing this quantity (see Methods). Note that only the mutation parameters were fitted, whereas the selection parameters β_i were calculated from the mutation model and the site-specific mean hydrophobicities given by Eq. (5), which only involves the two parameters *A* and *B*. These were calculated from the mean and the standard deviation of the hydrophobicity in the whole set of globular proteins, without any fitting procedure.

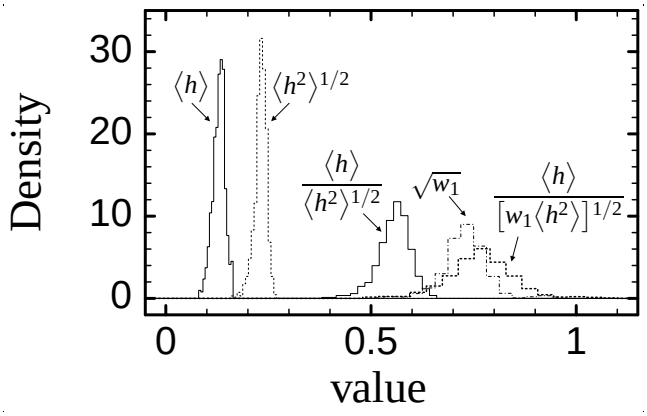

Figure 5

Distributions of hydrophobicity related quantities from a non-redundant subset of the PDB: Mean hydrophobicity; Root mean square hydrophobicity; Ratio between mean and root mean square hydrophobicity, τ, ratio between τ = <*h*>/$\sqrt{\langle h^2\rangle}$ and $\sqrt{W_1} = \sqrt{N}$ <*c*>. All these quantities are narrowly distributed, with standard deviations of the order of less than 1/10 of the average value.

We performed the computations described here using several hydrophobicity scales, listed in Methods. For all mutation models considered, the best fit was always obtained with the IH hydrophobicity scale [48], closely followed by the CH scale [48] and the buriability scale [59]. All other scales gave considerably worse results. In the following we refer to the IH scale when not otherwise stated.

We considered the following five mutation and selection schemes:

(1) No selection ($\beta_i \equiv 0$), nucleotide mutation matrices satisfying detailed balance with the equilibrium frequencies as free parameters. This scheme is labeled as 'opt. freq. at $\beta_i \equiv 0$'. Because of the normalization condition, there are only three free parameters. The optimal equilibrium frequencies are 0.265, 0.327, 0.175, and 0.233 (T, A, C, and G), and yield <*r*> = 0.56.

(2) Selection and uniform mutation probabilities at the amino acid level $P^{(i)}(a, b) \equiv 1$ without the genetic code being taken into account, i.e. $w_{AA}(a) \equiv 1$. This scheme is labeled as 'constant'. For this case we obtained <*r*> = 0.70 without any free parameter. It appears therefore that properly considering the selection process in the mean-field model gives better results than taking into account the genetic code but disregarding the amino acid properties (hydrophobicity), as in scheme (1).

(3) Independent and identical nucleotide mutation matrices satisfying detailed balance with equal equilibrium frequencies for all nucleotides, with selection from the mean-field model. This scheme is labeled as '#codons', since it holds $w_{AA}(a)$ = number of codons. The nucleotide frequencies are 0.25, 0, 25, 0.25, and 0.25 (T, A, C, and G). In this case, both the genetic code and the selection process are considered, and we obtained <*r*> = 0.80, again without any free parameter.

(4) Independent and identical nucleotide mutation matrices satisfying detailed balance with equilibrium frequencies as free parameters, and selection taken from the mean-field model. This scheme is labeled as 'opt. freq.'. The optimal nucleotide frequencies are 0.243, 0.312,

0.189, and 0.257 (T, A, C, and G) and yield $\langle r \rangle = 0.86$ with three free parameters.

(5) Same scheme as above, with an additional free parameter that expresses the enhancement of the mutation rate at CpG dinucleotides. Notice that this mutation scheme does not fulfil detailed balance, and mutations at different sites in the DNA sequences are not anymore independent. We only considered CpG dinucleotides within the same codon, otherwise the resulting mean-field model would not be anymore independent at different positions along the protein sequence. We label this scheme as 'CpG'. The optimal nucleotide frequencies are 0.193, 0.316, 0.210, and 0.281 (T, A, C, and G), and the enhancement of the mutation rate at CpG is $k_{CpG} = 5.6$. These optimal parameters yield $\langle r \rangle = 0.90$.

Fig. 6 shows one example of observed and predicted site-specific amino acid distributions. For illustration, we chose the class of sites with $c_i/\langle c \rangle \in [0.435,0.545]$. These are sites with small PE component, favoring amino acids with low hydrophobicity. Predictions (empty circles) were derived from the mean-field model with various mutation schemes (1, 2, 4 and 5, in the numeration above) and parameters that optimally fit the observed distributions at all site classes. Observed distributions sampled from protein sequences in the PDB are shown as full circles.

In order to make the plot more illustrative, the amino acid frequencies, both predicted and observed, were divided by the frequencies expected under mutation alone, $w_{AA}(a)$, which depend on the mutation model considered. In this way, for the mean-field models with mutation process satisfying detailed balance (1, 2 and 4), the plot represents in logarithmic scale the selection factor $Z^{-1}\exp[-\beta_i h(a)]$ (Z is a normalization constant). Since the horizontal scale represents the amino acid hydrophobicity $h(a)$, one can directly see from the slope of the plot the Boltzmann factor β_i and notice that it indeed depends on the mutation model considered. The mutation scheme 5 ('CpG', Fig. 6 bottom right) does not obey detailed balance, therefore the mean-field predictions, $\log(\pi^{\text{pred}}_{c_i/\langle c \rangle}(a)/w_{AA}(a))$ (open circles), do not lay on a straight line as a function of $h(a)$. Nevertheless, Eq. (10), represented as a line in the figure, is still a good approximation for the mean-field amino acid distribution.

In Fig. 7 we show the observed frequencies $\pi^{\text{obs}}_{c_i/\langle c \rangle}(a)$ versus the probabilities $\pi^{\text{pred}}_{c_i/\langle c \rangle}(a)$ predicted through the mean-field model with optimal mutation parameters. All amino acid types and all sites are represented and, as in the previous figure, all frequencies are divided by the expected frequencies under mutation alone $w_{AA}(a)$. The four frames refer to mutation schemes 2, 3, 4 and 5. For model 1, which is not shown, the predicted probabilities coincide with the mutation factors (i.e., $\pi^{\text{pred}}_{c_i/\langle c \rangle}(a) \equiv w_{AA}(a)$), so their ratio, represented on the horizontal line, is always one, and all points would lie on a vertical line and would yield no correlation between observed and predicted data. The best fit is obtained for models 4 and 5.

Discussion

Approximation used

Our analytic theory is based on a model of protein folding thermodynamics where the contact energy is approximated through $E(\mathbf{C}, \mathbf{A}) \approx -\sum_{ij} C_{ij} h(A_i) h(A_j)$, and the twenty parameters $h(a)$ can be associated with an effective hydrophobicity. The thermodynamically optimal sequence for this model, for a fixed structure **C** of a single-domain globular protein, is the sequence whose HP $h(A_i)$ has correlation coefficient close to one with the PE of the contact matrix **C**. This result also holds with very good approximation for models with contact interactions, where the contact interaction matrix $U(a, b)$ is well approximated by its main eigenvector $h(a)$. This is the case of most contact interaction matrices used in protein folding studies, as for instance the Miyazawa and Jernigan interaction matrix [60], and it is well known that the main eigenvector represents hydrophobicity [40,41]. For the case of the interaction matrix used in this study [38], the correlation between the matrix elements $U(a, b)$ and $h(a)h(b)$ is larger than 0.80, therefore contributions other than hydrophobicity are not completely negligible. However, for single-domain globular proteins and for mutation models that are not extremely biased, the average hydrophobicity profile observed in SCN simulations with full contact interaction matrix is practically undistinguishable from the optimal HP predicted on the basis of the reduced matrix $h(a)h(b)$, which justifies our theory *a posteriori*. The parameters $h(a)$ are obtained from the main eigenvector of the interaction matrix, therefore they should not be interpreted as hydrophobicity in a strict biochemical sense, since they also take into account other kinds of interactions. For instance, aromatic amino acids have very large $h(a)$, in part due to the strength of the interactions between aromatic rings. This is perhaps why the hydrophobicity scale that we use performs, for the purpose of predicting site-specific amino acid distributions, better than empirical hydrophobicity scales.

