A peer-reviewed version of this preprint was published in PeerJ on 5 October 2018.

View the peer-reviewed version (peerj.com/articles/5549), which is the preferred citable publication unless you specifically need to cite this preprint.

Jiménez-Santos MJ, Arenas M, Bastolla U. 2018. Influence of mutation bias and hydrophobicity on the substitution rates and sequence entropies of protein evolution. PeerJ 6:e5549 https://doi.org/10.7717/peerj.5549

Influence of mutation bias and hydrophobicity on the substitution rates and sequence entropies of protein evolution

María José Jiménez-Santos [1], **Miguel Arenas** [2], **Ugo Bastolla** Corresp. [1]

[1] Bioinformatics Unit, Center for Molecular Biology Severo Ochoa, CSIC-UAM, Madrid, Spain

[2] Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain

Corresponding Author: Ugo Bastolla
Email address: ubastolla@cbm.csic.es

Protein sites present different amino acids during their evolution, whose number reflects the selective constraints operating on them. This evolutionary variability is strongly influenced by the structural properties of the site in the native structure, and it is quantified either through sequence entropy or through substitution rates. However, while the sequence entropy only depends on the equilibrium frequencies of the amino acids, the substitution rate also depends on the exchangeability matrix that describes mutations in the mathematical model of the substitution process.

[p]Here we apply a mathematical model of protein evolution with selection for protein stability, both against unfolding and against misfolding, and find that sites with the same sequence entropy present different substitution rates depending on whether the site is prevalently hydrophobic or hydrophylic. For equal sequence entropy, polar sites evolve faster than hydrophobic sites. This is a consequence of the differential exchangeability associated with hydrophobic or polar amino acids. Accordingly, the model predicts that more polar proteins present, on the average, a faster substitution rate. However, these results change if we compare proteins that evolve under different mutation biases, such as orthologous proteins in different bacterial genomes. In this case, the substitution rates are faster in genomes that evolve under mutational bias that favour hydrophobic amino acids by preferentially incorporating the nucleotide Thymine that is more frequent in hydrophobic codons. In our model, the mutation bias influences both the sequence entropies and the substitution rates of protein sites. The sequence entropy is maximal for the mutational biases that reproduce the observed amino acid distributions and strongly decreases when extreme mutational biases are approached. The hydrophobicity for which the entropy is maximal is close to the mean hydrophobicity of the twenty amino acids and independent of the mutation bias. In contrast, the substitution rate and the hydrophobicity for which the substitution rate is maximal tend to increase when the mutation bias favours hydrophobic amino acids. Thus, changes of the mutational bias lead to deep effects on the biophysical properties of the protein (hydrophobicity) and on its evolutionary properties (sequence entropy and substitution rate) at the same time. The program Prot_evol is freely available for download at the url https://ub.cbm.uam.es/prot_fold_evol/prot_fold_evol_soft_main.php#Prot_Evol.[p]

Influence of mutation bias and hydrophobicity on the substitution rates and sequence entropies of protein evolution

María José Jiménez[(1)], Miguel Arenas[(2)], and Ugo Bastolla[(1,3)]

[(1)] Centro de Biologia Molecular "Severo Ochoa"

CSIC-UAM Cantoblanco, 28049 Madrid, Spain

[(2)] Department of Biochemistry, Genetics and Immunology,

University of Vigo, Vigo, Spain

[(3)] E-mail: ubastolla@cbm.csic.es

April 12, 2018

Abstract

Protein sites present different amino acids during their evolution, whose number reflects the selective constraints operating on them. This evolutionary variability is strongly influenced by the structural properties of the site in the native structure, and it is quantified either through sequence entropy or through substitution rates. However, while the sequence entropy only depends on the equilibrium frequencies of the amino acids, the substitution rate also depends on the exchangeability matrix that describes mutations in the mathematical model of the substitution process.

Here we apply a mathematical model of protein evolution with selection for protein stability, both against unfolding and against misfolding, and find that sites with the same sequence entropy present different substitution rates depending on whether the site is prevalently hydrophobic or hydrophylic. For equal sequence entropy, polar sites evolve faster than hydrophobic sites. This is a consequence of the differential exchangeability associated with hydrophobic or polar amino acids. Accordingly, the model predicts that more polar proteins present, on the average, a faster substitution rate. However, these results change if we compare proteins that evolve under different mutation biases, such as orthologous proteins in different bacterial genomes. In this case, the substitution rates are faster in genomes that evolve under mutational bias that favour hydrophobic amino acids by preferentially incorporating the nucleotide Thymine that is more frequent in hydrophobic codons.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

In our model, the mutation bias influences both the sequence entropies and the substitution rates of protein sites. The sequence entropy is maximal for the mutational biases that reproduce the observed amino acid distributions and strongly decreases when extreme mutational biases are approached. The hydrophobicity for which the entropy is maximal is close to the mean hydrophobicity of the twenty amino acids and independent of the mutation bias. In contrast, the substitution rate and the hydrophobicity for which the substitution rate is maximal tend to increase when the mutation bias favours hydrophobic amino acids. Thus, changes of the mutational bias lead to deep effects on the biophysical properties of the protein (hydrophobicity) and on its evolutionary properties (sequence entropy and substitution rate) at the same time. The program `Prot_evol` is freely available for download at the url https://ub.cbm.uam.es/prot_fold_evol/prot_fold_evol_soft_main.php#Prot_Evol.

Introduction

The evolutionary variability of an amino acid site in a protein family is an important indicator of the selective constraints that the site experiences. This variability, quantified either through the sequence entropy or through the substitution rate, is strongly influenced by the structural properties of the site in the native state of the protein (Echave, Spielman and Wilke, 2016). In particular, the substitution rate changes dramatically between exposed and buried sites, in such a way that buried sites tend to evolve more slowly than exposed sites, which is generally attributed to the fact that natural selection imposes stronger constraints on buried sites (Franzosa & Xia, 2009). It was later shown that the number of native inter-residue contacts formed by a protein site, which is negatively correlated with the solvent accessibility, is a stronger predictor of the substitution rate (Yeh et al. 2014).

Two different models rationalize why sites that form many contacts are subject to stronger selective constraints. The first kind of model, which we call stability-constrained fitness model, models the fitness as the fraction of protein found in the native state, which is a sigmoidal function of the folding free energy ΔG, i.e. $f = 1/(1+\exp(-\Delta G/kT))$ (see Goldstein 2011; Serohijos & Shakhnovich 2014; Bastolla Dehouck & Echave 2017). The second kind of model is the structurally-constrained model of protein evolution, which estimates how mutations affect the structure of the native state and computes the fitness from this predicted structural change (Echave 2008). Note in the literature the stability-constrained model is sometimes called structurally-constrained, but we think that this wording is misleading. In fact, for technical reasons, stability-constrained models estimate the change in free energy upon mutation assuming that the native structure does not change, whereas structurally-constrained models model the mutation as a perturbation applied to the wild-type structure and predict the extent by which the structure is perturbed through the Elastic Network Model (ENM, Tirion 1996) and linear response theory, but assume that the stability does not change. Thus, stability-constrained models predict the effect of mutations through the predicted stability change but neglect the cor-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

responding structure change, and structure-constrained models adopt the complementary perspective. Of course mutations modify both the stability and the precise structure of the native state, but current models of fitness cannot compute both effects.

