Structure and stability constrained substitution models outperform traditional substitution models used for evolutionary inference

Ivan Lorca[1], Miguel Arenas[2,3] and Ugo Bastolla[1]
1 Centro de Biologia Molecular "Severo Ochoa", CSIC-UAM Cantoblanco, 28049 Madrid, Spain.
2 CINBIO, University of Vigo, 36310 Vigo, Spain.
3 Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain.
E-mail: ubastolla@cbm.csic.es

September 7, 2022

Abstract

The current knowledge about how protein structures influence sequence evolution is rarely incorporated into substitution models adopted for phylogenetic inference, which are commonly based on independent with the same substitution process and ignore the known variation of the evolutionary rates across sites with different structural properties. In previous works, we presented site-specific substitution models of protein evolution based on selection on the folding stability of the native state (Stab-CPE), which predict more realistically the evolutionary variability across protein sites. However, those Stab-CPE present qualitative differences from observed data, probably because they ignore changes in the native structure, despite empirical studies suggest that conservation of the native structure is a strong selective force.

Here we present novel structurally constrained substitution models (Str-CPE) based on Julián Echave's model of the structural change due to a mutation as the linear response of the protein to a perturbation and on the explicit model of the perturbation generated by a specific amino-acid mutation. Compared to our previous Stab-CPE models, the novel Str-CPE models are more stringent (they predict lower sequence entropy and substitution rate), provide higher likelihood to multiple sequence alignments (MSA) of the wild-type protein, and better predict the observed substitution rates. Next, we combine Str-CPE and Stab-CPE models to obtain structure and stability constrained substitution models (SSCPE) that fit the empirical MSAs even better. Importantly, these SSCPE models present a relevant improvement of the phylogenetic likelihood for all ten protein families that we analyzed with the program RAxML-NG. We implemented the SSCPE models in the program Prot_evol, freely available at https://github.com/ugobas/Prot_evol.

Substitution models of protein evolution are a key element of probabilistic phylogenetic inference. However, they necessarily trade-off realism for computational treatability. In particular, traditional empirical substitution models assume that protein sites evolve independently of each other and under the same substitution process, although these assumptions are clearly violated by the physical interactions among amino acids that underlie protein structural integrity and functional dynamics (Liberles et al. 2012, Wilke 2012, Sikosek and Chan 2014). In particular, the substitution rate has been shown to vary substantially across protein sites, with sites that are exposed in the native state varying much more rapidly than buried sites (Franzosa and Xia 2009, Echave et al. 2016). The failure to take into account this variability of the substitution rate among

sites may affect the accuracy of branch lengths estimated through substitution models (Yang, 1996; Buckley et al. 2001), which are essential ingredients of phylogenetic inference both in the context of distance-based methods such as Neighbor Joining (Saitou and Nei 1987) and of probabilistic methods such as maximum likelihood (ML) (Felsenstein 1981) and Bayesian (Yang et al. 1997). As an approximation, Yang (1993, 1996) proposed to adopt a gamma distribution (+G) to model among-site rate variation in phylogenetics, but still this could be insufficient to fit the observed variation, for instance if the distribution of buried and exposed sites does not follow a gamma distribution or if the approximations for computing the average over the gamma distribution are problematic.

Two classes of models have been proposed to explain why the site-specific substitution rate depends on the structural properties of the site in the native state.

(1) Stability constrained models of protein evolution (Stab-CPE) represent fitness as the probability to find the protein in the native state (Goldstein 2011, Serohijos and Shakhnovich 2014, Bastolla et al. 2017), which can be estimated from the predicted free energy difference ΔG between the native and non-native states of the protein. For computational simplicity, these models assume that only the folding free energy changes upon mutation, while the average structure of the native state remains fixed.

(2) Structurally constrained models of protein evolution (Str-CPE) consider fitness as a decreasing function of the deviation of the average structure of the native state from a reference structure available in the Protein Data Bank (PDB) (Echave 2008), and predict this deviation through linear response theory applied to the Elastic Network Model (Tirion 1996, Atilgan et al. 2001) (ENM) that approximates the native ensemble of the protein, while they assume that the folding free energy remains constant. Although Str-CPE models were originally referenced as stability constrained models, we recommend that it is important to differentiate their name from strict Stab-CPE because they are based on different assumptions.

Importantly, Stab-CPE models can provide site-specific amino acid exchangeability matrices that can be used in phylogenetic inference, while retaining high computational efficiency due to the approximation that different sites evolve independently (Bastolla et al. 2006, Arenas et al. 2015). These Stab-CPE substitution processes adopt the Halpern-Bruno approach (Halpern and Bruno 1998) for associating to each site-specific vector of amino acid frequencies a substitution matrix that mathematically represents the combined effect of mutation and the Kimura's fixation probability (Kimura 1962). The standard Halpern-Bruno approach obtains the site-specific frequencies from empirical data, resulting in a large number of fitted parameters that entail a strong risk of overfitting. The Stab-CPE Halpern-Bruno approach that we developed, in contrast, deduces the site-specific amino acid frequencies based on simplified models of the stability of the native state against both unfolded and misfolded states. In this approach there is only one global fitted parameter that represents the strength of natural selection and includes global amino acid frequency parameters as in the +F empirical substitution models of protein evolution. These Stab-CPE models assigned much higher likelihoods to real multiple sequence alignments than empirical substitution models that are the same at all positions, indicating that they represent selection on protein structures more realistically (Arenas et al. 2015, Arenas et al. 2019),

Computer simulations based on Stab-CPE and Str-CPE models showed that both models predict that the sites that form more native contacts, which are more buried in the native state, are more constrained by negative selection and tend to evolve more slowly, which is consistent with observations from empirical data (Echave et al. 2015). This occurs because mutations at sites with

many contacts strongly affect both protein folding stability and the overall protein structure of the native state. However, the Stab-CPE model adopted in these simulations did not consider the stability of the native state against incorrectly folded (misfolded) conformations. Previous studies showed that stability against misfolding is an important selective force in protein evolution, whose footprint can be identified in natural protein structures (Berezovsky et al. 2007, Noivirt-Brik et al. 2009, Minning et al. 2013). Consistently, Stab-CPE models that neglect stability against misfolding predict protein sequences that are on the average more hydrophobic than natural protein sequences (since hydrophobicity stabilizes compact conformations belonging to both the native and misfolded states) and assign lower likelihood to real protein sequences than models that consider stability against misfolding (Arenas et al. 2015). The Stab-CPE models that consider misfolding, which are better supported by empirical data, predict that the protein sites that are more variable and evolve more rapidly are amphiphilic sites with an intermediate number of contacts, which can host both hydrophobic and polar residues (Jimenez et al. 2018). This prediction contrasts with the observation that the most variable sites are exposed sites, and supports Str-CPE against Stab-CPE.

Str-CPE are also supported by studies that suggest that protein structure is more conserved than protein sequence when the protein function is conserved (Illergard et al. 2009, Pascual-Garcia et al. 2010). This implies that mutations that can modify protein stability but do not modify the structure are more likely tolerated, which suggests that negative selection targets protein structure more strongly than protein sequence. Conversely, pairs of proteins with different molecular function show variations of the substitution rate attributable to positive selection. These evolutionary accelerations are stronger for protein structure changes than protein sequence changes, again suggesting that positive natural selection targets protein structure more strongly than protein sequence.

The success of elastic network models in predicting the native dynamics of proteins solely based on the average structure of the native state (Tirion 1996, Atilgan et al. 2001) provides additional support to Str-CPE, since it suggests that the functional dynamics of proteins, which is an important factor for protein fitness, is largely determined by their native structure, underlying its importance for selection.

All these arguments concur in supporting Str-CPE. Nevertheless, so far no Str-CPE model can generate site-specific substitution models for phylogenetic inference, essentially due to the lack of an explicit model that predicts how any specific amino acid change at a given protein site modifies the native structure.

Here we develop such a Str-CPE model for predicting site-specific substitution matrices. We found that this model is more restrictive than our previous Stab-CPE models, since it generates substitution processes with lower sequence entropy and lower substitution rates. In addition, the new Str-CPE model correctly predicts that completely exposed sites are more variable and evolve faster than partially exposed sites, and it outperforms our previous Stab-CPE models in terms of the likelihood of multiple sequence alignment (MSA) that include the protein of known structure. Finally, when applied to maximum-likelihood phylogenetic inference with the program RAxML-NG (Kozlov et al. 2019), the Str-CPE models yield higher likelihood than both empirical models with rate variation represented through the gamma distribution and Stab-CPE models.

By combining the Str-CPE model with our previous Stab-CPE model, we obtain an improved model characterized by lower tolerance to mutation, higher protein stability, higher likelihood of real MSA and higher phylogenetic likelihood, which we named SSCPE model (structure and stability constrained protein evolution model). All these findings suggest that SSCPE models represent selection on protein sequences and structures more realistically than both empirical substitution

models and our previous Stab-CPE models.

The program Prot_evol, which computes the site-specific substitution matrices for phylogenetic inference using as input a protein structure, its MSA, and the output of the program TNM, is freely available at https://github.com/ugobas/Prot_evol and it runs from seconds to minutes. The program tnm is freely available at https://github.com/ugobas/tnm and it runs in minutes.