There are other kinds of interactions that can not be well approximated in this simple contact scheme, such as elec-

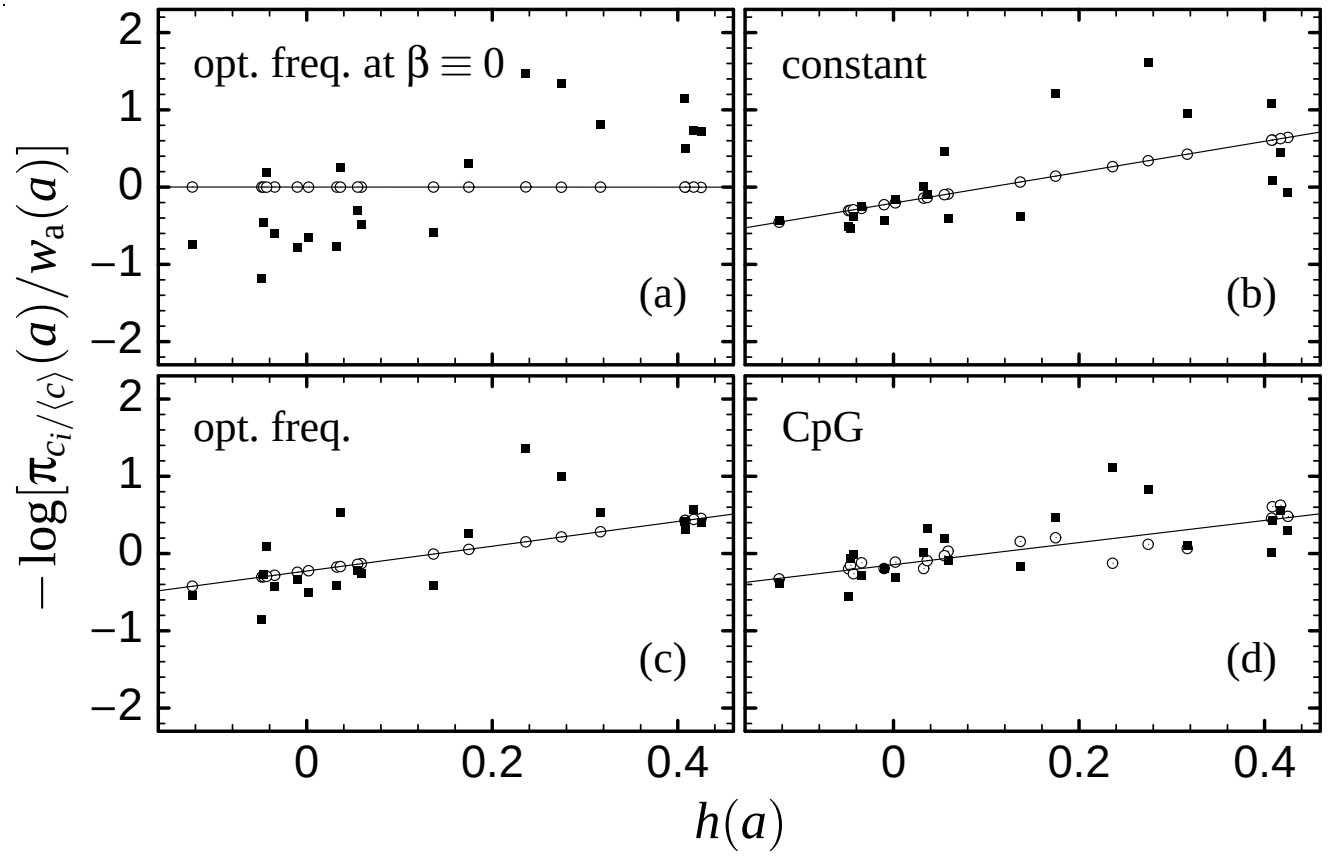

Figure 6

Observed and predicted site-specific amino acid distribution $\pi_{c_{i/<c>}}(a)$, divided by the expected frequencies under mutation alone $w_{AA}(a)$, for (a) the mutation models 1 ('opt. freq. at $\beta \equiv 0$'), (b) mutation model 2 ('constant'), (c) mutation model 4 ('opt. freq.'), and (d) mutation model 5 ('CpG'). For the theoretical models where mutation satisfies detailed balance, $\pi_{c_{i/<c>}}(a)/w_{AA}(a) \propto \exp[-\beta_i h(a)]$, so that the slope of the plot represents β_i at this site class. For illustration, site class with $c_i/\langle c\rangle \in [0.435, 0.545]$ was selected. Full symbols show the observed distributions obtained from sequences in the PDB, whereas the open symbols and the lines display the mean-field model.

trostatic interactions or local propensities for secondary structures, although the latter could be easily incorporated in the present scheme, and we are working in this direction. The good agreement between predicted and observed distributions, however, indicates that the energy function used captures a major component of the forces stabilizing protein folding.

Using the approximate free energy function, Eq. (2), and a continuous approximation for the hydrophobicity values $h(A_i)$, it is possible to determine analytically the sequence with minimal energy subject to the constraint of constant normalized energy gap. The HP of this optimally stable sequence is given by Eq. (17). For single-domain globular proteins, this optimal HP is almost parallel to the main eigenvector of the contact matrix (the PE).

It may be useful to recall some properties of the PE, in order to clarify its interpretation. The PE is the vector c_i which maximizes the quadratic form $Q = \sum_{ij} C_{ij} c_i c_j$ for fixed value of the norm $\sum_i c_i^2$. Therefore, it can be interpreted as an effective connectivity, since positions i with large c_i are in contact with as many as possible positions j with large c_j. Indeed, c_i correlates with the number of contacts formed by site i. Nevertheless, c_i depends not only on the local contacts, but also on the global structure of the pro-