In a recent work, we have shown that stability-constrained models that take into account negative design for destabilizing misfolded conformations (Berezovsky, Zeldovich & Shakhnovich 2007; Noivirt-Brik, Horovitz & Unger 2009; Minning, Porto & Bastolla 2013) predict that both the substitution rate and the entropy are maximal not at exposed sites with few contacts, as observed, but at sites where the number of contacts is intermediate, which can accomodate both hydrophobic and polar amino acids and are predicted to be extremaly tolerant to mutations (Jimenez, Arenas & Bastolla, 2018). On the other hand, when stability with respect to misfolding is neglected, stability-constrained models predict that the variability is maximal at exposed sites with few contacts (Scherrer, Meyer & Wilke 2012; Echave, Jackson & Wilke, 2015), but these kinds of models overestimate both the tolerance to mutations and the average hydrophobicity at almost all positions (Jimenez, Arenas & Bastolla 2018) and they much score worse than models that consider misfolding in likelihood calculations (Arenas & Bastolla 2015), so that models that consider misfolding have to be preferred. In contrast, structure-constrained models correctly predict that the variability is inversely related with the number of native contacts (Huang et al. 2014). These results support the view that the structural effect of mutations cannot be neglected, in particular at sites with intermediate numbers of contacts that are extremely tolerant to mutations under the point of view of the stability.

Here we adopt the stability-constrained mean-field (MF, Arenas & Bastolla 2015; Bastolla et al. 2006) and wild-type (WT, Jimenez, Arenas & Bastolla 2018) models of protein evolution that we used in the above-mentioned study. These models assume that sites in the protein evolve independently in a site-specific manner, and determine their site-specific properties by imposing a global constraint on the thermodynamic stability of the known native state against both unfolding and misfolding. The MF model significantly improves the likelihood of inferred evolutionary events with respect to empirical models that do not take into account the structural properties of each site (Arenas & Bastolla 2015), and it improves the reconstruction of the stability properties of ancestral sequences (Arenas et al. 2017). The WT models shows even better performances on several data sets (Arenas & Bastolla, in preparation). Both models exploit the formal analogy between the Boltzmann distribution in statistical physics, in which the probability of each conformation depends on the energy changed of sign and on the inverse of the temperature, and the stationary distribution of a protein family in which the probability of each sequence depends on its fitness and on the effective population size (Sella & Hirsh 2005, Mustonen & Lassig 2005). After the stationary distribution has been determined, the full substitution process is constructed applying the Halpern and Bruno formulas (Halpern & Bruno 1998), which impose that the fixation probabilities agree with Kimura's formulas (Kimura 1962. Both formulas are reproduced below). In the MF model, the effect on stability of amino-acid a at site i is predicted self-consistently against the MF distribution at all other sites, in the spirit of mean-field models in statistical mechanics. In turn,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

the WT model predicts the effect on stability and fitness of mutations of the wild-type sequence towards amino acid a at site i. Thus, in theory the WT model is more suited for short evolutionary divergences and the MF model is more suited for long evolutionary divergences (Arenas & Bastolla, in preparation).

Here we address the question whether the sequence entropy and the substitution rate are equivalent measures of the evolutionary variability of a position. We find that these two measures are not equivalent, since the sequence entropy is only influenced by the equilibrium distribution of amino acids while the substitution rate is also influenced by the mutation process that acts in evolution. As we shall see, both measures are deeply but differently influenced by hydrophobicity, both at the level of the individual protein sites and at the level of the average hydrophobicity induced by the mutational process.

Materials and methods

Stability constrained fitness model

Stability constrained models of protein evolution assume that the fitness of a protein with sequence $\mathbf{A}$ is proportional to the fraction of protein that is in the native state, which can be computed from the folding free energy as (Goldstein 2011; Serohijos & Shakhnovich 2014)

$$f(\mathbf{A}) = \mathrm{e}^{-\Delta G(\mathbf{A})/kT} / \left(1 + \mathrm{e}^{-\Delta G(\mathbf{A})/kT}\right) . \tag{1}$$

The computation is performed assuming that the native contact matrix C^{nat} does not change in evolution. Upon single mutation, the free energy change $\Delta G_{\mathrm{mut}} = \Delta G_{\mathrm{wt}} + \Delta\Delta G$ is predicted adopting some models of protein stability (see below).

Equilibrium distribution

Another approximation that is often used in these models is that the mutation rate is extremely slow ($N\mu \ll 1$) so that at every time there is only one mutant gene that "competes" with the wild-type gene for fixation in the population with effective population size of N individuals. Under this scenario, the probability that the mutation gets fixed in the population can be computed with Kimura's formula (Kimura 1962) as

$$P_{\mathrm{fix}}\left(\mathbf{A}^{\mathrm{wt}} \to \mathbf{A}^{\mathrm{mut}}\right) = \frac{\mathrm{e}^{-\left(\varphi(\mathbf{A}^{\mathrm{mut}}) - \varphi(\mathbf{A}^{\mathrm{wt}})\right)} - 1}{\mathrm{e}^{-N\left(\varphi(\mathbf{A}^{\mathrm{mut}}) - \varphi(\mathbf{A}^{\mathrm{wt}})\right)} - 1} \tag{2}$$

where $\varphi(\mathbf{A}) = \log(f(\mathbf{A}))$ is the logarithmic fitness associated with the amino acid sequence $\mathbf{A}$, Eq.(1). As it is well known, the fixation probability tends to the neutral limit $P_{\mathrm{fix}} = 1/N$ when $\Delta\varphi$ tends to zero, it tends exponentially to zero when $\Delta\varphi$ is negative and large, and it tends to $1 - \mathrm{e}^{-\Delta\varphi}$ when $\Delta\varphi$ is positive. Nearly neutral mutations with selective effect $|\Delta\varphi| \approx 1/N$ are likely to be fixed even when their effect is deleterious (Ohta 1976). Importantly, the above fixation probability defines a Monte Carlo process in sequence

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

space that fulfils detailed balance, so that its stationary distribution can be computed exactly (Sella & Hirsh 2005; Mustonen & Lassig 2005), except for the normalization constant, which would require a sum over 20^L possible sequences $\mathbf{A} = A_1 \cdots A_L$:

$$P(A_1 \cdots A_L) \propto \exp\left((N-1)\varphi(A_1 \cdots A_L)\right) \tag{3}$$

Note the analogy between this formula and the Boltzmann distribution with energy equal to $-\varphi$ and temperature equal to $1/(N-1)$. This explicit formula holds when the mutation process is unbiased, so that all sequences are equally probable under the mutation model. In the presence of mutation bias, the stationary distribution can be determined as the distribution with minimal Kullback-Leibler divergence from the mutational distribution, $d_{\rm KL} = \sum_A P^{\rm mut}(\mathbf{A})\left[\log(P^{\rm mut}(\mathbf{A})) - \log(P(\mathbf{A}))\right]$, with a constraint on the average fitness $\sum_A P(\mathbf{A})\varphi(\mathbf{A})$. This condition generalizes the Boltzmann principle, and it was adopted for developing the mean-field model of protein evolution (Arenas & Bastolla, 2015).