Materials & Methods

Model of structure change upon mutation

To model the change of protein structure produced by a mutation, we adopt the linearly forced elastic network model (LfENM) proposed by Julián Echave (2008) In this model, a mutation is represented as a force $\vec{F}$ that perturbs the native structure of the wild-type protein and the structural change of the mutation is predicted as the linear response of the elastic network model that represents the native structure:

$$\Delta r_i = \sum_j H_{ij}^{-1} \vec{F}_j^{(\mathrm{mut})} \tag{1}$$

where H^{-1} is the inverse of the Hessian matrix that describes the internal forces in the elastic network model and $\vec{F}^{(\mathrm{mut})}$ is the force (perturbation) generated by the specific mutation. This force is modelled as the sum of components directed along the contacts that join the mutated residue m to its neighbours in the native state. These components were randomly drawn in the original LfENM approach. In contrast, here we model each component as the combined effect of the changes of amino acid size, contact stability and optimal contact distance produced by the mutation, in such a way that the mutation force depends on the amino-acids that are interchanged.

Furthermore, the original LfENM approach adopted an ENM that only moves the $3L$ Cartesian coordinates of the alpha carbons, while here we adopt the torsional network model (TNM, Mendez and Bastolla 2010) that uses as degrees of freedom the torsion angles of each residue and constraints all other internal coordinates to their native values. This approach can move all heavy atoms of the protein using one, two, three or more degrees of freedom per residue if only phi and psi, phi, psi and chi or also side-chain angles are allowed to move, respectively. In this work, we adopt the version that adopts phi and psi, but other variants can be readily run.

Importantly, the TNM applies a torsional regularization as a quadratic potential that constraints the fluctuations of the free torsion angles as $\kappa_\varphi K \sum_a (\Delta\varphi_a)^2$, where a labels the torsion angles and K is an overall force constant that determines the length scale of the thermal fluctuations. This regularization is essential for limiting fluctuations that may violate the harmonic approximation on which all ENM approaches are based, and it is particularly important when more torsion angles are moved, increasing the risk to violate the harmonic approximation (Dehouck and Bastolla 2021). Here we adopt $\kappa_\varphi = 0.2$, a value that produced good results in previous studies.

We can express the structure change Eq.(1) in the base of the normal modes as

$$\Delta r_i = \sum_\alpha \frac{v_i^\alpha}{\omega_\alpha^2} \sum_j v_j^\alpha F_j^{\mathrm{mut}} \tag{2}$$

where i and j label the Cartesian components of all heavy atoms (not just alpha carbons), α labels the normal mode, ω_α is the frequency of mode α, which is inversely proportional to the amplitude

of the thermal fluctuations of the normal mode (i.e. lower frequency normal modes contribute more strongly to thermal fluctuations), v_j^α is the Cartesian component j of normal mode α (Mendez and Bastolla 2010), and we use the property of the normal modes $Hv^\alpha = M\omega_\alpha^2 v^\alpha$. In the following, we denote by $F_\alpha^{\rm mut} = \sum_j v_j^\alpha F_j^{\rm mut}$ the projection of the mutation force on normal mode α.

Model of the perturbation induced by a mutation

All ENMs adopt a structure-based, minimally frustrated energy function where all pairs of residues i and j that form a contact in the input native structure are assumed to be at the distance that minimizes the energy. In the TNM a contact, designed as $C_{ij} = 1$, exists if the two closest heavy atoms belonging to residues i and j, indicated as i' and j', are closer than 4.5Å, and the vector joining the two residues is represented by the vector joining the two closest atoms, $\vec{r}_{ij} = \vec{r}_{i'} - \vec{r}_{j'}$.

We indicate each specific mutation at a protein site with the 3 indexes (amb), meaning that amino acid a (e.g. Leucine) at the mutated site m changes to amino acid b (e.g. Alanine). There are $19L$ mutations from the wild-type sequence. Each specific mutation generates a perturbation at each residue k in contact with the mutated site m, i.e. $C_{mk} = 1$. We indicate with A_k the amino acid at site k in the wild-type sequence. We consider three types of changes generated by these mutations: (a) Change of the amino acid size $s(a) \to s(b)$. (b) Change of the stability of the contact, which is expressed through the contact free energy matrix U, $U(a, A_k) \to U(b, A_k)$. (c) Change of the optimal (i.e. most frequent) equilibrium distance d between amino acids, $d(a, A_k) \to d(b, A_k)$. Each of these changes generates a mutation force that is modelled as

$$\begin{aligned} F_k^{(amb,\rm size)} &= C_{mk} \frac{(\vec{r}_{m'} - \vec{r}_{k'})}{|\vec{r}_{m'} - \vec{r}_{k'}|^{1+g}} \, (s(b) - s(a)), \\ F_k^{(amb,\rm stab)} &= C_{mk} \frac{(\vec{r}_{m'} - \vec{r}_{k'})}{|\vec{r}_{m'} - \vec{r}_{k'}|^{1+g}} \, (U(b, A_k) - U(a, A_k)), \\ F_k^{(amb,\rm dist)} &= C_{mk} \frac{(\vec{r}_{m'} - \vec{r}_{k'})}{|\vec{r}_{m'} - \vec{r}_{k'}|^{1+g}} \, (d(b, A_k) - d(a, A_k)). \end{aligned} \tag{3}$$

The components of the force are such that they tend to repel the two residues in contact if the mutated amino acid b is larger than the wild-type a, or the interaction energy $U(b, A_k) > U(a, A_k)$ (meaning that the contact is destabilized), or the optimal distance increases, $d(b, A_k) > d(a, A_k)$. In this work we adopt the exponent $g = 2$, with the meaning that the mutation force is stronger at closer contacts. As in (Echave, 2008), the force at the mutated residue m is the reaction force of all of its contacts, $F_m^{(amb)} = -\sum_k F_k^{(amb)}$.

The parameters that appear in Eq.(3) are derived from protein statistics. $s(a)$ is given by the number of atoms of each amino acid a, $U(a, b)$ is the contact energy function derived in (Bastolla et al. 2000) and adopted for computing stability changes upon mutation in our previous Stab-CPE model (Bastolla et al. 2006, Arenas et al. 2015), and the distances $d(a, b)$ are derived from a non-redundant subset of the PDB. All these parameters are available as Supplementary material.

For predicting the structural effect of mutations, we optimized three coefficients that weight each type of contribution:

$$\vec{F}^{(amb)} = K R_0^g \left[W_{\rm size} \vec{F}^{(amb,\rm size)} + W_{\rm stab} \vec{F}^{(amb,\rm stab)} + W_{\rm dist} \vec{F}^{(amb,\rm dist)} \right]. \tag{4}$$

The multiplicative factor K (overall force constant) provides that the extent of the structural change does not vary when the force constant K is increased, which inversely decreases the fluctuations predicted by the TNM model. The factor R_0^g with $R_0 = 3.25$Å provides that the average extent of the structural change is almost independent of the exponent g.

The main difficulty for assessing and optimizing the mutation model is that the typical RMSD between the X-ray structures of the wild-type and of a mutant with only one amino acid change is of the same order as the RMSD between different X-ray structures with exactly the same sequence (Carpentier et al. 2019), and we do not know which part of the structural change is due to the mutation and which part is due to slightly different experimental conditions.

Here we take advantage of the fact that linear response theory also allows predicting the structural change in the absence of mutations. We recently proposed a null model of the random conformation changes produced by any change of experimental condition and that does not present any correlation with the native dynamics of the protein (Dos Santos et al. 2013). This null model is also based on linear response theory, and it assumes that each normal mode contributes to the conformation change as much as it contributes to thermal dynamics. In other words, the squared projection of the conformation change on normal mode α is proportional to the squared amplitude $1/\omega_\alpha^2$. We verified this null model on a large set of proteins with same sequence and different structures (Dos Santos et al. 2013).

From Eq.(2), we can see that the null model requires a perturbation $\vec{F}$ whose projection on each normal mode α is $F_\alpha \equiv \sum_j v_j^\alpha F_j^{\mathrm{NULL}} = r_\alpha \omega_\alpha$, with r_α independent of the normal mode amplitude, for instance Gaussianly distributed. Echave (2008) found that also the force F^{mut} that models the mutation has projections on normal modes that scale with ω_α. This is so because the mutation propagates along the network of native contacts with a structure similar to that of the Hessian matrix H of the ENM. We may invoke a similar argument for justifying our null model. The final model for the observed mean square deviation (MSD) is

$$\mathrm{MSD} = \frac{1}{M}\sum_i m_i \Delta r_i^2 = \mathrm{MSD}^{\mathrm{mut}} + \mathrm{MSD}^{\mathrm{nomut}} + \mathrm{MSD}^{\mathrm{cross}} \tag{5}$$

$$\mathrm{MSD}^{\mathrm{mut}} = \sum_\alpha \left(\frac{F_\alpha^{\mathrm{mut}}}{\omega_\alpha^2}\right)^2, \ \mathrm{MSD}^{\mathrm{nomut}} = \sum_\alpha \frac{1}{\omega_\alpha^2}, \ \mathrm{MSD}^{\mathrm{cross}} = 2\sigma_p \sqrt{\sum_\alpha \left(\frac{F_\alpha^{\mathrm{mut}}}{\omega_\alpha^3}\right)^2}, \tag{6}$$

where i labels the Cartesian coordinates of the protein atoms, m_i is their mass, M is the total mass, α labels the normal modes, F_α^{mut} are the projections of the mutation force on the normal modes, $\sigma_p = \pm 1$ is a binary random number, and we used the orthonormality conditions $\sum_i m_i v_i^\alpha v_i^\beta = \delta_{\alpha\beta}$. According to this model, the MSD of the mutant structure is given by three terms: (1) The mean squared thermal fluctuations (nomut), which are inversely proportional to the global force constant parameter K. (2) The MSD produced by the mutation force (mut), which depends on the mutation weights. (3) The cross term given by the scalar product of the deformations produced by the mutation and by the null model.