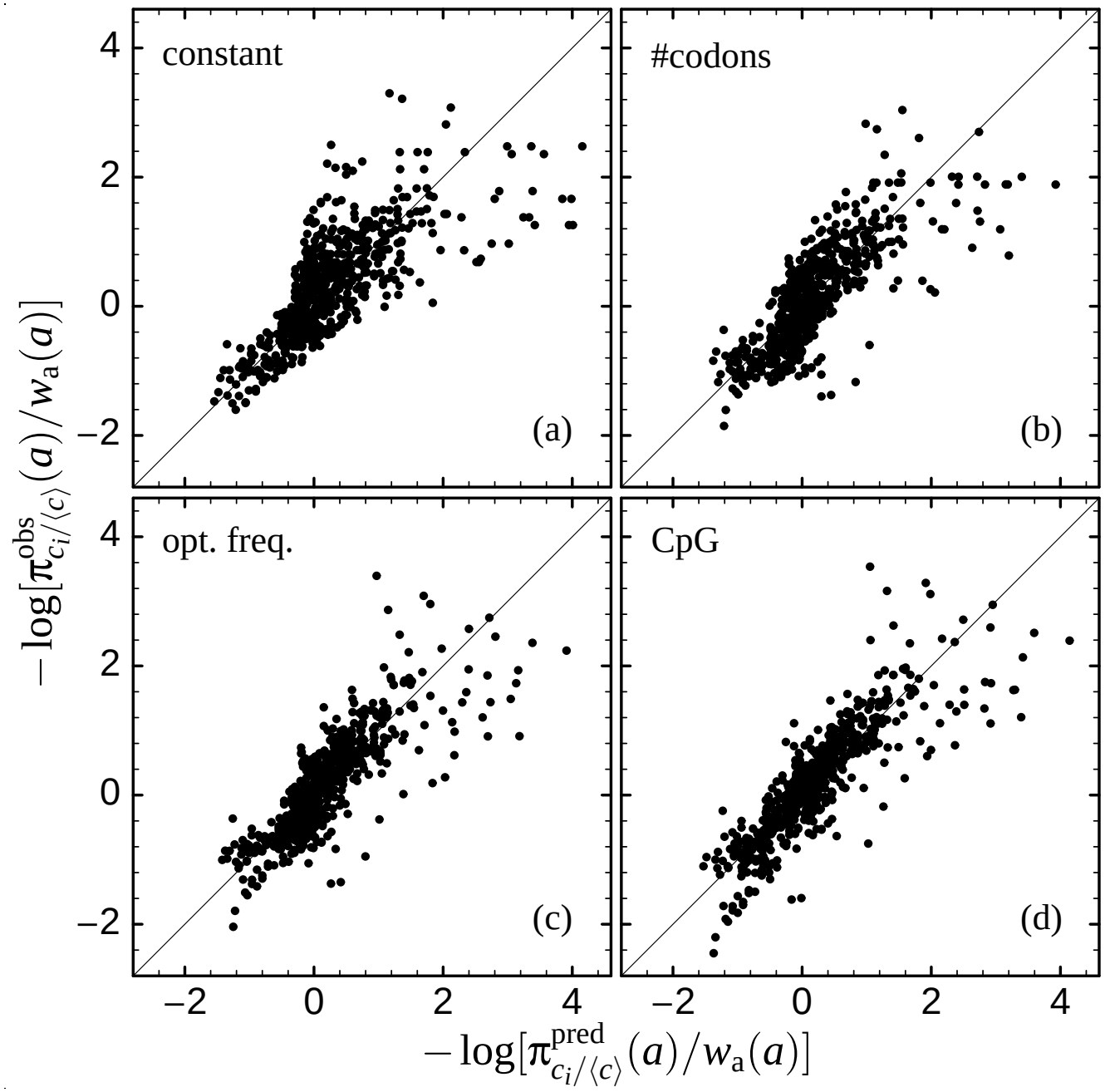

Figure 7

Site-specific amino acid frequencies sampled from the PDB, $\pi^{\mathrm{obs}}_{c_i/\langle c\rangle}(a)$, versus the probabilities $\pi^{\mathrm{pred}}_{c_i/\langle c\rangle}(a)$ predicted through the mean-field model with optimal mutation parameters. All amino acids and all sites are shown. Observed and predicted frequencies are divided by the frequencies expected under mutation alone $w_{\mathrm{AA}}(a)$. The four frames refer to (a) mutation model 2 ('constant'), (b) mutation model 3 ('#codons'), (c) mutation model 4 ('opt. freq.'), and (d) mutation model 5 ('CpG'), respectively.

tein chain, for instance its modularity, and it gives a much richer information than the simple number of contacts. It has been shown that a detailed knowledge of the PE is sufficient to reconstruct the full contact matrix [61], whereas different contact matrices may be associated to the same contact vector [62], whose components are the number of contacts at each site.

Influence of mutation on evolutionary and thermodynamic properties

It was observed several years ago that the nucleotide content at first and second codon positions, which influences the coded amino acid and the amino acid usage, is strongly correlated with the nucleotide content at third codon position, where transitions (A-G and T-C mutations) in most cases do not modify the coded amino acid [54,63]. As a consequence, the amino acid usage in proteins of different bacterial species, which may evolve with different mutation bias, is strongly dependent on the mean nucleotide content of the genome, which is thought to reflect essentially the mutation bias. However, this dependence is not as strong as one would predict from a model based on mutation alone [64], equivalent to our model 1 discussed in the Results section. This deviation from a pure mutation model reflects, at least in part, selection at the amino acid level.

SCN simulations of globular proteins, and the corresponding mean-field models, which take into account both mutation and selection, reproduce qualitatively these results. Our results show that the amino acid composition, and the GC content at first and second codon position, reflecting selection at the amino acid level, depend on the mutation bias and they are strongly correlated with the GC content at third codon position, but this correlation is weaker than one would expect under a mutation model alone.

SCN simulations also reveal the deep influence that the mutation process exerts on protein evolution. Selection for the stability of the native state has to fulfil two partially contrasting requirements: stability against unfolding and stability against compact misfolded conformations [39,47]. The mutation bias influences the balance between these two kinds of stabilities. For fixed selection parameters, mutation processes favoring GC rich codons favor protein sequences that are less hydrophobic, and which are predicted to be more stable against misfolded states but less stable against unfolding with respect to mutation processes favoring AT rich codons.

This trend of decreasing unfolding stability and increasing misfolding stability versus GC content occurs in SCN simulations for all the proteins that we studied, in agreement with a previous statistical analysis of bacterial proteomes [39]. However, for two of the proteins that we studied, the mean hydrophobicity was not a monotonous function of the GC content.

The mutation bias also influences very strongly the fraction of mutations that are eliminated by purifying selection. The optimum acceptance rate is observed for a slight bias towards GC. The nucleotide frequencies that optimize the match between the mean-field model and the observed distributions yield a nearly optimal acceptance rate.

However, in the SCN model the dependency between amino acid usage and mutation bias is significantly stronger than the one observed in the genes of different bacterial species coding for globular proteins. In particular, the GC content at the first and second codon position depends on the GC content at the third codon position more strongly in the SCN model than in bacterial genes. We discuss four possible explanations for this discrepancy, that indicate possible extensions of the model.

First, and most important in our opinion, the selection criterium used in SCN simulations is based on a contact free energy function, whose main contribution comes from the hydrophobicity effect and van der Waals interactions. Selection in protein evolution, on the other hand, also depends on functional constraints, on other stability constraints, as for instance secondary structure, and on constraints arising from protein dynamics. The ability of the mean-field model to reproduce site-specific amino acid distributions in the PDB, where these effects are averaged out, suggests that it captures important features of the selection process, but quite probably other selective forces are relevant as well.

Second, in the SCN simulations presented here we have considered mutation probabilities where we enforced the symmetry between nucleotides related through Watson and Crick pairing, $f(A) = f(T)$ and $f(G) = f(C)$, called type 2 parity rule [54]. The type 2 parity rule holds globally in a double-stranded DNA molecule. However, it is well known that the mutation process in the two DNA strands is different, leading to asymmetric strand composition called GC-skew [65]. Therefore, since the distribution of coding sequences in leading and lagging strands may be biased [65], there is no strict reason why this symmetry should hold for protein coding sequences, and in fact we found that the nucleotide contents under mutation alone that optimally fit the observed amino acid distributions do not follow parity rule 2. The results of the SCN model, such as for instance the C+G content at different codon positions, depend also on the proportion of C with respect to G and A with respect to T, and not only on the

cumulative content of C+G. Therefore the results presented here, obtained imposing the type 2 parity rule, are only indicative of a qualitative trend and can not be compared quantitatively with bacterial genes coding for globular proteins.