Mean-field model of protein evolution

The mean-field (MF) model assumes that the equilibrium amino acid distribution is the product of independent distributions at each protein site,

$$P(A_1, \cdots A_L) = \prod_{i=1}^{L} P^i(A_i)\,. \tag{4}$$

Of course this assumption is not realistic, since different sites determine protein stability through their interactions, but it is needed for performing likelihood computations in an efficient way. Our strategy consists in determining the effect of a mutation at site i self-consistently, with respect to the MF distribution at all other sites. For simplicity, we shall sometimes use the vectorial notation P^i_a for indicating $P^i(a)$, where a denotes one of the twenty amino acid types.

The mean-field distribution is determined by minimizing the Kullback-Leibler divergence (distance between distribution) with respect to a global mutational distribution $P^{\rm mut}_a$, i.e. $\sum_{ia} P^i_a \log\left(P^i_a/P^{\rm mut}_a\right)$. We impose a constraint on the average fitness, which is transformed into a constraint on the folding free energy ΔG. This condition on stability is imposed through the Lagrange multiplier Λ that represents the strength of selection and is related with the effective population size. Furthermore, we impose the normalization constraints $\sum_a P^i_a = 1$ at all sites.

Since the parameters that determine the folding free energy are fixed for all proteins (see below), the only free parameters of the model are Λ and $P^{\rm mut}_a$. The frequencies are generally determined from the observed sequences in the protein of known structures and the other sequences of the protein family, while Λ is determined by maximizing the log-likelihood of the PDB sequence, $\sum_i \log\left(P^i(A^{\rm PDB}_i)\right)$, which yields a well-defined single maximum. The pre-computation of the moments of the contacts makes the computation very fast, it runs in a few minutes even for proteins of several hundreds of amino acids. For further computational details see (Arenas & Bastolla 2015).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

Wild-type model of protein evolution

In the wild-type model (Jimenez, Arenas & Bastolla 2018), we also assume that sites evolve independently. We further assume that the site-specific distribution P_a^i of amino acid a at position i is proportional to the background distribution $P_a^{\rm mut}$ multiplied by the exponential of the logarithmic fitness of the corresponding mutation in which the wild-type amino acid in the PDB A_i^{WT} is substituted by the new amino acid a:

$$P_a^{{\rm WT},i} \propto P_a^{\rm mut} \exp\left(\Lambda \varphi\left({\rm mut}(A_i^{WT} \to a)\right)\right) , \tag{5}$$

The fitness of a sequence is computed as in Eq.(1). The parameter Λ is again determined by maximizing the likelihood of the wild-type sequence, $\sum_i \log\left(P^{{\rm WT},i}(A_i^{\rm WT})\right)$.

Sequence entropy

The sequence entropy at position i measures the variability of this position as

$$S_i = -\sum_{a=1}^{20} P_a^i \log(P_a^i) , \tag{6}$$

where P_a^i is obtained either from the evolutionary model (mean-field or wild-type) or from a MSA or from pooled amino acids at equivalent structural positions with the same number of contacts.

Halpern-Bruno exchangeability matrices

To fully specify the site-specific substitution processes, besides the site-specific frequencies P_a^i we need to compute consistent exchangeability matrices with the Halpern-Bruno formulas (Halpern & Bruno 1998).

Given a site-specific amino acid distribution that reflects selective constraints, the Halpern-Bruno method allows computing the rate matrices of the associated site-specific substitution processes $Q_{ab}^i = E_{ab}^i P_b^i$ that are consistent with the Kimura's fixation probability, Eq.(2), and with a background global (not site-specific) mutation process.

Without loss of generality, we parametrize the rate matrix of the global mutation process as $Q_{ab}^{\rm mut} = E_{ab}^{\rm mut} P_b^{\rm mut}$, where $P_a^{\rm mut}$ is the stationary matrix of the mutation process and $E_{ab}^{\rm mut}$ is its exchangeability matrix. To simplify formulas, here we assume detailed balance, i.e. we assume that $E_{ab}^{\rm mut}$ is a symmetric matrix (this condition can be easily relaxed). We write the rate matrices as $Q_{ab}^i = Q_{ab}^{\rm mut} P_{\rm fix}(f_a^i, f_b^i)$, where f_a^i is the "fitness" of amino acid a at site i. We impose that $P_{\rm fix}$ is the fixation probability Eq.(2). Halpern and Bruno showed that the site-specific fitness can be inferred from the stationary distribution

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

from $P_a^i = P_a^{\text{mut}} \left(f_a^i\right)^N$, yielding the following site-specific substitution process

$$Q_{ab}^i = E_{ab}^i P_b^i \tag{7}$$

$$E_{ab}^i = E_{ab}^{\text{mut}} \left(\frac{\ln(F_b^{\text{sel},i}) - \ln(F_a^{\text{sel},i})}{F_b^{\text{sel},i} - F_a^{\text{sel},i}} \right) \tag{8}$$

$$\text{with} \quad F_a^{\text{sel},i} = \frac{P_a^i}{P_a^{\text{mut}}} \tag{9}$$

The selective factors $F_a^{\text{sel},i}$ quantify how much the site-specific distribution P_a^i deviates from the background distribution P_a^{mut} induced by mutation alone.

It can be immediately seen that the exchangeability matrices E_{ab}^i are symmetric, which implies that detailed balance holds and P_a^i is the stationary distribution.

Evolutionary rates

For neutral substitutions with $F_a^{\text{sel},i} = F_b^{\text{sel},i}$, in particular synonymous substitutions $a = b$, applying l'Hopital's rule we find $E_{ab}^i = E_{ab}^{\text{mut}}/F_b^{\text{sel},i}$ and $Q_{ab}^i = Q_{ab}^{\text{mut}}$, i.e. the rate of synonymous substitutions equals the mutation rate, in agreement with Kimura's theory. If the amino acid b is favoured by selection with respect to amino acid a, $F_b^{\text{sel},i} > F_a^{\text{sel},i}$, then the substitution rate is enhanced with respect to the neutral rate, and it is decreased in the opposite case. Because of detailed balance, the flux in one direction and the other are equal, $R_{ab}^i = P_a^i P_b^i E_{ab}^i = R_{ba}^i$, with

$$R_{ab}^i = \left(P_a^{\text{mut}} P_b^{\text{mut}} E_{ab}^{\text{mut}}\right) F_a^{\text{sel},i} F_b^{\text{sel},i} \frac{\ln(F_b^{\text{sel},i}) - \ln(F_a^{\text{sel},i})}{F_b^{\text{sel},i} - F_a^{\text{sel},i}} \tag{10}$$

In the above equation, the flux is partitioned into a global component that is attributed to the mutation process (superscript mut) and a site-specific component that is attributed to selection (superscript sel), which allows analysing the contributions of mutation and selection separately. The flux is maximal for substitutions ab that have large and almost equal selective factors $F_a^{\text{sel},i} \approx F_b^{\text{sel},i}$ and have large mutational flux $P_a^{\text{mut}} P_b^{\text{mut}} E_{ab}^{\text{mut}}$. The site-specific substitution rates are computed as the weighted average of the substitution rate matrix $Q_{ab} = E_{ab}^i P_b^i$,

$$R^i = \sum_{a \neq b} P_a^i E_{ab}^i P_b^i = \sum_{a \neq b} R_{ab}^i \tag{11}$$

Since the flux between any pair of amino acids a and b decreases when their difference of fitness increases, Halpern and Bruno argued that the substitution rate R^i is higher at position with higher sequence entropy (Halpern & Bruno 1998). However, this expectation is not strictly fulfilled, and in fact we observe that the substitution rate is not a strictly increasing function of sequence entropy (see Fig.1).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

Mutation process

Finally, we have to define the global exchangeability matrix E_{ab}^{mut} that characterizes the mutation process. For this, we consider four types of mutational models. To compare the resulting substitution rates, in all cases we fix the scale of the exchangeability matrix equating the substitution rate under mutation alone, $\sum_{a\neq b} P_a^{\text{mut}} P_b^{\text{mut}} E_{ab}^{\text{mut}} = 1$.