The cross term term cannot be neglected, and its sign cannot be predicted. We estimate it through its root mean square (RMS) value, Eq.(6) multiplied times the "random" sign $\sigma_p = \pm 1$ that is specific of each wild-type-mutant pair. We do not need to know σ_p for predicting the MSD due to the mutation, but only for assessing the prediction. To this end, we assign σ_p according to whether the observed MSD is higher or lower than the expected MSD based on the null model

alone. Except for this sign, the MSD only depends on the mutation force and on the force constant K.

An additional problem is that the sometimes observed RMSD may be attributed to functionally distinct conformations more than to the effect of the mutation. Luckily, we can detect such instances and omit them from the training and test sets. Some pairs of structures significantly violate the null model of random conformational changes, since they present low frequency normal modes that contribute to the motion more than expected according to the null model. This implies that the harmonic free energy barrier that opposes to the motion is smaller than expected and the motion is more likely to happen spontaneously, suggesting that the observed conformation change is functionally important, such as an opening-close transition or ligand binding (Dos Santos et al. 2013, Bastolla and Dehouck 2019). Therefore, we eliminated conformational changes that present predicted harmonic energy barriers more than 50% lower than expected.

Model of protein stability change upon mutation

This part of the model is the same as the one presented in (Arenas et al. 2015), and is recalled here for completeness. We estimate the stability of the experimentally known native state of a protein adopting the contact matrix representation of its structure, $C_{ij} = 1$ if the residues i and j are in contact and zero otherwise. We define two residues to be in contact if any pair of their heavy atoms, either in the side chain or in the main chain, 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 $|i - j| \leq 2$. Note that we use the same contact matrix representation for computing the Hessian matrix of the TNM model, but with $C_{ij} = 1$ also for short range contacts.

The free energy of a protein in the mesoscopic state described by C_{ij} is modelled as the sum of contact interactions, $E(C, A) = \sum_{i<j} C_{ij} U(A_i, A_j)$, which depend on the types of amino acids in contact A_i and A_j and on 210 contact interaction parameters $U(a, b)$. We use the parameters $U(a, b)$ determined in (Bastolla et al. 2000).

For simplicity, we neglect the conformational entropy of the folded native state and estimate its free energy as $G_{\text{nat}}(C^{\text{nat}}, A) \approx \sum_{i<j} C^{\text{nat}}_{ij} U(A_i, A_j)$. 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 and Orland 1988, Shakhnovich and Gutin 1989), the resulting free energy is approximately described by the Random Energy Model (REM) (Derrida 1981),

$$G_{\text{misf}} \equiv -T \log \left(\sum_C \mathrm{e}^{-\sum_{i<j} C_{ij} U(A_i, A_j)/T + S(C)} \right) \approx \langle E \rangle - \frac{\left\langle (E - \langle E \rangle)^2 \right\rangle}{2T} - LS_C T$$

where LS_C is the logarithm of the number of compact contact matrices and $\langle . \rangle$ represents the average over the set of alternative compact contact matrices of L residues. 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. This

approximation is in good agreement with explicit computations of the free energy obtained through threading and real protein sequences present significantly lower stability of the misfolding ensemble than randomly reshuffled sequences (Minning et al. 2013). The previous estimate only holds above the freezing temperature of the REM (Derrida 1981), while the free energy is kept constant below the freezing temperature.

Note that, although the free energy is defined in terms of Boltzmann averages, the result of the calculation depends only on the first two moments of the contact energies. We compute these moments from the corresponding moments of the contacts. The first moment of the energy, $\langle E \rangle = \sum_{i<j} \langle C_{ij} \rangle U(A_i, A_j)$, depends on the moments of the contacts $\langle C_{ij} \rangle$, which are pre-calculated in order to accelerate the computation, which is important to achieve computational efficiency since the program must iteratively calculate averages over the site-specific distributions. The moments of the contacts are computed over the set of compact conformations of a chain of L residues by threading the chain over a non-redundant subset of the PDB, considering only compact submatrices whose number of contacts is larger than a threshold that depends on L. To reduce the number of parameters, ee assume that $\langle C_{ij} \rangle$ depends only on the contact range $|i-j|$ and on chain length L, $\langle C_{ij} \rangle = a(L) f(|i-j|)$ and determine the factor $a(L)$ so that the average number of contacts for structures of length L is the same as for native structures (Minning et al. 2013). For the contact correlation term, $\sum_{ij,kl} \left(\langle C_{ij} C_{kl} \rangle - \langle C_{ij} \rangle \langle C_{kl} \rangle \right) U_{ij} U_{kl}$, we compute approximate averages over two sites and consider three kinds of terms. (1) $ij = kl$: $\sum_{ij} \left(\langle C_{ij} \rangle - \langle C_{ij} \rangle^2 \right) U_{ij}^2$, where we used that $\left\langle C_{ij}^2 \right\rangle = \langle C_{ij} \rangle$. (2) $\sum_{ij} \left(\langle n_i n_k \rangle - \langle n_i \rangle \langle n_k \rangle \right) U_i U_k$, where $n_i = \sum_j C_{ij}$ is the number of contacts of site i and $U_i = \sum_j \langle C_{ij} \rangle U_{ij}/n_i$ is its average energy. (3) $\sum_{ij} \left(\langle C_{ij} N_c \rangle - \langle C_{ij} \rangle \langle N_c \rangle \right) U_{ij} U$, with $N_c = \sum_{ij} C_{ij}$ total number of contacts and $U = \sum_{ij} \langle C_{ij} \rangle U_{ij}/N_c$ average energy of the whole chain.

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

$$\Delta G(C^{\mathrm{nat}}, A) = G_{\mathrm{nat}} - kT \left(\mathrm{e}^{-G_{\mathrm{misf}}/kT} + \mathrm{e}^{-G_U/kT} \right) , \tag{7}$$

where the free energy of the non-native state is computed as a Boltzmann average, which is essentially equal to G_{misf} when the sequence is hydrophobic ($G_{\mathrm{misf}} - G_U/kT \ll -kT$) and is essentially equal to G_U when the sequence is hydrophilic ($G_{\mathrm{misf}} - G_U/kT \gg kT$). When we neglect stability against misfolding, we estimate $\Delta G = G_{\mathrm{nat}}(C^{\mathrm{nat}}, A) + LS_U$ as if all sequences were hydrophilic.

Selection on protein structure and stability

The SSCPE models that we present here are based on the formal analogy between the Boltzmann distribution in statistical physics, where conformations are weighted by minus the exponential of the energy divided $k_B T$ i.e. $P(C) = \exp(-E(C)/k_B T)/Z$, and the stationary distribution in sequence space that arises from an evolutionary process with Kimura's fixation probability. The latter distribution is proportional to the exponential of the logarithm of fitness $\varphi(A)$ multiplied times a selection parameter Λ related with the effective population size N, $P(A) = \exp(\Lambda \varphi(A))$ (Sella and Hirsch 2005, Halpern and Bruno 1999).

Similarly to the Boltzmann distribution, the sequence distribution can be formally defined as the distribution that maximizes sequence entropy with the constraint of given average value of the logarithm of fitness, $\sum_A P(A)\varphi(A)$. This formulation is valid when all amino acids a have

the same probability of being realized through the mutation process. Under a realistic mutation process, each amino acid will have a probability $P^{\mathrm{mut}}(a)$ of being realized under mutation alone, and the condition of maximum sequence entropy must be generalized by the condition that the sequence distribution has minimal Kullback-Leibler divergence from the distribution under mutation alone, $\prod_i P^{\mathrm{mut}}(A_i)$.

The above definition is only formal, because it is impossible to compute the normalization factor due to the huge number of possible sequences A. Nevertheless, this computation can be performed under the approximation that protein sites are independent (Arenas et al. 2015), which is almost necessary for computing the phylogenetic likelihood, and is commonly used in molecular evolution studies. If the fitness only depends on the properties of each individual protein site i, i.e. we can define a function $\varphi^i(A_i)$ that evaluates the fitness of amino acid A_i at position i independent of all other amino acids, the stationary amino acid distribution may be represented as

$$P(A_1 \cdots A_L) = \prod_i P^i(A_i, \Lambda) \tag{8}$$

$$P^i(A_i, \Lambda) = \frac{1}{Z_i} P^{\mathrm{mut}}(A_i) \exp(\Lambda \varphi^i(A_i)) \tag{9}$$

with $Z_i = \sum_a P^{\mathrm{mut}}(a) \exp(\Lambda \varphi^i(a))$.

Stab-CPE mean-field model

In (Arenas et al. 2015) we introduced the mean-field (MF) model of stability-constrained protein evolution, in which fitness is proportional to the fraction of protein that is in the native state, which can be computed from the folding free energy of the protein ΔG (Goldstein 2011, Serohijos and Shakhnovich 2014, Bastolla et al. 2006) and the logarithm of the fitness is proportional to $-\Delta G$. In the spirit of mean-field models of physics, the function $\varphi^i(A_i)$ is determined self-consistently by averaging the exponential factor over the distributions generated by all other sites. In this way, we recover the influence of interactions between sites in a model where sites are formally independent:

$$e^{\Lambda \varphi^{i,\mathrm{MF}}(a)} = \sum_{\{A\}, A_i = a} \prod_k P^{k,\mathrm{MF}}(A_k, \Lambda) e^{-\Lambda \Delta G(\{A\})} \,. \tag{10}$$

The MF distribution can be computed rapidly through an iterative algorithm whose complexity scales as the square of the number of sites, taking advantage of the precomputation of the statistics of misfolded conformations. We perform the computation with $\Lambda = 1$, obtain $\varphi^{k,\mathrm{MF}}(a)$ and then determine the optimal selection parameter Λ as explained in the next subsection.

