Site-specific structure and stability constrained substitution models improve phylogenetic inference

Ivan Lorca-Alonso[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

June 28, 2024

Abstract

In previous studies, 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 suggesting that conservation of the native structure is a stronger selective force than selection on folding stability.

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) that include one or more known structures, and better predict the observed conservation across sites. The models that combine Str-CPE and Stab-CPE models are even more stringent and fit the empirical MSAs better. We refer collectively to our models as structure and stability constrained substitution models (SSCPE). Importantly in comparison to the traditional empirical substitution models, the SSCPE models infer phylogenetic trees of distantly related proteins more similar to reference trees based on structural information. We implemented the SSCPE models in the program SSCPE.pl, freely available at https://github.com/ugobas/SSCPE, which infers phylogenetic trees under the SSCPE models with the program RAxML-NG from a concatenated alignment and a list of protein structures that overlap with it.

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 (i.e. frequencies and exchangeabilities are the same), 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, buried sites are much proner to be ocupied by hydrophobic amino acids, and the substitution rate varies 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 across 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 inference (Yang et al. 1997). For modelling substitution rate variation across sites, Yang (1993, 1996) proposed to adopt a gamma distribution (+G), fitting its shape parameter from the data. This approach greatly improves the fit of the data, but still it assumes the same substitution process at each site. Another approach to model rate variation was proposed by Halpern and Bruno (1998). Given a site where natural selection produces specific patterns of amino acid frequencies, they proposed how to derive an exchangeability matrix that represents the combined effect of mutation and the Fisher's fixation probability (Fisher 1930). This approach was later extended by Yang and Nielsen (2008) to selection for codon usage. However, this approach obtains the site-specific frequencies from the data, with the risk of substantial overfitting, instead of modelling how the substitution process is influenced by the protein structure.

More recently, several groups proposed biophysical models that predict how the site-specific evolutionary properties, such as the sequence variability and substitution rate, depend on the structural properties of the site in the native state, i.e. the biologically active state in which the protein has a well-defined average folded conformation (except for intrinsically disordered proteins, see Uversky 2011). These models belong to two main classes.

(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). These models predict this deviation through linear response theory applied to the Elastic Network Model (ENM; Tirion 1996, Atilgan et al. 2001) that approximates the native ensemble of the protein, and 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.

Some time ago, we developed Stab-CPE models that compute site-specific substitution processes that can be used in phylogenetic inference based on a simplified model of the stability of the native state against both unfolded and misfolded states (Bastolla et al. 2006, Arenas et al. 2015). These substitution models retain high computational efficiency due to the approximation that different sites evolve independently, and they only fit from the data one global parameter that represents the strength of natural selection and 19 global amino acid frequencies as in the +F empirical substitution models of protein evolution. Our Stab-CPE models assigned better AIC and BIC scores to real multiple sequence alignments than empirical substitution models that are the same at all positions (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 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 misfolding are worse supported by empirical data, since they assign lower likelihood to real protein sequences and they predict protein sequences that are on the average more hydrophobic than natural proteins and are not stable against misfolding (since hydrophobicity stabilizes compact conformations belonging to both the native and misfolded states), while models that consider stability against misfolding score better under all these points of view (Arenas et al. 2015). The Stab-CPE models that consider misfolding 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 models 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 of the protein sequence that do not modify the structure are more likely tolerated, which suggests that negative selection targets protein structure more strongly than protein sequence. Nevertheless, 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 (Pascual-Garcia et al. 2019), again suggesting that positive natural selection targets protein structure more strongly than protein sequence.

The success of the ENM 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 processes. We find 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 provides higher likelihood to multiple sequence alignment (MSA) that include at least one protein of known structure. However, different from Stab-CPE models, the sequences represented by the stationary sequences of Str-CPE models on the average do not have a stable folded state.

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 folding stability. These findings suggest that the combined models represent selection on proteins more realistically than both models that ignore site-specificity and our previous Stab-CPE models. They also support the view that selection on protein structure conservation, which we represent through the parameter Λ_{str}, is stronger than selection on protein folding stability Λ_{stab}.

We refer collectively to our models as SSCPE (structure and stability constrained protein evolutioni) models. Next, we assessed the usefulness of the SSCPE models, finding that they infer more accurate phylogenetic trees than those inferred under the empirical substitution models, in the sense that the SSCPE-based trees are more similar to the phylogenetic trees of very distantly related proteins inferred with structural information.

Phylogenetic trees inferred under empirical models with among-sites substitution rate variation according to a gamma distribution (+G) sometimes show higher likelihood than the SSCPE models, in particular for alignments that contain more than 100 sequences. However, we show here that this higher likelihood is explained by their longer branches, which are positively correlated to the log likelihood score. This advantage disappears if we regularize the maximum likelihood (ML) fit of branch lengths, jointly optimizing the log likelihood and the sum of branch lengths, in the spirit of the minimum evolution principle and the maximum posterior probability principle. The assessment of the inferred trees presented here supports the view that this regularization, which we call regularized maximum likelihood and minimum evolution (REGMLAME), improves the results of phylogenetic fitting, as it is well-known for other types of regression.