As a third possible discrepancy between SCN simulations and biological data, we have calculated stationary amino acid distributions when the substituion process has reached equilibrium. However, in SCN simulations even for short proteins equilibrium is attained after a very long transient phase, of the order of 10^5 substitutions. This number of substitutions is far larger than the estimated number of substitutions that took place since the split of the major domains of life [66]. In a recent study, Jordan *et al*. [66] have proposed that the substitution process in proteins has not yet attained equilibrium, and that one can still find the fingerprint of ancestral amino acid distributions by looking at the present day evolutionary process. Last, we recall that the SCN simulations reported here have been performed in the rare mutation regime where the product of population size times mutation rate is $M\mu$ <<> 1. In this regime, the population is genetically homogeneous. If this hypothesis does not hold, one observes a trend towards increased mutational robustness for increasing $M\mu$ [33-35]. However, this "selection" for robustness, without any explicit selective force, is expected to result in an increased correlation between the HP of selected sequences and the optimal HP. In fact, it has been observed by Bornberg-Bauer in simulations of neutral protein evolution that one can identify a prototype sequence that is maximally stable both thermodynamically and mutationally [21,22]. In our model, the prototype sequence coincides with the sequence with the optimal HP, which is strongly correlated with the PE. Consistently, the mutational robustness increases as the HP of the sequence gets closer to the optimal HP, predicted through the PE, as we have verified in previous simulations [67]. Therefore, we do not expect that selection for mutational robustness modifies qualitatively the results presented here, as far as equilibrium properties are concerned.

Amino acid distributions in the PDB

The mean-field model developed in this paper gives a satisfactory fit of the site-specific amino acid distributions observed in single-domain globular proteins. In applying the model to a representative subset of the PDB, we assumed that: (a) Its selection parameters, $\langle h \rangle$ and $\langle h^2 \rangle$, are roughly the same for all proteins. This hypothesis is reasonable, since both quantity are narrowly distributed, with standard deviations smaller than 1/10 of the average value. (b) The mutation process is the same for all the genes from which PDB proteins are derived. This hypothesis is clearly not valid, since mutation patterns depend on the organism considered and, within the same organism, they depend on the DNA strand and on the distance from the origin of replication in bacterial genomes [65], and they are thought to vary broadly within the large eukaryotic genomes.

Nevertheless, the fact that our model fits well the observed distributions may suggest the existence of a general mutational pattern valid for different organisms. Freeman and coworkers [68] found a strong correlation between excess of coding sequences and excess of purine basis (i.e., frequency of G+A) in bacterial and, later, eukaryotic genomes. This pattern agrees quite well with the frequencies that optimal fit our models to the observed distributions, which are $f(\mathrm{T}) = 0.243$, $f(\mathrm{A}) = 0.312$, $f(\mathrm{C}) = 0.189$, $f(\mathrm{G}) = 0.257$ if the nucleotide frequencies are the only free parameters and detailed balance is assumed, and $f(\mathrm{T}) = 0.193$, $f(\mathrm{A}) = 0.316$, $f(\mathrm{C}) = 0.210$, $f(\mathrm{G}) = 0.281$ if the mutation rate is enhanced at CpG dinucleotides. In both cases, we find that the frequency of A is larger than that of T and the frequency of G is larger than that of C, i.e. the fitted parameters suggest that there is purine excess in protein coding genes.

However, for some bacterial genomes the coding excess is better correlated with the excess of G+T (keto excess) than with the purine excess. This is consistent with the fact that in bacterial chromosomes there are two types of mutational patterns: (1) Predominance of G over C (positive GC skew) and of A over T (positive AT skew) in the leading strand, the opposite in the lagging strand, implying purine excess in coding sequences that are overrepresented in the leading strand; (2) Predominance of G over C (positive GC skew) and of T over A (negative AT skew) in the leading strand, the opposite in the lagging strand, implying keto excess in coding sequences [69]. In agreement with our SCN simulations, these mutational bias strongly influence the amino acid frequencies for proteins coded in the two strands [70]. The protein sequences that we studied in this work were derived from the PDB, where nucleotide sequences are not stored. We did not find any study of nucleotide frequencies in genes coding for proteins in the PDB. Therefore, we could not compare our fitted parameters to nucleotide frequencies in the genes coding for the proteins in the PDB. It would be interesting to know whether the purine excess that our study suggests is observed in these genes.

Conclusion

We have shown that an evolutionary model with independently evolving sites is able to reproduce in a quantitative way the results of simulations where conservation of the thermodynamic stability of a protein native state is explicitly enforced, and therefore sites evolve in a correlated way. This mean-field model with independent sites

is also able to reproduce site-specific amino acid distributions at sites with specific values of the principal eigenvector of the contact matrix, in good agreement with the distributions observed in the whole PDB. As site-independent evolutionary models are readily amenable to computation, we expect these results to be widely applicable in the context of the reconstruction of evolutionary histories. Our results also demonstrate that the mutation process has a deep influence on protein evolution. It modifies the balance between stability against the unfolded state and stability against misfolded compact conformations, and it modifies the fraction of mutations which are eliminated by purifying selection. In our mean-field model, mutations and selection are strictly interrelated, as the effective selection probabilities depend on the mutation process.

Methods

SCN model of neutral evolution

In the Structurally Constrained Neutral (SCN) model of protein evolution [29-32] amino acid mutations are proposed randomly, and accepted according to a stability criterion. The stability of the folded protein, is defined by an effective free energy function, $E(A, C)$, based on contact interactions,

$$E(\mathbf{A},\mathbf{C}) = \sum_{i<j} C_{ij} U(A_i, A_j), \qquad (13)$$

where **A** represents the protein sequence, **C** is the contact map of the native structure, and **U** is a 20 × 20 symmetric matrix whose element $U(a, b)$ represents the effective interaction, in units of $k_B T$, of amino acids of types a and b; we use the interaction matrix derived by Bastolla *et al*. [38]. For most protein chains in the PDB this interaction matrix assigns lower effective free energy, Eq. (13), to the native structure than to decoys generated by threading, and it produces a well correlated free energy landscape. We then estimate two parameters: (i) The effective energy per residue, $E(\mathbf{A}, \mathbf{C})/N$, Eq. (13), where N is the protein length. This quantity correlates with the folding free energy per residue for a set of 18 small proteins that are folding with two-states thermodynamics (correlation coefficient $r = 0.91$; UB, unpublished result); (ii) The normalized energy gap α, which characterizes fast folding sequences [71] with well correlated energy landscapes [17,72-75]. In the SCN model, mutated sequences are considered thermodynamically stable if both stability parameters are above predetermined thresholds. Synonymous mutations are always accepted, whereas mutations to stop codons are always rejected.

The normalized energy gap $\alpha(\mathbf{A})$, estimating misfolding stability, is defined as the minimal value of the difference between the energy of the native configuration, **C***, and the energy of any compact configuration **C** satisfying the constraints of chain connectivity, excluded volume and hydrogen bonding, normalized times the absolute native energy and divided by the structural dissimilarity between the structures **C*** and **C**, 1 - q(**C***,**C**), where q represents the contact overlap,

$$\alpha(\mathbf{A}) = \min_{\mathbf{C}} \frac{E(\mathbf{A},\mathbf{C}) - E(\mathbf{A},\mathbf{C}^*)}{|E(\mathbf{A},\mathbf{C}^*)|\,[1 - q(\mathbf{C},\mathbf{C}^*)]}. \qquad (14)$$

The normalized energy gap is estimated in our simulations from a set of alternative configurations. It depends strongly on the size of this set: We typically use hundreds of thousands of structures that can be generated by threading the protein sequence on all non-redundant structures in the PDB.