Stab-CPE wild-type model

Next, we define the stability constrained wild-type model (WT) that defines the fitness as minus the change in folding free energy between the wild type sequence A^{WT} and the mutant sequence with amino acid a at site i:

$$\varphi^{i,\mathrm{WT}}(a) = -\Delta\Delta G(A_i^{\mathrm{WT}} ia) \tag{11}$$

where $A_i^{\mathrm{WT}} ia$ is the mutant protein in which amino acid A_i^{WT} at site i is changed to a, and $\Delta\Delta G$ is the corresponding change of folding free energy. The WT model is mathematically simpler than

the MF model, since it avoids the self-consistent calculation of the selection factors, but produces results that agree even better with the amino acid frequencies of extant protein families (Arenas et al. 2019). In fact note that, despite sequence space is exhaustively explored within the limit of the MF approximation, the native structure is maintained fixed, reducing the possible advantages of the more general MF approach.

Str-CPE RMSD model

The structurally constrained models that we introduce here are similar in spirit to the WT model, since they are based on the predicted structural changes from the WT structure. In the RMSD model the logarithm of the fitness is modelled as minus the RMSD predicted by the TNM program for any specific mutation, i.e. the fitness decreases as the mutated structure becomes more different from the wild-type structure:

$$\varphi^{i,\mathrm{RMSD}}(a) = -RMSD(A_i^{\mathrm{WT}} ia)\,. \tag{12}$$

Str-CPE DE model

We also evaluated the fitness function proposed by Echave (2008), whose logarithm is defined as minus the change in harmonic energy (DE) due to the mutation:

$$\varphi^{i,\mathrm{DE}}(a) = -\Delta E^{\mathrm{harm}}(A_i^{\mathrm{WT}} ia)\ , \tag{13}$$

$$\Delta E^{\mathrm{harm}} = (\Delta r, H\Delta r) = \left(F^{\mathrm{mut}}, H^{-1}F^{\mathrm{mut}}\right) = \sum_{\alpha} \left(\frac{F_\alpha^{\mathrm{mut}}}{\omega_\alpha}\right)^2\ , \tag{14}$$

where F^{mut} is the perturbation generated by any specific mutation. The new version of the program TNM, available at https://github.com/ugobas/tnm, computes and outputs the RMSD and the DE produced by all point mutations of the wild-type sequence if this computation is requested in the input file.

SSCPE model

Finally, we consider the four possible combinations of Stab-CPE and Str-CPE models, which we call SSCPE models:

$$\Lambda\varphi^{i,\mathrm{SSCPE}}(a) = \Lambda_1\varphi^{i,\mathrm{StabCPE}}(a) + \Lambda_2\varphi^{i,\mathrm{StrCPE}}(a) \tag{15}$$

The SSCPE models have two selection parameters Λ_1 and Λ_2.

Cut-off probability

In phylogenetic applications it is important that the amino acid frequencies are not zero for any amino acid, thus we enforce the minimum frequency $P^i(a,\Lambda) \geq \epsilon = 0.001$. The drawback is that with this condition the amino acid distribution is not an analytic function of Λ.

Selection parameter and regularization

The most important parameter of the site-specific substitution processes is the global selection parameter Λ, which we determine by maximizing the log-likelihood of the site-specific amino acid frequency observed in the input MSA:

$$\mathrm{LL}(\Lambda) = \sum_i \sum_a f^{i,\mathrm{MSA}}(a) \log \left(P^i(a, \Lambda) \right) \tag{16}$$

The Log-likelihood can be maximized analytically with cut-off parameter $\epsilon = 0$ as the value where the derivative vanish, and then recalculated with the actual value $\epsilon = 0.001$. The same calculation can be performed when there are two selection parameters Λ_1 and Λ_2.

As with all optimization problems, it is very important to regularize the parameter that is optimized in order to limit overfitting (it is actually surprising, in our opinion, that the regularization problem is so often overlooked in phylogenetic methods based on ML). The overfitting is particularly evident when there are few sequences in the MSA, or they are similar to the wild-type sequence. Since in the RMSD and DE models the most favored amino acid is always the wild-type one, in this case the Λ parameter increases without limits in order to maximize the probability of the wild-type sequence, which is a clear example of overfitting. In order to limit the optimal value of Λ, we apply one of two following regularizations.

Tykhonov regularization with specific heat criterion

This approach is often applied in the context of least squares, where it is called ridge regression (Hoerl and Kennard 1970). In the likelihood context, it can be expressed as maximizing $\mathrm{LL}(\Lambda) - R\Lambda^2$where R is the regularization parameter. Note that R is not a new fitted parameter, on the contrary it limits the likelihood achieved through the fit of the selection parameter Λ. Also note that this parameter is often denoted as Λ in the context of ridge regression, while here Λ is the parameter to be optimized, which we determine by numerically solving through Newton's method the regularized equation with cut-off parameter $\epsilon = 0$:

$$\frac{\partial \mathrm{LL}}{\partial \Lambda} - 2R\Lambda = \sum_i n_i \sum_a P^i(a, \Lambda)\varphi^i(a) - \Lambda \sum_{ia} f^i(a)\varphi^i(a) - 2R\Lambda = 0. \tag{17}$$

In (Dehouck and Bastolla 2017) one of us noted the analogy between regularization and thermodynamics, in the sense that the error of the fit can be interpreted as an energy, which increases when increasing the regularization parameter R, interpreted as a temperature. Adopting this analogy, in the likelihood context we can define a specific heat C_V as

$$C_V \equiv -\frac{\partial \mathrm{LL}(\Lambda, R)}{\partial R} > 0 \tag{18}$$

We found that, in many least square problems, the specific heat presents a maximum that separates a small R regime where overfitting is strong from a large R regime dominated by bias. The maximum of the specific heat presents good performances for different regression problems (Dehouck and Bastolla 2017, Bastolla and Dehouck 2019), and we adopt this criterion for determining the R parameter. Here we take advantage of the fact that the specific heat can be computed analytically for cut-off $\epsilon = 0$.

$$C_V = \left(\frac{\partial \mathrm{LL}}{\partial \Lambda}\right) \frac{2\Lambda}{2R + \sum_{ia} f^i(a)\varphi^i(a) + \sum_i n_i \left(\langle \varphi^i(a)^2\rangle - \langle \varphi^i(a)\rangle^2\right)} \tag{19}$$

Pseudo-counts

A popular way to regularize observed amino acid frequencies when few sequences are available consists in using pseudo-counts of amino acids that are not observed at this position. These regularized frequencies are given by $\left(n_i f^{i,\mathrm{MSA}}(a) + R f^{\mathrm{glob}}(a)\right)/(n_i + R)$, where n_i is the number of amino acids observed at site i, R is the regularization parameter, and $f^{\mathrm{glob}}(a)$ is the global amino acid distribution observed at all sites or over a large database of proteins. We then define the log-likelihood as

$$\mathrm{LL}'(\Lambda) = \sum_i \sum_a \frac{n_i f^{i,\mathrm{MSA}}(a) + R f^{\mathrm{glob}}(a)}{n_i + R} \log\left(P^i(a,\Lambda)\right) \tag{20}$$

We determine Λ by maximizing LL'. Again, the real likelihood Eq.(16) decreases with R, and we can define a "specific heat" through Eq.(18), which we calculate numerically. We choose the value of R that maximizes this specific heat.

The two regularization procedures produce qualitatively similar results, but they are slightly better for the pseudo-count regularization, probably due to the small error induced by the cut-off parameter $\epsilon = 0.001$ that is neglected in the maximization. The presented results are based on the pseudo-count method.

Substitution process

Up to now, we have only defined the site-specific amino acid frequencies of the substitution process. To obtain a full substitution process Q^i_{ab}, we define the exchangeability matrix E^i_{ab} at site i through the Halpern-Bruno theory (Halpern and Bruno 1998):

$$Q^i_{ab} = E^i_{ab} P^i(b) \tag{21}$$

$$P^i(a) = \frac{1}{Z_i} P^{\mathrm{mut}}(a) e^{\Lambda \varphi^i(a)} \tag{22}$$

$$E^i_{ab} = E^{\mathrm{mut}}_{ab} \left(\frac{\Lambda\left(\varphi^i(b) - \varphi^i(a)\right)}{\frac{1}{Z_i}\left(e^{\Lambda\varphi^i(b)} - e^{\Lambda\varphi^i(a)}\right)}\right) \tag{23}$$

This substitution process can be interpreted as the combination of a global mutation process with frequencies $P^{\mathrm{mut}}(a)$ and exchangeability E^{mut}_{ab} and a site-specific and site-independent selection processes with logarithm of fitness $\varphi^i(a)$ and fixation probability given by Kimura's formula. The resulting substitution matrices Q^i_{ab} can be used in likelihood-based phylogenetic analyses.

Global frequencies and exchangeability

Finally, the global mutation frequencies are defined by maximizing the likelihood of the observed amino acids, which yields

$$P^{\mathrm{mut}}(a) = \frac{\sum_i f^{i,\mathrm{MSA}}(a)}{\sum_i \frac{1}{Z_i} e^{\Lambda\varphi^i(a)}} \tag{24}$$

We first determine $P^{\rm mut}(a)$ with $\Lambda = 0$, use these frequencies to determine Λ, update $P^{\rm mut}(a)$ and iterate the process. Two steps are usually enough to achieve the maximum likelihood.

To determine the global exchangeability matrix $E^{\rm mut}_{ab}$, we consider the flux $R^i_{ab} = P^i_a P^i_b E^i_{ab}$ between any pair of amino acids a and b and we impose that the flux averaged over all sites i equals the flux measured in an empirical model of the substitution process, $\sum_i R^i_{ab}/L = \left(P^{\rm emp}_a P^{\rm emp}_b E^{\rm emp}_{ab}\right)$. We get

$$P^{\rm mut}(a)P^{\rm mut}(b)E^{\rm mut}_{ab} = \frac{P^{\rm emp}(a)P^{\rm emp}(b)E^{\rm emp}_{ab}}{\frac{1}{L}\sum_i \frac{1}{Z_i} e^{\Lambda(\varphi^i(a)+\varphi^i(b))} \frac{\Lambda(\varphi^i(b)-\varphi^i(a))}{e^{\Lambda(\varphi^i(b)}-e^{\Lambda(\varphi^i(a)}}} \tag{25}$$

Results presented here are obtained with the JTT matrix (Jones et al. 1992), but the WAG matrix (Whelan and Goldman 2001) is also implemented.