1. In the first model, the global exchangeability matrix is equal to the empirical exchangeability matrix (WAG, Whelan & Goldman 2001; or JTT, Jones Taylor and Thornton 1992), i.e. $E_{ab}^{\text{mut}} \equiv E_{ab}^{\text{emp}}$. We call this model the empirical (emp) exchangeability matrix. Since empirical substitution processes include information both on mutation and selection, we expect that they strongly correlate with the selection process.

2. In the second model, we remove the effect of selection from the empirical substitution model by imposing that for each pair of amino acids, the flux predicted by the global model and averaged over all positions is equal to the empirical flux $P_a^{\text{emp}} P_b^{\text{emp}} E_{ab}^{\text{emp}}$, which is the observational data from which empirical models are deduced:

$$\left(P_a^{\text{mut}} P_b^{\text{mut}} E_{ab}^{\text{flux}}\right) \frac{1}{L} \sum_i F_a^{\text{sel},i} F_b^{\text{sel},i} \frac{\ln(F_b^{\text{sel},i}) - \ln(F_a^{\text{sel},i})}{F_b^{\text{sel},i} - F_a^{\text{sel},i}} = P_a^{\text{emp}} P_b^{\text{emp}} E_{ab}^{\text{emp}} \qquad (12)$$

   where we use more compact matricial notation. We call the corresponding exchangeability matrix E_{ab}^{flux} the flux matrix (flux).This mutation model yields optimal results in phylogenetic inference (Arenas & Bastolla, 2015).

3. Thirdly, we model the mutational process at the nucleotide level, using the genetic code and parameterizing the process through the nucleotide frequencies and the transition-transversion ratio κ. The four free parameters are fixed by imposing that the resulting background distribution P_a^{mut} yields amino acid frequencies as close as possible to those observed in the data, P_a^{obs} (Arenas & Bastolla, 2015), as detailed below. We call the corresponding exchangeability matrix the optimized nucleotide (nuc_opt) matrix.

4. The last model is identical to the nuc_opt model, except that the nucleotide frequencies are not optimized but they are input parameters. In this way, we can vary the average hydrophobicity of the complete model by varying the Thymine content, since hydrophobic amino acids are enriched in the T base at second codon position. We call this model the nuc_var model.

In the nuc models, for any set of nucleotide frequencies and transition-tranversion rate we combine the substitution process at the nucleotide level with a selection process that assigns fitness one to sense codons and fitness zero to stop codons. Detailed balance is fulfilled at the nucleotide level, but it is only approximated at the codon level because

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

of this selection against stop codons, therefore the transition to transversion rate can influence the stationary frequencies and we have to compute the stationary distribution of the 61 sense codons numerically.

More precisely, we model the mutation rate between two codons differing at one position, say the third one ($n_1n_2n_3$ and $n_1n_2n_3'$) as $\mu\kappa(n_3, n_3')f^{\rm nuc}(n_3')S(n_1n_2n_3')$, where μ is a global rate parameter, $\kappa(n_3, n_3')$ is one if n_3, n_3' are related through a transversion and is the transition-tranversion rate otherwise, $f^{\rm nuc}(n_3')$ is the stationary frequency of the new nucleotide and $S(n_1n_2n_3')$ is zero if $n_1n_2n_3'$ is a stop codon, one otherwise. After the frequencies of the 61 sense codons evolve to their equilibrium state, the stationary frequencies of amino acids $P_a^{\rm mut}$ are computed summing over codons and the exchangeability matrix is computed from the equilibrium fluxes between pairs of codons that code for any pair of amino acids. In the nuc_opt model, the score of each set of mutation parameters is computed as the likelihood of the observed number of amino acids, $\sum_a n^{\rm obs}(a)\log(P_a^{\rm mut})$, and the parameters that maximize the likelihood are chosen.

Data and observed substitution rates

We performed our computations on 213 proteins that were examined in a previous study (Echave, Jackson & Wilke 2015). The results were qualitatively identical from one protein to the other.

The observed substitution rates of 213 proteins that we show for comparison were estimated in (Echave, Jackson & Wilke 2015) from the MSA of homologous sequences through the program Rate4Site (Pupko et al. 2002), which builds the phylogenetic tree using a neighbour-joining algorithm (Saitou & Nei 1987) and estimates rates with an empirical Bayesian approach adopting the JTT model of sequence evolution (Jones, Taylor & Thornton 1992). The multiple sequence alignments were generously provided by Julián Echave and are publicly available at the url https://github.com/wilkelab/therm_constraints_rate_variation/.

Modelling stability against unfolded and misfolded states

Finally, for completeness we descibe here how we estimate the folding free energy ΔG of the experimentally known native state of a protein.

For this purpose, we adopt the contact matrix representation of the protein structure, consisting in the following: For each pair of residues at positions i and j along the polypeptidic chain, C_{ij} equals one if the residues are in contact and zero otherwise. We define two residues to be in contact if any pair of their heavy atoms are closer than 4.5Å. Since contacts with $|i-j| \leq 2$ are formed in almost all structures, they do not contribute to the free energy difference between the native and the misfolded ensemble, and we set $C_{ij} = 0$ if $|i-j| \leq 2$. The free energy of a protein in the mesoscopic structure described by C_{ij} is modelled as a sum of contact interactions, $E(C, A) = \sum_{i<j} C_{ij}U(A_i, A_j)$, which depends on the type of amino acids in contact A_i and A_j and on 210 contact interaction parameters $U(a, b)$, for which we adopt the parameters determined in (Bastolla, Vendruscolo &

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

Knapp, 2000).