For the analytic calculation reported below, but not for the simulations, we use an estimate of the normalized energy gap [47] based on the Random Energy Model [76,77]

$$\alpha(\mathbf{A}) \approx \frac{\langle \mathrm{U} \rangle_{\mathbf{A}} - \sigma_{U,\mathbf{A}} \sqrt{2\log(m_N)/N_c} - E(\mathbf{A},\mathbf{C}^*)/N_c}{|E(\mathbf{A},\mathbf{C}^*)/N_c|\,(1 - q_0)} \qquad (15)$$

Here, N_c is the number of contacts in the native structure, $\langle U \rangle_{\mathbf{A}}$ and $\sigma_{U,\mathbf{A}}$ are the mean and standard deviation of all possible contact interactions in the protein sequence **A**, both native and non-native, $q_0 = 0.1$ is a parameter representing the typical overlap of unrelated structures, N is chain length, m_N is the number of alternative structures compatible with the above constraints, estimated through the empirical formula $\log(m_N) \approx 0.1 \times N + 4$.

The above estimate may be improved considering that the probability of contact formation decreases with sequence distance [78]. However, this correction is small [47], and it will not be considered here because it would not allow us to get the analytic expression derived in the following section.

Optimal hydrophobicity profile and the principal eigenvector of the contact matrix

We report here for completeness the calculation originally developed in [48] on the relationship between sequence and structural profiles induced by stability conditions.

The contact interaction matrix can be approximated through the main component of its spectral decomposition, $U(a, b) \approx - h(a)h(b)$. Here $h(a)$ is the eigenvector of the matrix $U(a, b)$ corresponding to the eigenvalue with the largest absolute value, which is negative. This eigenvector is very strongly correlated with the hydrophobicity of amino acid a, as for typical amino acid interaction matrices based on contacts [40,41]. We call the 20 param-

eters $h(a)$ obtained from the principal eigenvector of the interaction matrix the *interactivity* hydrophobicity scale (IH), and we call the N-dimensional vector $h(A_i)$ the Hydrophobicity Profile (HP) of sequence **A** [48].

We use in the simulations the complete free energy function, Eq. (13), but in the analytic calculations we approximate it with the hydrophobic component,

$$H(\mathbf{A},\mathbf{C}) \equiv -\sum_{i<j} C_{ij} h(A_i) h(A_j), \qquad (16)$$

Using this approximation of the free energy, it is possible to derive an analytic relationship between the protein sequence and the protein structure. To this end, we calculate the optimally stable hydrophobicity profile, that minimizes the approximate effective free energy Eq. (16), for a given contact matrix, with a large normalized energy gap.

Using the REM estimate for the normalized energy gap, Eq. (15), we see that it depends on the protein sequence only through three parameters: the native energy and the mean and the standard deviation of the non-native contact interactions. Therefore, in order to minimize the native energy with a large normalized energy gap, we have to maintain fixed $\langle U\rangle_{\mathbf{A}}$ and $\sigma_{U,\mathbf{A}}$. Within the hydrophobic approximation of the contact interaction energy, it holds $\langle U\rangle \approx \langle h^2\rangle$ and $\langle U^2\rangle_{\mathbf{A}} \approx \langle h^2\rangle^2$.

In conclusion, we look for the hydrophobicity profile h_i that minimizes the effective hydrophobic energy, Eq. (16), for a given contact matrix, and for given first and second moment of the hydrophobicity vector, $\langle h\rangle = N^{-1}\sum_i h(A_i)$ and $\langle h^2\rangle = N^{-1}\sum_i h(A_i)^2$. In the calculation, we neglect the discretization in twenty values corresponding to natural amino acids.

The PE $v_i^{(1)}$ is the solution of the related minization problem in which no condition on $\langle h\rangle$ is imposed. We denote the eigenvalues of the contact matrix C_{ij} by λ_α and the corresponding eigenvectors by $v_i^{(\alpha)}$. These eigenvectors constitute an orthonormal basis. Expressing the constraints on $\langle h\rangle$ and $\langle h^2\rangle$ through Lagrange multipliers, one finds that the optimal HP is given by the following implicit expression,

$$h_i^{\mathrm{opt}} = \sqrt{N\langle h^2\rangle}\,\frac{\sum_\alpha \frac{\sqrt{w_\alpha}}{\Lambda-\lambda_\alpha} v_i^{(\alpha)}}{\sqrt{\sum_\alpha \frac{w_\alpha}{(\Lambda-\lambda_\alpha)^2}}} \qquad (17)$$

where the multiplier Λ is obtained through the constraint

$$\tau \equiv \frac{\langle h\rangle}{\sqrt{\langle h^2\rangle}} = \frac{\sum_\alpha \frac{w_\alpha}{\Lambda-\lambda_\alpha}}{\sqrt{\sum_\alpha \frac{w_\alpha}{(\Lambda-\lambda_\alpha)^2}}}. \qquad (18)$$

N represents the number of residues in the protein and $w_\alpha \equiv N\langle v^{(\alpha)}\rangle^2 = \langle v^{(\alpha)}\rangle^2/\langle (v^{(\alpha)})^2\rangle$, is the ratio between squared mean and mean square of the components of eigenvector α. Since the $v_i^{(\alpha)}$ constitute an orthonormal basis, it holds $\sum_\alpha w_\alpha = 1$. The weight w_1 of the principal eigenvector is the largest of the w_α. The projection of the optimal HP along the direction of the PE is thus given by

$$\frac{\sum_i h_i^{\mathrm{opt}} v_i^{(1)}}{\sqrt{N\langle h^2\rangle}} = \left(1 + \frac{(\Lambda-\lambda_1)^2}{w_1}\sum_{\alpha>1}\frac{w_\alpha}{(\Lambda-\lambda_\alpha)^2}\right)^{-1/2} \qquad (19)$$

For $\Lambda = \lambda_1$, the optimal HP is parallel to the PE, i.e. the coefficients of $v_i^{(\alpha)}$ vanish for $\alpha > 1$, as for the optimization without any constraint on $\langle h\rangle$. In this case, $\tau = \sqrt{w_1}$, i.e. $\langle h\rangle = \langle v^{(1)}\rangle \sqrt{\langle h^2\rangle/\langle (v^{(1)})^2\rangle}$, which means that the ratio between squared mean and mean square is the same for the HP and for the PE, and the energy is the same as in the absence of constraints on $\langle h\rangle$, which is the lowest energy for all possible values of the constraint $\langle h\rangle$.

A structure-derived quantity that estimates the importance of the minor eigenvectors with $\alpha > 1$ in determining the optimal hydrophobicity profile is

$$\eta_2 = \frac{1}{w_1}\sum_{\alpha>1}\frac{w_\alpha}{(\lambda_1-\lambda_\alpha)^2}, \qquad (20)$$

In the SCN model, the values of $\langle h\rangle$ and $\langle h^2\rangle$, and therefore of τ and Λ, are not fixed by stability requirements but vary in a broad range depending on the mutation process and on the selection parameters. Numerical results show that, when η_2 is small (smaller than, say, 0.1), the contribution of minor eigenvectors can be neglected in all the (quite broad) simulated range of parameters $\langle h\rangle$ and $\langle h^2\rangle$. This is the case of many single-domain proteins, since the w_α corresponding to minor eigenvectors tend to be small in these cases. On the other hand, for modular (for instance multi-domains) structures, the w_α of the eigenvectors that correspond to the minor domains are large, and these eigenvectors give a non-negligible contribution to the optimal HP.

For $\Lambda \approx \lambda_1$ and $\tau \approx \sqrt{w_1}$, it can be easily seen that the correlation coefficient between the optimal HP and the PE deviates from unity by a term of second order in $1 - r(\mathbf{h}_i^{\mathrm{opt}}, \mathbf{v}_i^{(1)}) \approx (\tau / \sqrt{w_1} - 1)^2$.

Even for extreme mutation bias, our simulation results yield $\tau \sqrt{w_1} \in [0.73, 0.97]$, which is close to one. Consistently, the constribution to minor eigenvectors is small in most of the range of parameters even for proteins for which η_2 is large.