Site-specific sequence entropy and substitution rate

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

$$S_i = -\sum_{a=1}^{20} P^i(a)\log(P^i(a)). \tag{26}$$

The site-specific substitution rates of the evolutionary models are computed as the weighted average of the substitution rate matrix $Q^i_{ab} = E^i_{ab}P^i(b)$,

$$R^i = \sum_{a\neq b} P^i(a)P^i(b)E^i_{ab} \tag{27}$$

Results

Predicting the amount of structure change caused by a mutation

We determined the weights $W_{\rm size}$, $W_{\rm stab}$ and $W_{\rm dist}$ in Eq.(4) by numerically fitting Eq.(6) over a training set of mutants with RMSD$>$ 0.15Å. From the starting set of 222 pairs, we eliminated 27 pairs of conformational changes that are likely to have a functional meaning (see Methods), obtaining 195 mutations that we split into training (156) and test (39) set with RMSD$>$ 0.15Å and with standard deviation 0.28ÅḊuring the optimization process, some borderline pairs may be omitted if the random realization of the null model does not fulfill the selection criteria.

To avoid the influence of the parameter σ_p, which is assessed from the observed RMSD (see Methods), we optimized the correlation between the RMSD predicted only with the mutation and the observed RMSD, which is not influenced by σ_p. Note that σ_p does not influence the predicted RMSD produced by the mutation, which is needed for modelling the substitution process. The parameters that result from the optimization are $W_{\rm size} = 3.05$, $W_{\rm dist} = 15.6$ and $W_{\rm stab} = 2.39$, yielding the correlation coefficient between observed and predicted RMSD $r = 0.86$ and root mean square error (RMSE) of 0.20Å over the training set from wild-type to mutant, and $r = 0.75$, RMSE=0.22Å from mutant to wild-type (the two directions differ since the torsional network model of the mutant is slightly different from the one of the wild-type). We evaluate the performances of the model on the test set, finding correlation coefficients $r = 0.95$ and $r = 0.97$ and RMSE 0.16Å and RMSE 0.11Å from wild-type to mutant and from mutant to wild-type, respectively (see

Supplementary Figure S1). These correlations were very high because the test set contained only one pair with large RMSD, which is more difficult to predict. Considering only the mutation model we obtained poorer correlations (0.54, 0.28, 0.68 and 0.65, respectively for the training set in the two directions and the test set), which supports that a combination of the mutation model, the null model and their cross product improves the prediction of the conformational change caused by a mutation. We will present more detailed results in a forthcoming publication.

For the sake of the evolutionary applications of the model, three general properties of the mutation model, independent of the particular values of the parameters, are important. They are shared with the random force model proposed by Julián Echave. The only advantage of the present model is that it can be explicitly applied to generate substitution models for phylogenetic inference.

(1) Due to Eq.(3), the more contacts the mutated site m has, the more terms the resulting mutation force has. Therefore, mutating sites with more contacts tend to generate larger perturbations, and as a consequence these mutations are less tolerated. This is exemplified in Fig.1A, which shows the average RMSD generated by a mutation as a function of the number of contacts of the mutated site.

(2) Due to Eq.(2), sites with more native contacts tend to be less deformed by the mutation because they produce lower flexibility in the ENM (their eigenvector components v_i^α are small in low frequency normal modes that contribute most to the thermal motion). As a consequence, the structural effect of mutation is weaker at highly connected sites, as observed in Fig.1B, which shows the RMSD of site i averaged over all possible mutations as a function of the number of contacts of the mutated site.

(3) Finally, Eq.(3) implies that the RMSD caused by a particular mutation is correlated with the corresponding change of native contact energy, and the extent of this correlation grows with the parameter W_{stab}.

Comparison of the likelihood of amino acid frequencies

In Fig.2, we compare the results obtained with the 8 different selection models on the MSA of 208 monomeric proteins. For all models, the Λ parameter was determined maximizing the likelihod of the MSA profile at the regularization value that maximizes the specific heat for that particular model. One can see from Fig.2A that the average likelihood can be ranked as MF<WT<DE<RMSD, i.e. Str-CPE models generally present higher likelihood than the Stab-CPE models. The SSCPE models present higher likelihood, the highest being for the RMSDWT model, although the improvement is not very large. The reversed ranking can be seen in Fig.2B for the sequence entropy, i.e Str-CPE models present lower entropy than Stab-CPE models, which indicates that they are less tolerant to mutations. Once again, we obtain the lowest entropy with the RMSDWT model, but the difference is not large. However, this model presents lower folding free energy than the RMSD model, Fig.2C, i.e. combining the RMSD model with the WT model we model more stable proteins. Finally, the mean hydrophobicity is equal to the mean hydrophobicity of the PDB sequences for the MF model, and it is slightly higher for the other models, see Fig.2D, evidencing a bias towards hydrophobic sequences, which is stronger for the WT model.

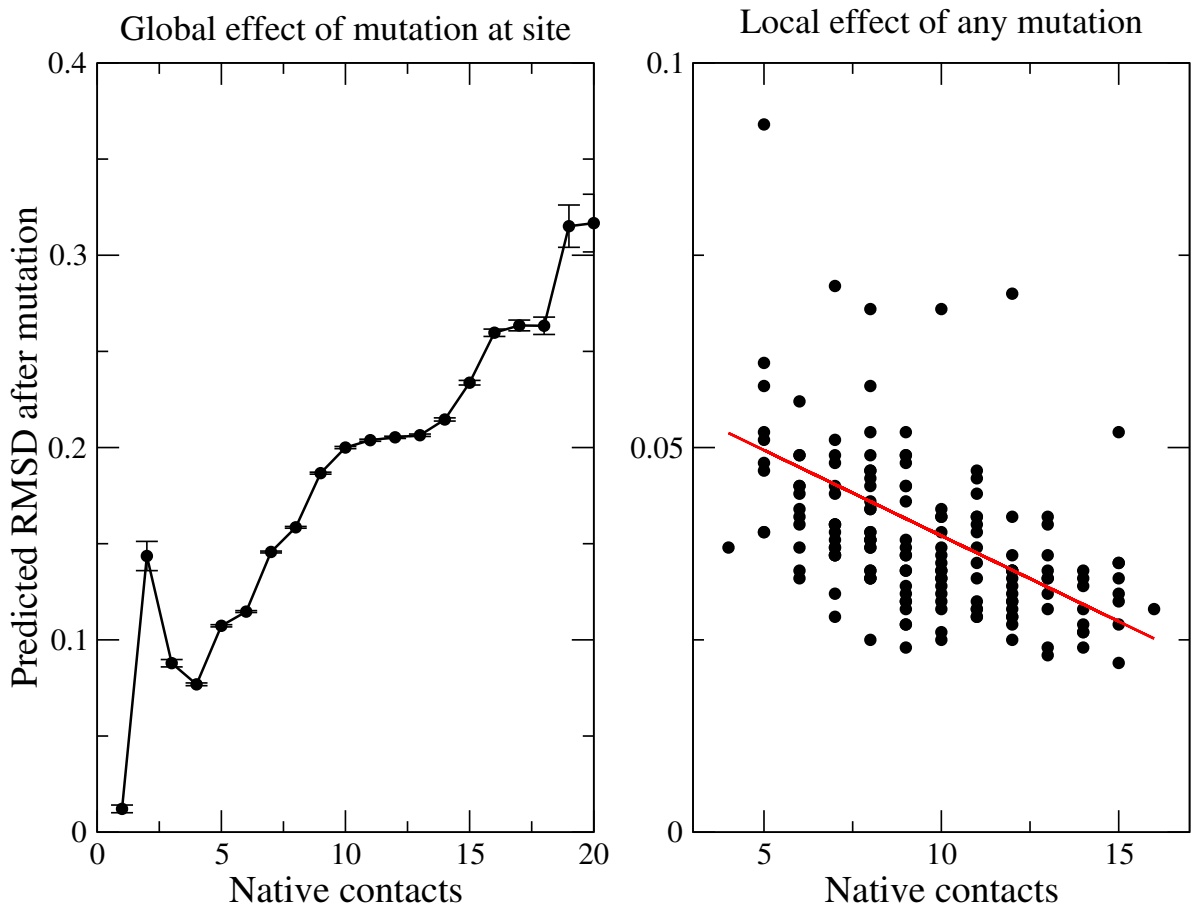

Figure 1: Effect of the mutation model as a function of the number of contacts of the mutated site (A) and the number of contacts at the position where the structure change is monitored (B).

Site-specific entropy and substitution rates

Fig.3 shows the average site-specific sequence entropy (A) and substitution rate (B) as a function of the number of native contacts for the highest likelihood Stab-CPE model (typically WT), Str-CPE model (RMSD) and SSCPE model (RMSDWT). The Stab-CPE model presents a maximum for intermediate number of contacts both for the sequence entropy and for the substitution rate, indicating that the model is too tolerant for sites with intermediate number of contacts. In contrast, both sequence entropy and substitution rate tend to decrease with the number of contacts for the Str-CPE and SSCPE model except for sites with very high number of contacts, in qualitative agreement with observations from MSAs. Real data for MSAs were also reported in (Jimenez-Santos et al. 2018), and are shown here for completeness. Nevertheless, all models are quantitatively more tolerant than real MSA data in terms of both sequence entropy and substitution rates.