For simplicity, we neglect the conformational entropy of the folded native state and estimate its free energy as $G_{\rm nat}(C^{\rm nat},A) \approx \sum_{i<j} C_{ij}^{\rm nat} U(A_i,A_j)$. Regarding the unfolded state, we neglect their contact interactions and estimate its free energy as $G_U \approx -TLS_U$, where T is the temperature in units in which $k_B = 1$, L is chain length and S_U is the conformational entropy per residue of an unfolded chain. We compute the free energy of the misfolded state from the partition function of the contact energy $E(C,A)$ over a set of compact contact matrices C of L residues that are obtained from the PDB. In agreement with previous studies (Garel & Orland 1988; Shakhnovich & Gutin 1989; Bryngelson et al. 1995), the resulting free energy is approximately described by the Random Energy Model (REM) (Derrida 1981), with the addition of the third moment of the contact energy (Minning, Porto & Bastolla 2013):

$$\begin{aligned} G_{\rm misf} &\equiv -T \log \left(\sum_C {\rm e}^{-\sum_{i<j} C_{ij} U(A_i,A_j)/T + S(C)} \right) \\ &\approx \langle E \rangle - \frac{\left\langle (E - \langle E \rangle)^2 \right\rangle}{2T} + \frac{\left\langle (E - \langle E \rangle)^3 \right\rangle}{6T^2} - LS_C T \end{aligned} \tag{13}$$

where LS_C is the logarithm of the number of compact contact matrices, $\langle . \rangle$ represents the average over the set of alternative compact contact matrices of L residues. This estimate only holds above the freezing temperature, while the free energy is kept constant below the freezing temperature (Derrida 1981). We assume for simplicity that the conformational entropy, $S(C_{ij})$, is approximately the same for all compact structures including the native one, and it can be neglected for computing free energy differences. The mean values of the energy can be computed from the mean values of the contacts, which are computed at the beginning and tabulated to accelerate the computation: $\langle E \rangle = \sum_{i<j} \langle C_{ij} \rangle U_{ij}$, $\langle (E - \langle E \rangle) \rangle = \sum_{i<j,k<l} \left(\langle C_{ij} C_{kl} \rangle - \langle C_{ij} \rangle \langle C_{kl} \rangle \right) U_{ij} U_{kl}$ with $U_{ij} = U(A_i,A_j)$. We also adopt the approximation that $\langle C_{ij} \rangle$ only depends on $|i-j|$ (Minning, Porto & Bastolla 2013).

Putting together these free energy estimates, we obtain the free energy difference between the native and the non-native states as

$$\Delta G(C^{\rm nat},A) = G_{\rm nat} - kT \log \left({\rm e}^{-G_{\rm misf}/kT} + {\rm e}^{-G_U/kT} \right) , \tag{14}$$

where the free energy of the non-native state is computed as a Boltzmann average, which is essentially equal to $G_{\rm misf}$ when the sequence is hydrophobic ($G_{\rm misf} - G_U/kT \ll -kT$) and is essentially equal to G_U when the sequence is hydrophylic ($G_{\rm misf} - G_U/kT \gg kT$). For neglecting stability against misfolding, we compute $\Delta G = G_{\rm nat}(C^{\rm nat},A) + LS_U$.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

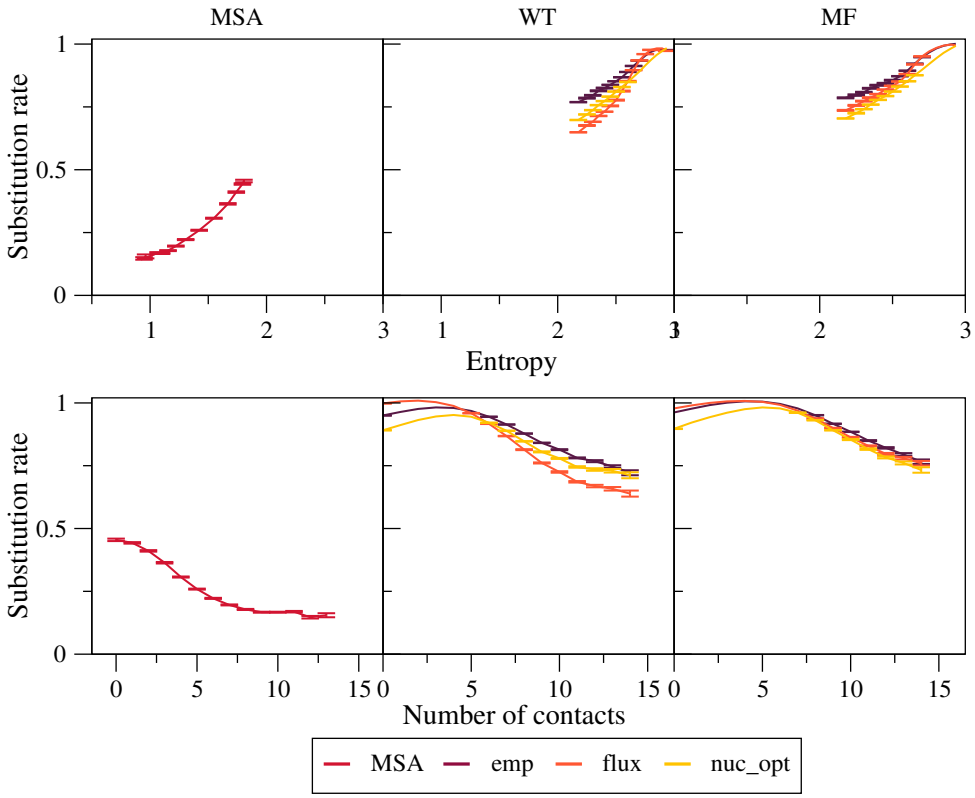

Figure 1: Effect of the exchangeability model on substitution rates. The plots represent substitution rate versus sequence entropy (top) and versus number of native contacts (bottom) for MSA (left), WT model (center) and MF model (right). The models are simulated with the emp, flux and nuc_opt models of the global exchangeability matrix. In all cases the emp model produces the highest substitution rates, consistent with the fact that this model also represents selection.

Results

Dependence of the substitution rate on the global exchangeability model

In this work, we found that the substitution rates depend not only on the selective forces that act specifically at each protein site, but also on the global exchangeability matrix that represents the mutation process.

We considered three models of global exchangeability matrices (see Materials and Methods): (1) Empirical (emp) exchangeability matrices, such as the familiar WAG (Whelan & Goldman 2001) or JTT (Jones, Taylor & Thornton 1992) matrices; (2) Flux exchangeability matrices (flux), which are obtained from empirical exchangeability matrices

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

removing the selective factors represented in the stability-constrained mean-field model, so that the average flux predicted by the model between any pair of amino acids coincides with the observed empirical flux, see Eq.(12); (3) Exchangeability matrices between amino acids obtained from a mutational process at the nucleotide level with parameters optimized by maximizing the likelihood of the observed amino acid composition (nuc_opt); (4) Exchangeability matrices obtained from a mutational process at the nucleotide level with varying parameters, that allows studying the effect of varying hydrophobicity (nuc_var).

We found that empirical exchangeability matrices (emp) produce the larger substitution rates (Fig.1). These matrices take into account both the mutation process and the selection process, since they have been obtained from substitutions that have been fixed through natural selection. From Eq.(10) we can see that the substitution rate is enhanced if the global exchangeability matrix is correlated with the fixation probability. This may explain why empirical exchangeability matrices yield high substitution rates.

The flux exchangeability matrices remove from the empirical exchangeability matrix the effect of natural selection that is represented in the mean-field model. Consistently, we find that the substitution rates determined through the flux model are smaller than those determined with the emp model. We also found in previous work that the flux model yields larger likelihood in phylogenetic inference (Arenas & Bastolla, 2015). Because of this, the flux model is our default exchangeability model.

Fig. 1 shows that the nuc_opt model with mutations at the nucleotide level and optimized parameters produces lower substitution rates than the flux model when associated with the MF model of selective constraints, which suggests that the flux model may still represent some selection. However, when the WT model of selection is applied, the nuc_opt model again produces lower substitution rates than the flux model for exposed sites with few contacts and high entropy, but the flux model produces lower substitution rates for buried sites with many contacts and low entropy. Note that the WT model represents stronger selective constraints than the MF model, since it generally predicts lower sequence entropies and substitution rates. Thus, these results suggest that the flux model associated with the WT model is effective in removing selective constraints for sites with many contacts, but less effective for sites with few contacts.