Therefore, we consider in this work the approximation that the correlation coefficient between the optimal HP and the PE is one, corresponding to the zeroth order in the τ- $\sqrt{w_1}$ expansion or to a situation where η_2 is very small. For simplicity of notation, the PE will be denoted in the following as $c_i \equiv v_i^{(1)}$. As a result, we get

$$h_i^{\mathrm{opt}} \approx \sqrt{\frac{\langle h^2 \rangle - \langle h \rangle^2}{\left(\langle c^2 \rangle - \langle c \rangle^2 \right)}} (c_i - \langle c \rangle) + \langle h \rangle. \qquad (21)$$

Mutation process

The SCN model was originally defined at the protein sequence level, with equally probable mutations from one amino acid to any other one [30-32]. We have modified the mutation process in order to take into account the genetic code and the mutation bias at the DNA level (see also Ref. [53]). We represent each amino acid site by 3 nucleotides, and consider two mutation schemes: (1) Independent and identical mutation processes at each nucleotide site, each one satisfying detailed balance. (2) Same process as in (1), but with an enhanced mutation rate at CpG dinucleotides contained within a codon.

The mutation process (1) consists of the HKY mutation matrix [79,80] with rates $P_\mu^{\mathrm{nuc}}(n, n') = \mu f(n')$ if the mutation from n to n' is a transversion and $P_\mu^{\mathrm{nuc}}(n, n') = \mu k f(n')$ if it is a transition. Transitions are changes between the two purines A and G, or between the two pyrimidines C and T. Transversions are changes between a purine and a pyrimidine. Since they change less the chemical nature of the DNA basis, transitions are far more frequent than transversions. For convenience of notation, we define $t(n)$ as the nucleotide obtained from n through a transition (t(A) = G, t(T) = C, $t(t(n)) \equiv n$). The diagonal elements of the mutation matrix are defined through the normalization condition $P_\mu^{\mathrm{nuc}}(n, n) = 1 - \sum_{n' \neq n} P_\mu^{\mathrm{nuc}}(n, n')$. This mutation process satisfies detailed balance, with stationary distribution given by the frequencies $f(n)$ independently of the transition-tranversion ratio k that, therefore, is expected to have no influence on the stationary amino acid distribution as well.

In SCN simulations, in order to reduce the number of parameters, we further imposed the condition that the mutation process is the same on the two DNA strands, so that f(G) ≡ f(C) and f(A) ≡ f(T). Therefore, the stationary distributions only depend on the GC bias f(G)/f(A).

The mutation process (2) starts from the same model as in (1), but every time a codon contains a CpG dinucleotide the rate of the mutations from C to T and from G to A are enhanced by a factor $k_{\mathrm{CpG}} \geq 1$. This model was considered with a transition-transversion ratio k = 1, and it was only used in calculations with the mean-field model, and not in SCN simulations. For simplicity, only CpG dinucleotides within a codon were considered, so that we obtained an independent mutation process for each codon. From the above definition, we computed the mutation matrix at the codon level to be used in the master equation (8), $P_\mu^{\mathrm{COD}}(n_1 n_2 n_3, n_1' n_2' n_3')$. The matrix element is set to zero if the two codons differ at more that one position and to $P_\mu^{\mathrm{nuc}}(n, n')$ if the two codons differ at one position where they contain respectively nucleotides n and n'. If the mutated nucleotide is either the C or the G of a CpG dinucleotide contained into codon $n_1 n_2 n_3$ and the mutation is a transition (C to T or G to A), then the matrix element is increased by a factor k_{CpG}.

In the SCN model, the mutation process is simulated extracting at random at each time step the site where a mutation takes place. The probability that a site is extracted depends on the nucleotide occupying it, and it is $p(n) = \sum_{n' \neq n} f(n') = (k - 1) f(t(n))$.

Calculation of the mean-field distributions

The mean-field amino acid distributions were computed in two steps. In a first step, we computed the site-specific mean hydrophobicities $[h_i]$ using Eq. (5), which needs as input the distribution of PE values, i.e. the fraction of sites with $c_i/\langle c \rangle$ in a given range, and the mean and standard deviation of the hydrophobicity, which were obtained from the protein sequences. In a second step, and for a given mutation model, the site-specific mean-field distributions were calculated as a function of β and a value β_i was associated to each site in such a way that the mean

hydrophobicity at the site coincides with the predicted one (numerically, the predicted value of $[h_i]$ was bounded between an upper and a lower bound corresponding to two β values, and β_i was found by interpolation).

The mean hydrophobicity was calculated as $[h]\beta = \sum_a h(a)\pi(\beta, a)$. For mutation models fulfilling detailed balance, we used Eq. (10), $\pi(\beta, a) \propto w_{AA} \exp[-(\beta h(a)]$, with weights $w_{AA}(a)$ obtained from the mutation model. For mutation models not obeying detailed balance, $\pi(\beta, a)$ was numerically computed as the stationary distribution of the Markov process, Eq. (8).

Computation of the acceptance rate

The rate of acceptance of a mutation at position i in the stationary state was calculated as

$$P_{\mathrm{acc},i} = \frac{\sum_{\mathbf{nn}'} P_i^{\mathrm{COD}}(\mathbf{n}) P_\mu^{\mathrm{COD}}(\mathbf{n},\mathbf{n}') \min(1, \exp[-\beta_i [h(\mathcal{A}[\mathbf{n}']) - h(\mathcal{A}[\mathbf{n}])]])}{\sum_{\mathbf{nn}'} P_i^{\mathrm{COD}}(\mathbf{n}) P_\mu^{\mathrm{COD}}(\mathbf{n},\mathbf{n}')}, \qquad (22)$$

where we use the notation introduced previously, $P_i^{\mathrm{COD}}(\mathbf{n})$ is the stationary frequency of codon **n** in the mean-field model, and the summations exclude stop codons.

Observed amino acid distributions

We compared our predictions to site-specific distributions sampled from a representative subset of the Protein Data Bank (PDB). We considered a non-redundant subset of single-domain globular proteins in the PDB, with a sequence identity below 25% [58]. Globularity was verified by imposing that the fraction of contacts per residue was larger than a length dependent threshold, $N_c/N > 3.5 + 7.8N^{-1/3}$. 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 so as to eliminate outliers with respect to the general trend, which are mainly non-globular structures. The condition of being single-domain was verified by imposing that the normalized variance of the PE components was smaller than a threshold, $(1 - N\langle c\rangle^2)/(N\langle c\rangle^2) < 1.5$. Multi-domain proteins have PE components which are large inside the main domain and small outside it. The PE components would be exactly zero outside the main domain if the domains do not share contacts (see for instance Ref. [61]). 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 selected 404 such structures with 200 or less amino acids. We counted the number of each of the 20 amino acids as a function of $c_i/\langle c\rangle$, where $\langle c\rangle$ denotes the average over a single structure. We used a bin-size of 0.05 for $c_i/\langle c\rangle \leq 2.5$ and a bin-size of 0.1 for $c_i/\langle c\rangle > 2.5$.

Similarity score between observed and predicted amino acid distributions

The accuracy of the predicted amino acid distributions was assessed by calculating the mean correlation coefficient between observed and predicted amino acid distributions for M structures, given by

$$\langle r\rangle = M^{-1}\sum_{i=1}^{M} r\left(\pi^{\mathrm{obs}}_{c_i/\langle c\rangle}, \pi^{\mathrm{pred}}_{c_i/\langle c\rangle}\right)$$

Optimization of the mutation parameters

The optimal values for the parameters of the different mutation models were found by maximizing the mean correlation coefficient $\langle r\rangle$ between observed and predicted amino acid distributions as defined in the previous subsection. First, we discretized the possible values of the free parameters within a reasonable range, using a step size of 0.001, and we numerically assessed all possible combinations. We then performed a (much faster) optimization by gradient descent, finding the same results up to relative precision of 10^{-3}.