Figure 3C represents the mean hydrophobicity of each site as a function of the number of contacts in the native state, for three representative models and for the real MSA. One can see that in all cases the hydrophobicity tends to increase with the number of contacts, as expected

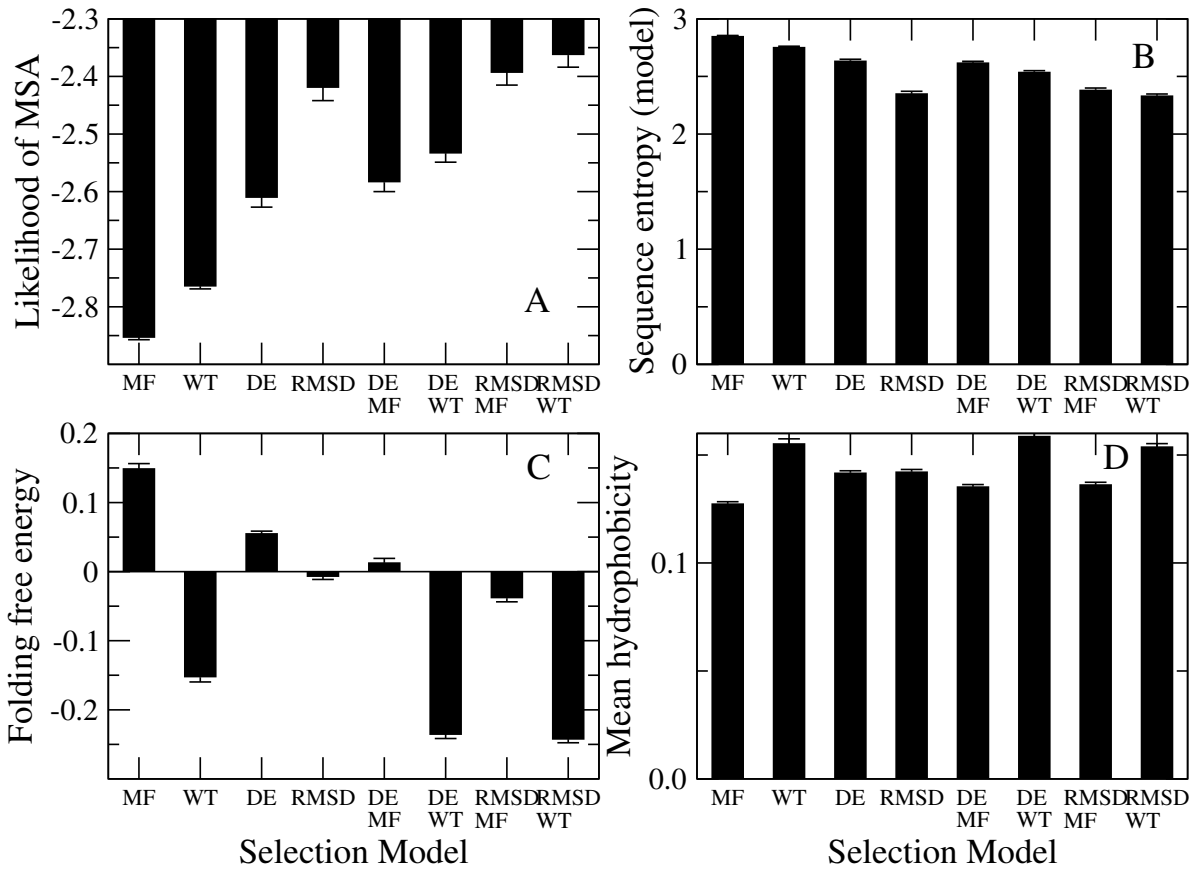

Figure 2: Comparison of results obtained with different selection models, in terms of likelihood of the observed MSA (A), sequence entropy (B), Average folding free energy (C) and average hydrophobicity (D) of the modelled protein sequences.

since this trend favors folding stability. However, interestingly the maximum hydrophobicity is attained before the maximum number of contacts (11 instead of 15) both for the MSA and for the models. In general, the models capture well the behaviors of the observed data, with the SSCPE model showing the best agreement. The mean hydrophobicity of the MSA, however, spans a wider range than the models. In particular, it is lower than the models for small number of contacts and higher for high number of contacts, again suggesting that the models may be more tolerant than real protein evolution.

Comparison of the phylogenetic likelihoods

Finally, we used the site-specific substitution processes to infer phylogenetic trees of 10 randomly chosen protein alignments of our database (Table 1) with the ML method implemented in the program RAxML-NG (Kozlov et al. 2019) and compared their AIC scores (double of minus the log-likelihood plus the number of fitted parameters) with those of phylogenetic trees reconstructed with the widely used empirical substitution model JTT +F +G (Jones et al. 1992). For all types of models we adopted the JTT exchangeability matrix and the option +F, therefore the difference

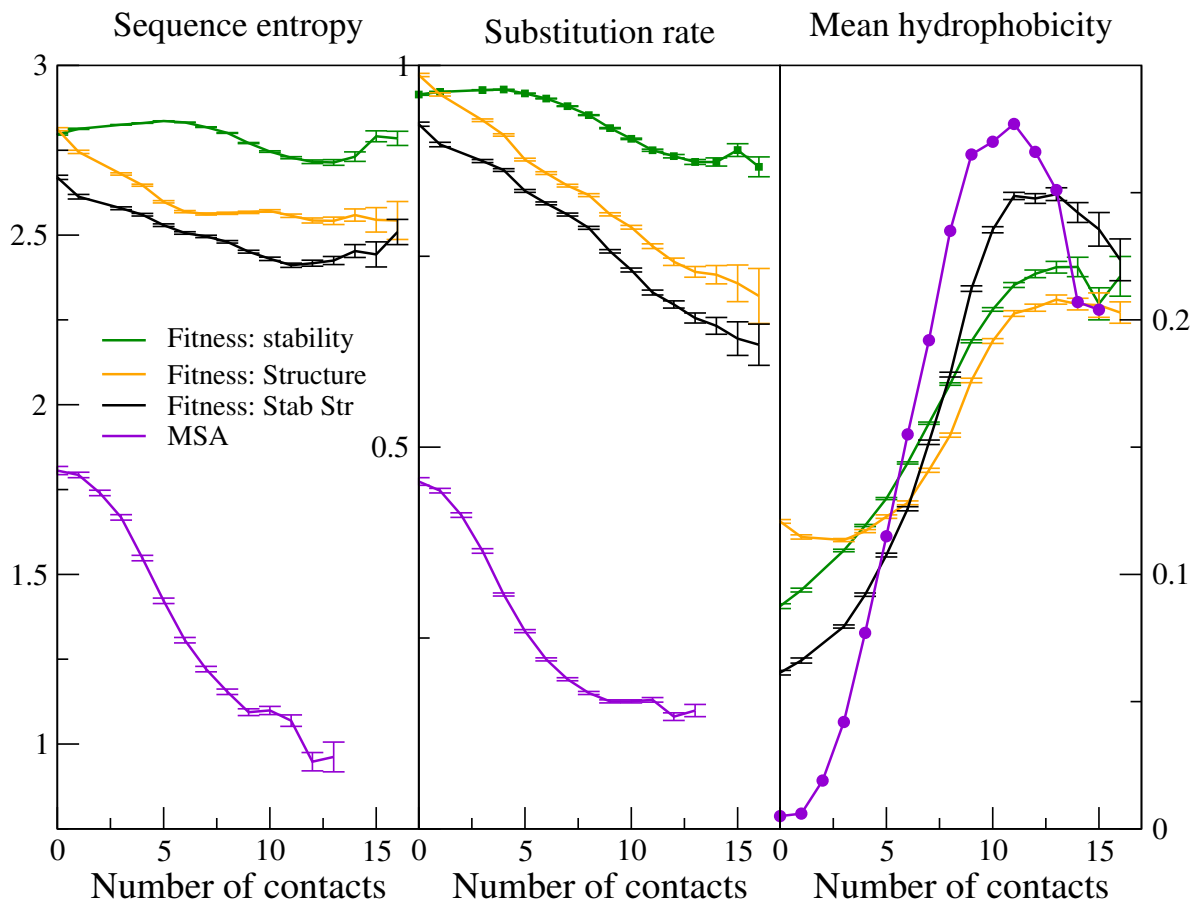

Figure 3: Average site-specific sequence entropy (A), substitution rate (B) and hydrophobicity (C) as a function of the number of contacts of the site in the native state of the wild-type sequence.

between the empirical and the site-specific models is that the latter consider selection on the protein structure instead of a gamma distribution (+G) for taking into account across-sites rate variability. The number of fitted parameters is the same for the empirical model and the Stab-CPE or Str-CPE models, with the parameter of the gamma distribution substituted by the parameter Λ of our models that represents the selection strength, while the SSCPE model has one additional parameter Λ_2.

Results are presented in Table 1 and Fig.4. For each protein family, we compared the empirical model JTT with our Stab-CPE, Str-CPE and SSCPE model, which we indicate by the symbols STAB, STR and STRSTAB, respectively. For each class of model, we choose the one that provides the highest likelihood to the input MSA and typically correspond to the WT, RMSD and RMSDWT models, respectively. To reduce the dependence on the protein length, the number of sequences in the alignment and the sequence identity, we normalized the AIC by the product of the number of sites (L) and the total branch length of the tree inferred with the empirical model.