Substitution rates are different for hydrophobic and hydrophylic sites with the same entropy

Next, we investigated more in detail the relationship between site-specific sequence entropies and substitution rates. These quantities are in general well correlated as predicted by Halpern and Bruno, as one can see in Fig.1 top plots that show that sites with larger entropy tend to have on the average larger substitution rates. Nevertheless, in Fig.2 one can see that the detailed plot of the substitution rate versus the sequence entropy of all sites presents two branches that correspond to different numbers of native contacts. For equal sequence entropy, polar sites with few contacts evolve faster than hydrophobic sites with many contacts. This happens for both models of the mutation process that do not

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

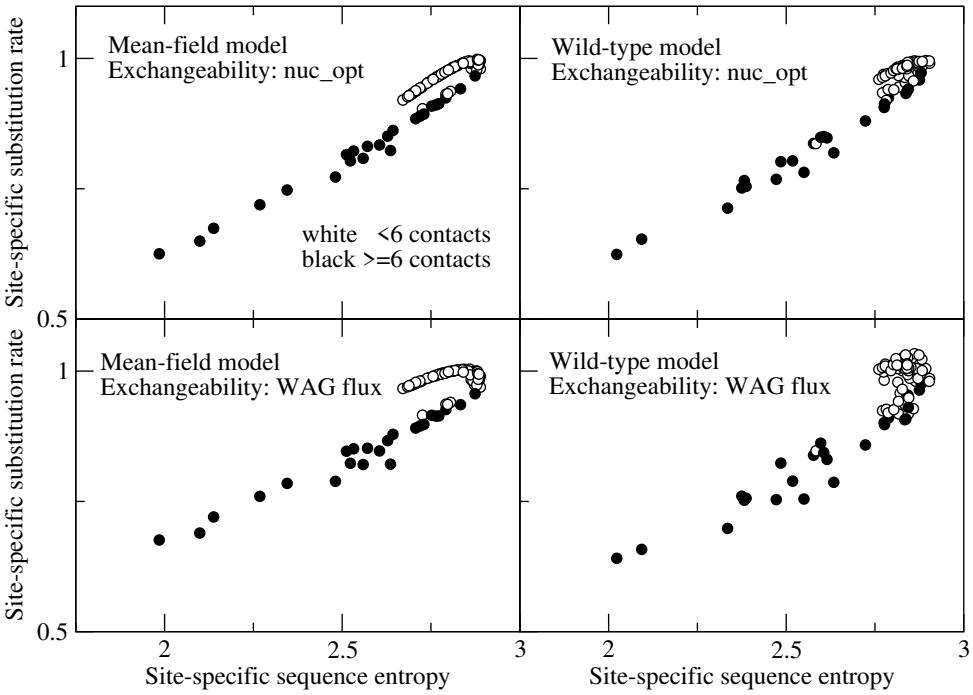

Figure 2: Each point represents sequence entropy and substitution rate for a site of the ribonuclease with PDB code 1pyl, which is representative of all of our data set. We show data for this protein since its small size makes the figure easier to interpret. One can spot two branches, corresponding to sites that evolve faster and slower for the same sequence entropy. The two branches correspond to polar sites with few contacts (white circles) and hydrophobic sites with many contact (black circles), respectively.

take into account natural selection, both flux and nuc_opt, and both for the MF and WT model of natural selection.

Since sequence entropy is a measure of the selective constraints, this difference between sites with equal sequence entropy should be attributed to the mutation process embodied in the exchangeability matrix, not to natural selection. Since sites with few contacts evolve faster and they have lower hydrophobicity, this relationship points at an inverse relationship between mutational fluxes and hydrophobicity.

More hydrophobic proteins substitute more slowly, but mutation bias towards hydrophobicity increases the substitution rates

After investigating the relationship between hydrophobicity and substitution rates for individual sites, we perform the same analysis between different proteins. For this purpose, we group the 213 proteins in our data set according to their predicted average hydrophobicity under the same mutational process and compare the substitution rates of groups characterized by different hydophobicity. In Fig.3, each point represents a group

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

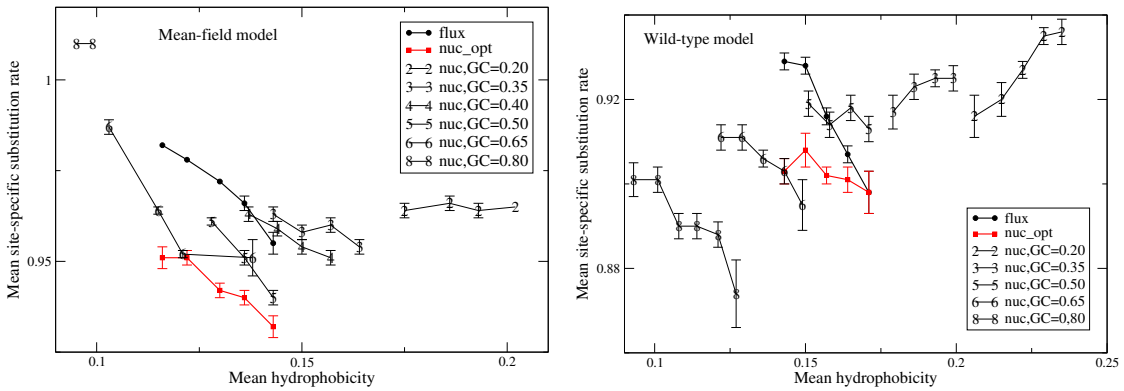

Figure 3: In this plot each point represent a group of proteins with similar mean hydrophobicity in the evolutionary model, and each curve is obtained by varying the background distribution and the exchangeability matrix, which represent the mutation process. One can see that, for the same mutation process, more hydrophobic proteins tend to evolve more slowly, except when the mutation process induces very high hydrophobicity, in which case the substitution rate becomes an increasing function of hydrophobicity. On the other hand, mutation processes with extreme properties (very high or very low hydrophobicity) tend to increase the substitution rate.

of proteins with similar mean hydrophobicity. Each curve is obtained for a mutation bias with different G+C content (nuc_var model), which produces a different background distribution P^{mut} and exchangeability matrix E^{mut}. Since Thymine at second codon position almost always codes for hydrophobic amino acids, there is a negative correlation between G+C content of the mutation model and the average hydrophobicity of the protein sequence. Varying the mutation bias we construct different sets of model proteins that present varying hydrophobicity and are characterized by different background amino acid frequencies and exchangeability matrices. In this way, we can investigate how the mutation bias influcence the biophysical properties (hydrophobicity) and the evolutionary properties (substitution rate, sequence entropy) of an evolving protein.

One can see that, for the same mutation process (connected points in the plot), more hydrophobic proteins tend to evolve more slowly, consistent with what we observed in Fig.1. Nevertheless, when the mutation process induces very high hydrophobicity, the substitution rate becomes an increasing function of hydrophobicity. This is easily rationalized by the fact that, when the background distribution P_a^{mut} is biased towards hydrophobic amino acids, the mutational flux $P_a^{\mathrm{mut}} P_b^{\mathrm{mut}} E_{ab}^{\mathrm{mut}}$ is higher between pairs of hydrophobic residues.