Hydropathy scales

In this work, mean-field distributions were obtained using interactivity (IH, see below) as hydrophobicity scale h(a), both for comparison with SCN simulations and with amino acid distributions sampled from the PDB. However, for the latter case we tested eleven hydropathy scales, finding that all other scales provide worse results. They are: (1) The KD82 hydropathy scale, derived to identify trans-membrane helices using diverse experimental data [81]; (2) The L76 hydropathy scale, which was derived by using experimental data and theoretical calculations [82]; (3) The R88 hydropathy scale, which is based on the transfer of solutes from water to alkane solvents [83]; (4) The augmented Whilmey-White (WW01) hydropathy scale, derived to improve recognition of trans-membrane helices [84]; (5) The G98 classification of amino acids into polar, hydrophobic, and amphiphylic classes, adopted by Gu et al. [85] to investigate the relationship between the hydrophobicity of a protein and the nucleotide composition of the corresponding gene; (6) The MP78 hydropathy scale, derived from statistical properties of globular proteins [86]; (7) The AV hydropathy scale, derived by averaging 127 normalized hydropathy scales published in the literature [87]; (8) The FP83 hydropathy scale, derived from the experimental measurement of octanol/water partition coefficients [42]; (9) The ZZ04 scale, also called buriability, proposed by Zhou and Zhou [59]; (10) The interaction scale IH, obtained from the main eigenvector of the interaction matrix $U(a, b)$ used in this work [48]; (11) The optimized interactivity

scale, or connectivity scale CH, which maximizes the correlation with the principal eigenvector of protein contact matrices for a non-redundant set of Protein Data Bank (PDB) structures [48].

Authors' contributions

UB wrote the code for the SCN simulations, developed the mean-field model, analyzed PDB sequences with the mean-field model, contributed to the analysis of the data and wrote the first version of the paper. MP performed the SCN simulations, analyzed PDB sequences with the mean-field model, and contributed to the analysis of the data and the writing of the paper. HER contributed to the analysis of the data and the writing of the paper. MV contributed to the analysis of the data and the writing of the paper.

Note

[1] Here and in the following, the angular brackets $(\cdot)$ denote the average over all sites in the protein.

[2]We also verified that, in agreement with Eq. (17), when other eigenvectors are relevant their scalar product with the average HP is correlated with the factor $w_\alpha/(\lambda_1 - \lambda_\alpha)$. For myoglobin the correlation coefficients, for the 15 most relevant eigenvectors excluding the PE, are in the range between 0.73 and 0.93, except for the extreme mutation bias, for ATPE they are always larger than 0.81.

Acknowledgements

UB acknowledges financial support from the I3P program of the Spanish CSIC, co-funded by the European Social Fund, and from the project FIS2004-05073-C04-04 of the Spanish Ministery of Education and Science. MP acknowledges financial support from the Deutsche Forschungsgemeinschaft via project PO 1025/1-1. We are grateful to an anonymous referee for valuable comments and suggestions.

References

1. Nei M, Kumar S: **Molecular evolution and phylogenetics.** Oxford Univ. Press; 2000.
2. Graur D, Li WH: **Fundamentals of molecular evolution.** Sinauer, Sunderland; 2000.
3. Felsenstein J: **Evolutionary trees from DNA sequences: A maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.
4. Lockless SW, Ranganathan R: **Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families.** *Science* 1999, **286:**295-299.
5. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R: **Evolutionary information for specifying a protein fold.** *Nature* 2005, **437:**512-518.
6. Parisi G, Echave J: **Structural constraints and emergence of sequence patterns in protein evolution.** *Mol Biol Evol* 2001, **18:**750-756.
7. Parisi G, Echave J: **The structurally constrained protein evolution model accounts for sequence patterns of the LH superfamily.** *BMC Evol Biol* 2004, **4:**41. doi:10.1186/1471-2148-4-41.
8. Fornasari MS, Parisi G, Echave J: **Site-specific amino acid replacement matrices from structurally constrained protein evolution.** *Mol Biol Evol* 2002, **19:**352-356.
9. Robinson DM, Jones DT, Kishino EL, Goldman N, Thorne JL: **Protein evolution with dependence among codons due to tertiary structure.** *Mol Biol Evol* 2003, **20:**1692-1704.
10. Rodrigue N, Lartillot N, Bryant D, Philippe H: **Site interdependence attributed to tertiary structure in amino acid sequence evolution.** *Gene* 2005, **347:**207-217.
11. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15:**910-7.
12. Kimura M: **Evolutionary rate at the molecular level.** *Nature* 1968, **217:**624-626.
13. Kimura M: **The neutral theory of molecular evolution.** Cambridge Univ. Press; 1983.
14. Schuster P, Fontana W, Stadler PF, Hofacker IL: **From sequences to shapes and back – A case-study in RNA secondary structures.** *Proc R Soc London B* 1994, **255:**279-284.
15. Huynen MA, Stadler PF, Fontana W: **Smoothness within ruggedness: The role of neutrality in adaptation.** *Proc Natl Acad Sci USA* 1996, **93:**397-401.
16. Fontana W, Schuster P: **Continuity in evolution: on the nature of transitions.** *Science* 1998, **280:**1451-1455.
17. Gutin AM, Abkevich VI, Shakhnovich EI: **Evolution-like selection of fast-folding model proteins.** *Proc Natl Acad Sci USA* 1995, **92:**1282-1286.
18. Govindarajan S, Goldstein RA: **Evolution of model proteins on a foldability landscape.** *Proteins* 1997, **29:**461-466.
19. Govindarajan S, Goldstein RA: **On the thermodynamic hypothesis of protein folding.** *Proc Natl Acad Sci USA* 1998, **95:**5545-5549.
20. Taverna DM, Goldstein RA: **The distribution of structures in evolving protein populations.** *Biopolymers* 2000, **53:**1-8.
21. Bornberg-Bauer E: **How are model protein structures distributed in sequence space?** *Biophys J* 1997, **73:**2393-2403.
22. Bornberg-Bauer E, Chan HS: **Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space.** *Proc Natl Acad Sci USA* 1999, **96:**10689-10694.
23. Babajide A, Hofacker IL, Sippl MJ, Stadler PF: **Neutral networks in protein space.** *Fol Des* 1997, **2:**261-269.
24. Bussemaker HJ, Thirumalai D, Bhattacharjee JK: **Thermodynamic stability of folded proteins against mutations.** *Phys Rev Lett* 1997, **79:**3530-3533.
25. Tiana G, Broglia RA, Roman HE, Vigezzi E, Shakhnovich EI: **Folding and misfolding of designed proteinlike chains with mutations.** *J Chem Phys* 1998, **108:**757-761.
26. Mirny LA, Abkevich VI, Shakhnovich EI: **How evolution makes proteins fold quickly.** *Proc Natl Acad Sci USA* 1998, **95:**4976-4981.
27. Dokholyan NV, Shakhnovich EI: **Understanding hierarchical protein evolution from first principles.** *J Mol Biol* 2001, **312:**289-307.
28. Dokholyan NV, Mirny LA, Shakhnovich EI: **Understanding conserved amino acids in proteins.** *Physica A* 2002, **314:**600-606.
29. Bastolla U, Roman HE, Vendruscolo M: **Neutral evolution of model proteins: Diffusion in sequence space and overdispersion.** *J Theor Biol* 1999, **200:**49-64.
30. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Lack of self-averaging in neutral evolution of proteins.** *Phys Rev Lett* 2002, **89:**208101/1-208101/4.
31. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Connectivity of neutral networks, overdispersion and structural conservation in protein evolution.** *J Mol Evol* 2003, **56:**243-254.
32. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Statistical properties of neutral evolution.** *J Mol Evol* 2003, **57:**S103-S119.
33. van Nimwegen E, Crutchfield JP, Huynen M: *Proc Natl Acad Sci USA* 1999, **96:**9716.
34. Taverna DM, Goldstein RA: **Why are proteins so robust to site mutations?** *J Mol Biol* 2002, **315:**479-484.
35. Wilke CO: **Molecular clock in neutral protein evolution.** *BMC Genetics* 2004, **5:**25. doi:10.1186/1471-2156-5-25
36. Ohta T, Kimura M: **On the constancy of the evolutionary rate of cistrons.** *J Mol Evol* 1971, **1:**18-25.
37. Gillespie JH: **The causes of molecular evolution.** *Oxford University Press* 1991.
38. 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.
39. Bastolla U, Moya A, Viguera E, van Ham RCHJ: **Genomic determinants of protein folding thermodynamics.** *J Mol Biol* 2004, **343:**1451-1466.