Our results show that the Stab-CPE model presents a substantial improvement of the phylogenetic fit (from 0.03 to 0.50 normalized AIC units) for 8 out of 10 protein families (Fig.4). There is no improvement for the family with PDB code 1onr, and only a small improvement for 1qam. In

Table 1: The ten protein families used for computing phylogenetic likelihood with the program RAxML-NG, ranked by sequence identity with the protein with known structure. ΔAIC/(Lt) denotes the difference of AIC score between the JTT+F+G model and the structure-aware models normalized by the product of the number of sites (L) and the total branch length of the tree (t) inferred with the JTT model.

Protein family	Num. of sequences	Residues in PDB	Mean seq. identity	PDB code	ΔAIC/Lt Stab-CPE	ΔAIC/Lt Str-CPE	ΔAIC/Lt SSCPE
Methyltransferase	302	235	0.25	1qam	0.008	0.011	0.047
Dehydrogenase	302	307	0.28	1s3i	0.030	0.062	0.122
Thioredoxins I	302	123	0.41	1bp2	0.081	0.104	0.191
Phospholipase	302	556	0.42	1m53	0.056	0.119	0.195
NAD+ ligase	302	310	0.52	1b04	0.046	0.198	0.270
Imidazole glycerol phosphate synthase	302	251	0.53	2a0n	0.110	0.253	0.293
Phosphotransferase	302	217	0.53	1zio	0.027	0.209	0.303
RuvB Hydrolase	302	314	0.59	1hqc	0.147	0.368	0.440
Transaldolase	302	316	0.61	1onr	−0.002	0.380	0.461
Transglutaminase	10	331	0.84	1iu4	0.496	7.138	6.505

contrast, the RMSD model presents a substantial improvement for 9 out of 10 MSA (the exception is again 1qam), and the SSCPE model was the best fitting model, except for the MSA of 1iu4, for which the RMSD model has higher likelihood, showing for all 10 MSA an improvement from 0.05 to 6.5 normalized AIC units. Note that, unlike the parameter of the gamma distribution of the empirical models, the Λ parameter is not determined so to maximize the phylogenetic tree but only the likelihood of the MSA, which is unaware of the phylogeny. One can see from Table 1 that the improvement of the AIC is positively correlated ($r > 0.7$) with the mean sequence identity with the protein in the PDB.

Discussion

Although the selective forces that act in protein evolution are complex and difficult to model, many years of modelling and observations have revealed some general biophysical principles that most natural proteins must obey (Liberles et al. 2012, Wilke 2012, Sikosek and Chan 2014). However, the empirical substitution processes that are commonly used for phylogenetic inference ignore these biophysical principles, and they fit the site-specific variability of substitution rates through the gamma model (Yang, 1993), which is unaware of the protein structure. Introducing biophysical realism in the models of protein evolution holds the promise of improving phylogenetic inferences.

The biophysical principle that received most attention is the folding stability of the native state, reviewed for instance in (Goldstein 2011, Serohijos and Shakhnovich 2014, Bastolla et al. 2017), which produced site-specific stability-constrained substitution models that can be used for phylogenetic inference (Arenas et al. 2015). Nevertheless, these Stab-CPE models assume for simplicity that the native structure is stictly conserved in protein evolution, and they cannot represent the selective process for maintaining the native structure, which is thought to be an

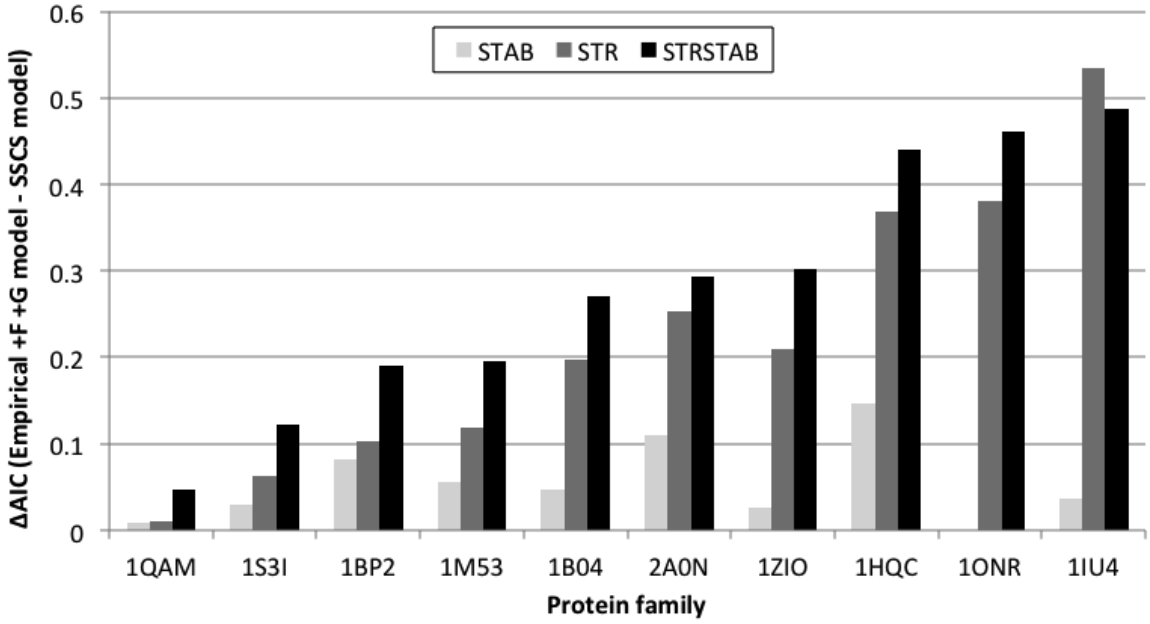

Figure 4: Difference between the AIC scores of the site-specific models with selection on stability and structure and the JTT+F+G empirical model computed by the RAxML-NG program. To facilitate comparisons, we normalized the difference by the product of the number of sites and the total length of the branches inferred with the empirical model. Protein families are ranked by sequence identity (Table 1). The horizontal labels are the PDB codes of the representative structure. The family 1iu4 is reduced by a factor 0.075 to allow presentation on the same scale, the real value is reported in Table 1.

important force both for negative selection (Illergard et al. 2009, Pascual-Garcia et al. 2010) and for positive selection (Pascual-Garcia et al. 2019) and that strongly influences the conservation of protein dynamics (Tirion 1996, Atilgan et al. 2001).

In this work, we introduced structurally constrained substitution processes based on the model by Julián Echave (2008) that predicts the structural changes due to mutations as the linear response of the elastic network that models the protein to the perturbation caused by the mutation. Here we advanced with respect to Echave's proposal, developing an explicit model of the perturbations generated by specific amino-acid mutations in terms of changes of size, stability and optimal distance of the native contacts formed by the mutated residue.

The resulting model has only 3 parameters (the weights of the three terms of the perturbation) that we optimized on a database of 190 point mutants with known experimental structure. A difficulty of this optimization is that the mean square deviation (MSD) between the wild-type and the mutant PDB structure is only partially caused by the mutation; a large part of it is caused by the change of experimental conditions (Carpentier et al. 2019) and by functional conformational changes. In order to develop an integrative model of the MSD between the wild-type and mutant protein structures, we exploited the fact that the change of experimental conditions between two structures of the same protein can be represented by a null model of linear response (Dos Santos et al. 2013). We also eliminated from the training and test set the pairs that significantly violate this

null model, which suggests that the conformation change is in part functional (Dos Santos et al. 2013). Unfortunately, one term of the resulting model, which consists of the scalar product between the deformation induced by the mutation and the one induced by the change of experimental conditions, has a sign that we cannot predict. However, we can estimate this sign depending on whether the observed MSD is larger or smaller than expected based on the null model, although this limits our ability to assess the model. Despite the intrinsic difficulty of the problem, we obtained satisfactory results, with a correlation coefficient between observed and predicted MSD of 0.80 over the training set and 0.96 over the test set.

Irrespective of the parameters, and similar to the original Echave's model, the model of linear response model predicts that mutations at sites with many contacts induce larger deformations (Fig.1A) and sites with many contacts tend to respond less to perturbations (Fig.1B). Furthermore, the specific form in which we modelled mutations implies that there is a correlation between the effect of any specific mutation on the native structure and on its folding stability.

We applied this model of the structural change due to mutations to define two structurally constrained fitness functions in which the logarithm of the fitness is represented by minus the predicted RMSD due to the mutation (RMSD model) and the predicted change in the harmonic energy of the elastic network (DE model), as proposed by Echave (2008). We then compared these Str-CPE models with two variants of our previous stability constrained substitution models in which the logarithm of the fitness is modelled as the folding free energy ΔG, evaluated with contact interactions, considering as non-native states both the unfolded state and compact but incorrectly folded states (Minning et al. 2013). These Stab-CPE models include the MF model, in which the fitness is self-consistently averaged with respect to the distribution generated by the model, and the WT model in which the fitness is approximated as the change of folding free energy $\Delta\Delta G$ between any point mutant and the wild-type sequence.

All these evolutionary models depend, besides global amino acid frequencies (+F), on a crucial parameter Λ that represents the strength of natural selection, and which in some sense corresponds to the parameter of the gamma distribution (+G) adopted by empirical models. We fit the Λ parameter by maximizing the likelihood of the site-specific amino acid frequencies of the MSA (which are phylogenny unaware) with respect to the model. To reduce overfitting, which would induce large values of Λ and vanishingly small probabilities for amino acids that are not observed, we regularized the fit either by the Tykhonov regularization scheme (Hoerl and Kennard 1970), also known as ridge regression, or by adding pseudo-counts to the MSA. In both cases, we choose the regularization parameter R that maximizes the derivative of minus log-likelihood with respect to R, which can be interpreted as the analogous of a specific heat (Dehouck and Bastolla 2017).