On the other hand, mutational processes with extreme bias (very high or very low G+C content and hydrophobicity) tend to increase the substitution rate.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

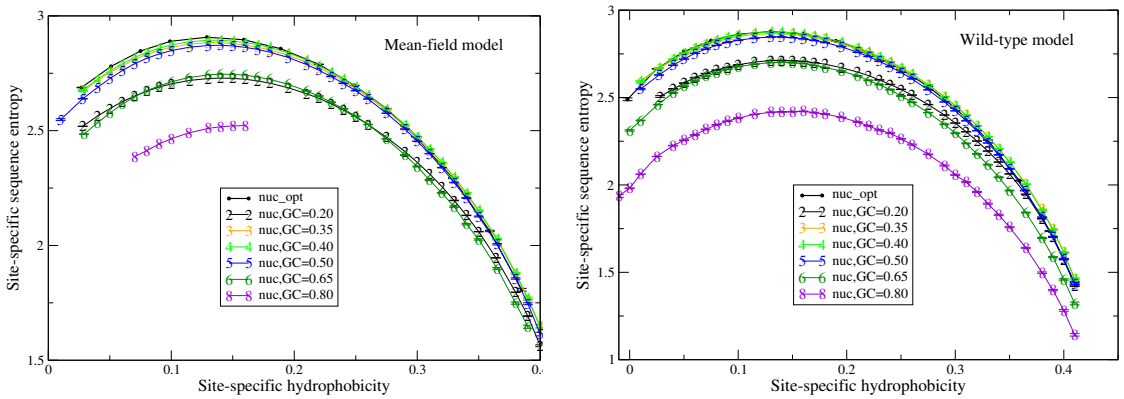

Figure 4: Each point represents a set of protein sites with similar hydrophobicity in the evolutionary model. The sequence entropy has a universal shape as a function of hydrophobicity, with a maximum when the hydrophobicity is approximately 0.14, which is the mean hydrophobicity of the equiprobable distribution of amino acids. Changes in the background distribution mostly shift the sequence entropy curves without changing the position of the maximum, but they affect the values of entropy. The largest entropies are obtained for the mutation model optimized for each protein sequence (thick black line) and for the mutation bias with GC content equal to 0.40, which yield only slightly hydrophobic sequences.

Influence of the mutation bias on sequence entropies

We next study how the shape of the entropy-hydrophobicity curve depends on the mutation bias. In Fig.4 each point represents a set of protein sites with similar hydrophobicity in the stability-constrained evolutionary model. The sequence entropy has an almost universal shape as a function of hydrophobicity, with a maximum when the hydrophobicity is approximately 0.14. This is the mean hydrophobicity of the equiprobable distribution of amino acids. Changes in the background distribution mostly shift the sequence entropy curves in the vertical direction, but they do not change the position of the maximum.

In contrast, the values of sequence entropy change systematically with the mutation bias. The largest entropies are obtained for the mutation model nuc_opt optimized separately from each PDB sequence (thick black line) and for the mutation bias with G+C content equal to 0.40, which has a small bias towards slightly hydrophobic sequences. Extreme mutation bias yield very reduced sequence entropies, which means that the selective contraints impose stronger constraints in order to preserve the average hydrophobicity needed for stable proteins. This result is consistent with the finding that, for equal population size, the average fitness achieved in evolution has a maximum as a function of the mutation bias, and it is low for extreme mutation bias either toward hydrophobic or towards hydrophylic sequences (Mendez et al 2010; Bastolla, Dehouck & Echave, 2017).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

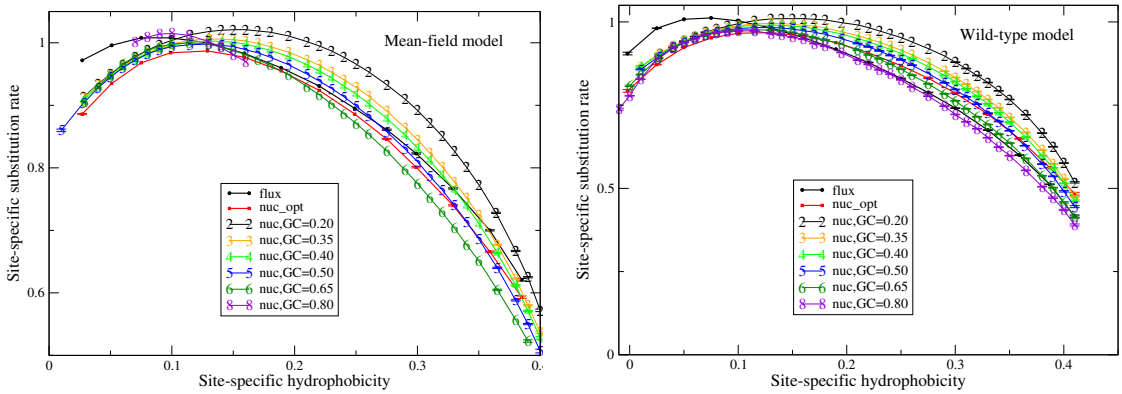

Figure 5: Each point represents a set of protein sites with similar hydrophobicity in the evolutionary model. The substitution rate shows a maximum whose position depends on the mutation process. The hydropobicity at which the maximum rate is achieved increases with the mean hydrophobicity of the mutation process (lower GC content). The substitution rates tend to increase for mutation processes that yield higher hydrophobicity (lower GC content), but for the MF model they also increase for extremely polar mutation bias (GC content 0.8).

Influence of the mutation bias on substitution rates

We now study the relationship between site-specific hydrophobicity and site-specific substitution rate. As in the previous figure, also in Fig.5 each point represents a set of protein sites with similar hydrophobicity in the stability-constrained evolutionary model, and we plot the substitution rate versus the hydrophobicity. Different from the shape of the sequence entropy, the shape of the substitution rate curve clearly depends on the mutation bias. The hydropobicity at which the maximum substitution rate is achieved decreases with the G+C content or, equivalently, it increases with the mean hydrophobicity of the mutation process, as expected. In other words, when the mutation process favours more hydrophobic amino acids, the maximum of the substitution rate is achieved at sites that are more hydrophobic. This result confirms that the mutation process has a strong influence on the substitution rates.

Consistent with Fig.3, the substitution rate at the maximum tends to increase for mutation processes that favour higher hydrophobicity (lower G+C bias), but for the MF model they also increase for extremely polar mutation bias (G+C content 0.8). As seen in Fig.1, the flux model of the exchangeability matrix (thick black line) predicts higher substitution rates than the nuc_opt model (red) when applied together with the MF model, but when it is applied together with the WT model it predicts lower substitution rates at hydrophobic sites with many contacts.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

Discussion and conclusions

Here we studied how the predicted evolutionary variabilities of different protein sites of a protein are influenced by the underlying mutation process, according to a model of stability-constrained protein evolution with selection on the stability of the native state against both unfolding and misfolding.