40. 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:**725-32.
41. 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.
42. Fauchere JL, Pliska V: **Hydrophobic parameters of amino acid side chain from the partitioning N-acetyl amino acid amides.** *Eur J Med Chem* 1983, **18:**369-375.
43. Dobson CM: **Protein folding and misfolding.** *Nature* 2003, **426:**884-890.
44. Rutherford SL, Lindquist S: **Hsp90 as a capacitor for morphological evolution.** *Nature* **396:**336-342.
45. Agashe VR, Hartl FU: **Roles of molecular chaperones in cytoplasmic protein folding.** *Semin Cell Dev Biol* 2000, **11:**15-25.
46. Fares MA, Ruiz-Gonzalez MX, Moya A, Elena SF, Barrio E: **GroEL buffers against deleterious mutations.** *Nature* 2002, **417:**398.
47. Bastolla U, Demetrius L: **Stability constraints and protein evolution: the role of chain length, composition, and disulphide bonds.** *Prot Eng Des and Sel* 2005, **18:**405-415.
48. Bastolla U, Porto M, Roman HE, Vendruscolo M: **The principal eigenvector of contact matrices and hydrophobicity profiles in proteins.** *Proteins* 2005, **58:**22-30.
49. 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. Erratum: *Mol Biol Evol* **22**:1156
50. Koshi JM, Goldstein RA: **Models of natural mutation including site heterogeneity.** *Proteins* 1998, **32:**289-295.
51. Koshi JM, Mindell DP, Goldstein RA: **Using physical-chemistry based substitution models in phylogenetic analysis of HIV-1 subtypes.** *Mol Biol Evol* 1999, **16:**173-179.
52. Finkelstein AV, Gutin AM, Badretdinov AY: **Boltzmann-like statistics of protein architectures. Origins and consequences.** *Subcell Biochem* 1995, **24:**1-26.
53. 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.
54. Sueoka N: **Intrastrand parity rules of DNA base composition and usage biases of synonymous codons.** *J Mol Evol* 1995, **40:**318-325. *J. Mol. Evol.* **42**:323
55. Ohta T: **Role of very slightly deleterious mutations in molecular evolution and polymorphism.** *Theor Pop Biol* 1976, **10:**254-275.
56. Berg J, Willmann S, Lassig M: **Adaptive evolution of transcription factor binding sites.** *BMC Evol Biol* 2004, **4:**42.
57. Sella G, Hirsch AE: **The application of statistical physics to evolutionary biology.** *Proc Natl Acad Sci USA* 2005, **102:**9541-9546.
58. Hobohm U, Sander C: **Enlarged representative set of protein structure.** *Protein Sci* 1994, **3:**522-524.
59. Zhou H, Zhou Y: **Quantifying the effect of burial of amino acid residues on protein stability.** *Proteins* 2004, **54:**315-322.
60. Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules* 1985, **18:**534-552.
61. Porto M, Bastolla U, Roman HE, Vendruscolo M: **Reconstruction of protein contact maps from their principal eigenvectors.** *Phys Rev Lett* 2004, **92:**218101/1-218101/4.
62. Vendruscolo M, Subramanian B, Kanter I, Domany E, Lebowitz JL: **Statistical properties of contact maps.** *Phys Rev E* 1999, **59:**977-984.
63. Bernardi G, Bernardi G: **Compositional constraints and genome evolution.** *J Mol Evol* 1986, **24:**1-11.
64. Lobry JR: **Influence of genomic G+C content on average amino acid composition of proteins from 59 bacterial species.** *Gene* 1997, **205:**309-316.
65. Lobry JR: **Asymmetric substitution patterns in the two DNA strands of bacteria.** *Mol Biol Evol* 1996, **13:**660-665.
66. Jordan IK, Konradshov FA, Adzhubei IA, Wolf YL, Koonin EV, Konradshov AS, Sunyaev S: **A universal trend of amino acid gain and loss in protein evolution.** *Nature* 2005, **433:**633-638.
67. Bastolla U, Porto M, Roman HE, Vendruscolo M: **The Structurally Constrained Neutral Model of Protein Evolution.** In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* Edited by: Bastolla, U, Porto M, Roman HE, Vendruscolo M. Springer Verlag; 2006.
68. Freeman JM, Plasterer TN, Smith TF, Mohr SC: **Patterns of genome organization in bacteria.** *Science* 1998, **279:**1827 [http://bmerc-www.bu.edu/genomeplot/].
69. McLean MJ, Wolfe KH, Devine KM: **Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes.** *J Mol Evol* 1998, **47:**691-696.
70. Rocha EPC, Danchin A, Viari A: **Universal replication biases in bacteria.** *Mol Microb* 1999, **32:**11-16.
71. Bastolla U, Frauenkron H, Gerstner E, Grassberger P, Nadler W: **Testing a new Monte Carlo algorithm for protein folding.** *Proteins* 1998, **32:**52-66.
72. Bryngelson JD, Wolynes PG: **Spin-glasses and the statistical-mechanics of protein folding.** *Proc Natl Acad Sci USA* 1987, **84:**7524-7528.
73. Goldstein RA, Luthey-Schulten ZA, Wolynes PG: **Optimal protein-folding codes from spin-glass theory.** *Proc Natl Acad Sci USA* 1992, **89:**4918-4922.
74. Abkevich VI, Gutin AM, Shakhnovich EI: **Free energy landscapes for protein folding kinetics – intermediates, traps and multiple pathways in theory and lattice model simulations.** *J Chem Phys* 1994, **101:**6052-6062.
75. Klimov DK, Thirumalai D: **Factors governing the foldability of proteins.** *Proteins* 1996, **26:**411-441.
76. Derrida B: **Random Energy Model: an exactly solvable model of disordered systems.** *Phys Rev B* 1981, **24:**2613.
77. Shakhnovich EI, Gutin AM: **Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach.** *Biophys Chem* 1989, **34:**187-199.
78. Govindarajan S, Goldstein RA: **Optimal local propensities for model proteins.** *Proteins* 1995, **22:**413-8.
79. 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.
80. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein coding DNA sequences.** *Mol Biol Evol* 1994, **11:**725-36.
81. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157:**105-132.
82. Levitt M: **A simplified representation of protein conformations for rapid simulation of protein folding.** *J Mol Biol* 1976, **104:**59-107.
83. Roseman MA: **Hydrophobicity of polar amino-acid side chains is markedly reduced by flanking peptide bonds.** *J Mol Biol* 1988, **200:**513-522.
84. Jayasinghe S, Hristova K, White SH: **Energetics, stability, and prediction of transmembrane helices.** *J Mol Biol* 2001, **312:**927-934.
85. Gu X, Hewett-Emmett D, Li WH: **Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria.** *Genetica* 1998, **102–103:**383-391.
86. Manavalan P, Ponnuswamy PK: **Hydrophobic character of amino acid residues in globular proteins.** *Nature* 1978, **275:**673-674.
87. Palliser CC, Parry DA: **Quantitative comparison of the ability of hydropathy scales to recognize surface beta-strands in proteins.** *Proteins* 2001, **42:**243-255.