The resulting Str-CPE models provide higher likelihood to observed MSA than the corresponding Stab-CPE models (Fig.2A) and lower sequence entropy, which implies that they are less tolerant to mutations (Fig.2B). Next, we combined the Str-CPE and Stab-CPE models, obtaining structure and stability constrained protein evolution (SSCPE) models with even better agreement with empirical MSAs, lower entropy, higher folding stability (Fig.2C) and more realistic hydrophobicity (Fig.2D) than Str-CPE models.

We examined site-specific evolutionary properties as a function of the number of contacts of the site in the native structure, finding that, while Stab-CPE models present maximum tolerance to mutations in terms of both sequence entropy and substitution rate at intermediate number of contacts, which qualitatively disagrees with observations (Jimenez-Santos et al. 2018), the Str-CPE and SSCPE models predict that the tolerance to mutations tends to decrease with the number of

contacts, agreeing with observed data (Fig.3A, B). Moreover, Str-CPE and SSCPE models also produce a more realistic range of mean hydrophobicity (Fig.3C).

Nevertheless, from the quantitative point of view, all these substitution models of protein evolution appear to be more tolerant to mutations than the MSAs that we examined, which present lower sequence entropies and substitution rates. This may indicate that the models still do not capture important selective constraints, but it may also be partially attributed to insufficient sampling of the protein sequences, which underestimates the site-specific variability (only 300 sequences were sampled for each protein structure).

Therefore, the Str-CPE and SSCPE models represent site-specific selection on protein sequences more realistically than Stab-CPE models, and much more than empirical models, that do not consider site-specific selection.

In order to see whether this constitutes an advantage for phylogenetic inference, we inferred with the program RAxML-NG (Kozlov et al. 2019) the phylogenetic trees of 10 randomly chosen MSAs from our database of 208 monomeric proteins. We used either the empirical model JTT+F+G, or the best (in terms of likelihood of the MSA) Stab-CPE model (usually WT), Str-CPE model (usually RMSD) or SSCPE model (usually RMSDWT). We compared the AIC score in order to take into account that the SSCPE model has an additional fitted parameters can be different, and normalized the AIC by the product of the number of protein sites and the total length of the branches of the phylogenetic tree inferred with the JTT model.

The results from this analysis showed that the Stab-CPE model presents a substantial improvement of the phylogenetic fit (from 0.03 to 0.50 normalized AIC units) for 8 out of 10 protein families, the Str-CPE model presents a substantial improvement (from 0.06 to 7) for 9 out of 10 MSA, and the SSCPE model was the best model for 9 out of 10 MSA, presenting an improvement between 0.05 and 6.5 normalized AIC units, see Table 1 and Fig.4. This happens despite Λ is not fitted to maximize the phylogenetic likelihood but the MSA likelihood, which is phylogeny unaware, and the fit is regularized to limit overfitting. As a consequence, the fit of the Λ parameter can be computed much more rapidly than the fit of the parameter of the gamma distribution for the empirical model.

We found that the improvement is positively correlated with the mean sequence identity of the MSA with the PDB sequence (Table 1), i.e. MSAs that differ more from the reference structure present a smaller improvement with respect to the empirical models. Nevertheless, the improvement is still sizeable for the most diverse MSAs with mean sequence identity of 0.25 (1qam, normalized AIC improvement 0.05) and 0.28 (1s3i, normalized improvement 0.12).

On the other hand, the protein family 1iu4 presented a very large improvement (Table 1), which obliged us to rescale its value in Fig.4. This family contains only 10 very similar proteins, one of which presents a stretch of 10 residues (CLTRASSATG) largely different from the corresponding stretch of the PDB protein (-VYESKFRNW), probably arising through insertion and deletion, so that the residues are likely not homologous. This is supported by the positive folding free energy predicted by our program (7.5 instead of -6.5 for the protein in the PDB), which suggests that the mutated protein presents significant structural differences, as we would expect since they are present in the same genome (the bacterium Streptomyces Morabaensis) and they are likely on the way to differentiate their function and probably also their structure (Pascual-Garcia et al. 2019). Note that such sequences with positive predicted ΔG are not used for determining the Λ parameter of the SSCPE model. As a result, the normalized likelihood is very low both for the empirical model and for the site-specific models, but the latter present a very sizable improvement that suggests that they might be more robust to aligning non homologous regions.

Since the computations of the site-specific substitution matrices only take from seconds to minutes and the site-specific matrices can be readily handled by the programs RAxML-NG or PAML, these results support the use of these structure-aware and selection-aware models for phylogenetic inferences in the many cases in which a representative structure is available in the PDB.

Data Availability

The program Prot_evol is freely available at https://github.com/ugobas/Prot_evol and the program TNM for predicting the RMSD upon mutation is available at https://github.com/ugobas/tnm. The package SSCPE.zip at https://github.com/ugobas/SSCPE/blob/main/SSCPE.zip provides all needed codes, compiles and installs TNM and Prot_evol, and provides the SSCPE.pl script that generates the site-specific models that can be used with the RAxML-NG program.

Supplementary figures and data are stored in the Dryad data repository at http://datadryad.org. They include: 1. Figure S1 that assesses the RMSD predicted through our mutation model. 2. Mutation parameters used for TNM predictions. 3. MSAs of the 208 studied proteins. 4. RMSD and 5. DE predicted by the TNM program for all possible mutations of the studied proteins. 6. Input files for the tnm program and the Prot_evol program. 7. Data generated for testing the model: Site-specific substitution rates, sequence entropy, mean hydrophobicity and number of contacts. 8. Summary data for all tested models and all proteins. 9. Site-specific substitution models for the 10 protein families studied for phylogenetic inference. 10. Outputs of the RAxML-NG program for the 10 studied proteins.

Funding

This work was supported by the grants PID2019-109041GB-C22/10.13039/501100011033 and PID2019-107931GA-I00/AEI/10.13039/501100011033 of the Spanish Agency of Research (AEI). Research at the CBMSO is facilitated by the Fundación Ramón Areces.

Acknowledgements

We thank Centro de Supercomputación de Galicia (CESGA) for the computer resources.

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 CC, Liberles DA, Bastolla U. 2017. ProtASR: An evolutionary framework for ancestral protein reconstruction with selection on folding stability. Syst Biol. 66:1054-64.

[3] Arenas M, Bastolla U. 2020. ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. Meth. Ecol Evol. 11:248-257.

[4] Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J.* 80, 505–515.

[5] Bastolla U, Vendruscolo M, Knapp EW. 2000. A statistical mechanical method to optimize energy functions for protein folding. Proc. Natl. Acad. Sci. USA 97:3977-81.

[6] Bastolla U, Porto M, Roman HE, 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.

[7] Bastolla U, Porto M, Ortiz AR. 2008. Local interactions in protein folding determined through an inverse folding model. Proteins 71:278-99.

[8] Bastolla U. 2014. Detecting selection on protein stability through statistical mechanical models of folding and evolution. Biomolecules 4:291-31.

[9] 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.

[10] Bastolla U, Dehouck Y. 2019. Can Conformational Changes of Proteins Be Represented in Torsion Angle Space? A Study with Rescaled Ridge Regression. J Chem Inf Mod 59:4929-4941.

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

[12] Buckley, T. R., Simon, C. and Chambers, G. K. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst Biol 50, 67-86.

[13] Carpentier M, Chomilier J. 2019. Protein multiple alignments: sequence-based versus structure-based programs. Bioinformatics 35:3970-3980.

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

[15] Dehouck Y, Bastolla U. 2017. The maximum penalty criterion for ridge regression: application to the calibration of the force constant in elastic network models. Integr Biol 9:627-641.

[16] Dehouck Y, Bastolla U. 2021. Why are large conformational changes well described by harmonic normal modes? Biophys J. 2021 Dec 7;120(23):5343-5354.

[17] Dos Santos HG, Klett J, Méndez R, Bastolla U. 2013. Characterizing conformation changes in proteins through the torsional elastic response. Biochim Biophys Acta. 1834:836-46.

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

[19] Echave J, Jackson EL, Wilke CO. 2015. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. Phys Biol. 12:025002.

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

[21] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17: 368-376.

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

[23] Garel T, Orland H. 1988. Mean-field model for Protein Folding. Europhys. Lett. 6:307-10.

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

[25] Goldstein RA, Pollock DD. 2017. Sequence entropy of folding and the absolute rate of amino acid substitutions. Nature Ecol Evol 1:1923-30.

[26] Hoerl AE, Kennard RW. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12: 55-67.

[27] 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. 14:78.

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

[29] Illergard K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. Proteins 77:499-508.

[30] Jack BR, Meyer AG, Echave J, Wilke CO. 2016. Functional sites induce long-range evolutionary constraints in enzymes. PLoS Biol. 14:e1002452.

[31] 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. 35:743-755.

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

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

[34] Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35: 4453-4455.

[35] Liberles DA, et al. 2012. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21:769-85.

[36] 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.

[37] Mendez R, Bastolla U. 2010. Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. Phys Rev Lett. 104:228103.

[38] 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.

[39] Pascual-Garcia A, Abia D, Mendez R, Nido GS, Bastolla U. 2010. Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins 78:181-196.

[40] Pascual-García A, Arenas M, Bastolla U. 2019. The Molecular Clock in the Evolution of Protein Structures. Syst Biol 68:987-1002.

[41] 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.

[42] Saitou N, Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

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

[44] 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 26:84-91.

[45] Shakhnovich EI, Gutin AM. 1989. Formation of unique structure in polypeptide chains, Biophys. Chem. 34: 187.

[46] Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface. 11:20140419.

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

[48] 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.

[49] Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396-1401.

[50] Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol Evol. 11: 367-372.

[51] Yang, Z.; Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol 14: 717-724.

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

[53] Wilke CO. 2012. Bringing molecules back into molecular evolution. PLoS Comput Biol. 8:e1002572.