We found that the sequence entropy and the substitution rate are not equivalent measures of the evolutionary variability of the protein sites. These measures tend to be correlated as expected (Halpern & Bruno 1998), because the substitution rate tends to increase for sites with higher sequence entropy (Fig.1). However, sites with the same sequence entropy are characterized by different substitution rates, which are systematically higher for polar sites than for hydrophobic sites (Fig.2). This difference is not due to different selective constraints, which are quantified by sequence entropy, but it is due to the different exchangeability of polar and hydrophobic amino acids, which influences the substitution rates but not the sequence entropy. The result robust with respect to changes of the selection model (WT or MF) and the mutation model (the flux between amino acids observed in empirical models or a codon model with optimized nucleotide frequencies). As a result, more polar proteins are predicted to evolve faster than proteins with large mean hydrophobicity (Fig.3).

These results hold when the exchangeability matrix, which represents the mutational process, is kept constant. When different mutation bias are simulated and compared, the substitution rates tend to be larger for mutation bias favoring hydrophobic residues (low G+C), and also for mutation bias favoring very polar amino acids (high G+C), but the latter happens only when the MF model of selection is applied (Fig.3). Thus, the comparison of proteins with different hydrophobicity under the optimal mutation model nuc_opt and the comparison between different mutation models (for instance, achieved in different organisms) yield contrasting results as far as the substitution rate is concerned: substitution rates tend to be higher for more polar proteins evolving under the same mutation process, but they tend to be higher in organisms with mutation bias towards A+T that favour hydrophobic residues. While the former result is attributable to the higher exchangeability of polar residues, thus to the mutational process, as we argued above, the latter result is likely caused both both by the mutational process and by selection.

In fact, changing the mutation bias severely affects the selective constraints imposed on the protein sites (Fig.4). The shape of the curve of the sequence entropy versus the hydrophobicity does not depend on the mutation bias, showing a maximum when the hydrophobicity is equal to the mean unweighted hydrophobicity of the 20 amino acids, which corresponds to an equiprobable distribution of amino acids. In contrast, the values of entropies strongly decrease when the mutation bias becomes extreme in both the hydrophobic or polar direction, and they are large when the mutation bias is G+C=0.40, corresponding to slightly hydrophobic residues. The sequence entropy of polar sites decreases more with the mutation bias than the entropy of hydrophobic sites.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

Thus, polar sites are affected by stronger selective constraints when the mutation bias is extreme.

The substitution rate is systematically influenced by the mutation bias (Fig.5), in such a way that when the mutation bias favours more hydophobic proteins (low G+C) the maximum of the substitution rate is achieved at sites that are more hydrophobic, confirming that the mutation bias has a direct influence on the substitution rates, and the substitution rate at the maximum tends to increase for mutation processes that favour higher hydrophobicity, confirming the prediction that low G+C mutation bias that favour hydrophobic sequences enhance the substitution rates.

Finally, we note that, under extreme mutation bias, the site-specific hydrophobicity at which the sequence entropy is maximal does not coincide with the hydrophobicity at which the substitution rate is maximum, producing larger discrepancies between the two measures of evolutionary variability.

Acknowledgements

We thank Julián Echave for providing us the empirical multiple sequence alignments that were used for the leftmost plots in Fig.1, and we acknowledge interesting discussions with him, Alberto Pascual-García and Nick Goldman. This work was supported by the Spanish Ministery of Economy through grants BIO2016-79043 and BFU2012-40020. MA was supported by the Ramón y Cajal Grant RYC-2015-18241 from the Spanish Government. Research at CBMSO is facilitated by Fundación Ramón Areces.

References

[1] Arenas M., Sanchez-Cobos A. and Bastolla, U. (2015) Maximum likelihood phylogenetic inference with selection on protein folding stability. Mol Biol. Evol. 32:2195-207.

[2] Arenas M., Weber C.C., Liberles D.A. and Bastolla, U. (2017) ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. Syst Biol. 66:1054-1064.

[3] Bastolla U., Porto, M., Roman, H.E.and Vendruscolo, M. 2006. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank BMC Evol. Biol. 6:43.

[4] Bastolla U, Dehouck Y, Echave J (2017) What evolution tells us about protein physics, and protein physics tells us about evolution. Curr Opin Struct Biol. 42:59-66.

[5] Bastolla, U.; Vendruscolo, M.; Knapp, E.W. (2000) A statistical mechanical method to optimize energy functions for protein folding. Proc. Natl. Acad. Sci. USA 97, 3977-3981.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

[6] Berezovsky IN, Zeldovich KB, Shakhnovich EI (2007) Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput Biol 3:e52.

[7] Bryngelson, J.-D.; Onuchic, J.-N.; Socci, N.-D.; Wolynes, P.-G (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins 21, 167-195.

[8] Derrida, B. (1981) Random Energy Model: An exactly solvable model of disordered systems. Phys. Rev. B 24, 2613.

[9] Echave J. (2008) Evolutionary divergence of protein structure: The linearly forced elastic network model. Chem Phys Lett 457, 413-416.

[10] Echave J., Jackson E.L. and Wilke C.O. (2015) Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. Phys Biol. 12:025002.

[11] Echave J, Spielman SJ, Wilke CO (2016) Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 17:109-21.

[12] Franzosa, E. A., and Y. Xia, 2009 Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26: 2387-2395.

[13] Ohta T (1976) Role of very slightly deleterious mutations in molecular evolution and polymorphism, *Theor. Pop. Biol.* **10**, 254-275.

[14] Garel, T.; Orland, H. (1988) Mean-field model for Protein Folding, Europhys Lett, 307-310.

[15] Goldstein RA (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins 79:1396-1407.

[16] Halpern A and Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. 15(7):910-917.

[17] Huang TT, del Valle Marcos ML, Hwang JK, Echave J (2014) A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol Biol. 9;14:78.

[18] Jimenez MJ, Arenas M, Bastolla U (2018) Substitution rates predicted by stability-constrained models of protein evolution are not consistent with empirical data. Mol Biol Evol. 35, 743755.

[19] Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-82.

[20] Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 4:713-719.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018

[21] Mendez R, Fritsche M, Porto M, Bastolla U (2010= Mutation bias favors protein folding stability in the evolution of small populations. PLoS Comput Biol 6:e1000767.

[22] Minning J, Porto M, Bastolla U (2013) Detecting selection for negative design in proteins through an improved model of the misfolded state. Proteins 81:1102-1112.

[23] Mustonen V, Lässig M (2005). Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc Natl Acad Sci USA 102:15936-41.

[24] Noivirt-Brik O, Horovitz A, Unger R (2009) Trade-off between positive and negative design of protein stability: from lattice models to real proteins. PLoS Comput Biol 5:e1000592.

[25] Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics. 18 Suppl 1:S71-7.

[26] Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-25.

[27] Scherrer MP, Meyer AG, Wilke CO (2012) Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol. 12:179.

[28] Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 102:9541-6.

[29] Serohijos AW, Shakhnovich EI (2014) Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Curr Opin Struct Biol 2014, 26:84-91.

[30] Shakhnovich, E.-I.; Gutin, A.-M. (1989) Formation of unique structure in polypeptide chains, Biophys. Chem. 34, 187.

[31] Tirion MM (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett. 77:1905-1908.

[32] Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-9.

[33] Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. Mol Biol Evol. 2014 31:135-9.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26868v1 | CC BY 4.0 Open Access | rec: 18 Apr 2018, publ: 18 Apr 2018