We implemented the SSCPE models, and their application to phylogenetic tree reconstruction, into a computer framework named SSCPE.pl that is freely available from https://github.com/ugobas/SSCPE.

Materials & Methods

Model of structure change upon mutation

To model the deformation of the protein structure produced by a mutation, we adopt the linearly forced elastic network model (LfENM) proposed by Echave (2008). This model represents a mutation as a force F^{mut} that perturbs the native structure of the wild-type protein and it predicts the associated structural deformation vector Δr_i as the linear response to this force of the elastic network model (ENM) that represents the equilibrium dynamics of the native state of the protein, i.e. the biologically active state, where the protein atoms perform small fluctuations around the average native structure represented in the PDB:

$$\Delta r_i = \sum_j (H^{-1})_{ij} F_j^{\text{mut}} \tag{1}$$

Here the indexes i and j label the $3L$ Cartesian coordinates of the C_α atoms that represent the L structured residues of the protein, H^{-1} is the inverse of the Hessian matrix of the second derivatives of the energy function (H^{-1} describes the dynamical correlations of the ENM) and F^{mut} is the force (perturbation) generated by the specific mutation.

Different from the original LfENM approach, which adopts an ENM that only moves the $3L$ Cartesian coordinates of the alpha carbons, here we adopt the torsional network model (TNM), an ENM in torsion angle space (Mendez and Bastolla 2010) implemented in the program tnm

available at https://github.com/ugobas/tnm. The TNM moves all heavy atoms of the protein using as degrees of freedom a set of its torsion angles and constrains all other internal coordinates to their native values. In this work, we adopt the option that moves only the phi and psi backbone torsion angles, amounting to $2L$ degrees of freedom, fewer than the $3L$ ones of Cartesian ENM.

The TNM considers a contact-based energy function. All ENMs approximate the energy function as a quadratic function of the displacements, which is only valid when these displacements are small. In the TNM approach we add a quadratic potential that constraints the fluctuations of the torsion angles, $\kappa_\psi K \sum_a (\Delta\psi_a)^2$, where a labels the torsion angles and K is an overall force constant. This additional potential allows to apply the TNM up to larger displacements representative of the thermal motion of the protein (Dehouck and Bastolla 2021). We adopt $\kappa_\psi = 0.2$, a value that maximizes the correlation between predicted atomic displacements and those observed in NMR ensembles (based on data from Dehouck and Mikhailov 2013).

The ENM provides an analytic description of the dynamics in the native state in terms of the normal modes, which are independent and collective harmonic motions of all protein atoms. Each normal mode α is characterized by its direction v_i^α and by the frequency ω_α of the harmonic motions, 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). The number of normal modes equals the number of degrees of freedom excluding rigid body motions, i.e. $3L-6$ for the Cartesian ENM and $2L-2$ for the TNM. We express the structure change Eq.(1) in the base of the normal modes as

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

where α labels the normal mode, j labels the Cartesian components of all heavy atoms (not just alpha carbons), 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.

Representation of a mutation as a perturbing force

The LfENM models the force as the sum of components directed along the contacts that join the mutated residue m to its neighbours in the native state. These components are 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 explicitly on the amino-acids that are interchanged.

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 (this value was also determined by maximizing the correlations between predicted and observed atomic fluctuations). 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). 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 equilibrium distance d between the amino acid pair, $d(a, A_k) \to d(b, A_k)$. Each of these changes generates the mutation force that is modelled as

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

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, $d(b, A_k) > d(a, A_k)$. These parameters are derived from protein statistics. $s(a)$ is 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 optimal distances $d(a, b)$ were determined as the most frequently observed distance in a non redundant subset of the PDB clustered at the 25% sequence identity level. They are available as Supplementary material. 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,

$$F_m^{amb} = -\sum_k C_{mk} F_k^{amb}. \tag{4}$$

Parameter optimization

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

$$F^{\text{mut},amb} = K(R_0)^g \left[W_{\text{size}} F^{\text{size},amb} + W_{\text{stab}} F^{\text{stab},amb} + W_{\text{dist}} F^{\text{dist},amb} \right]. \tag{5}$$

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 is not correlated

with the native dynamics of the protein (Dos Santos et al. 2013). This null model is based on linear response theory, and it assumes that the contribution of each mode to the conformation change c_α^2 (i.e. the squared projection of the conformation change on normal mode α) is proportional to its contribution to thermal dynamics given by $1/\omega_\alpha^2$. We verified this null model on a large set of proteins with same sequence and different structures (Dos Santos et al. 2013), finding that c_α^2 and $1/\omega_\alpha^2$ are indeed correlated for all pairs of conformations with identical sequence, but for functional conformational changes the low frequency normal modes contribute more than expected based on the null model, with the effect of reducing the harmonic energy barrier that opposes the conformation change (see also Dehouck and Bastolla 2021).

From Eq.(2), we see that the null model requires a perturbation F^{NULL} whose projection on each normal mode α is $F_\alpha'^{\mathrm{NULL}} \equiv \sum_j v_j^\alpha F_j^{\mathrm{NULL}} = r_\alpha \omega_\alpha$, with r_α independent of the normal mode amplitude. 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 is directed 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

$$\mathrm{MSD} = \frac{1}{M}\sum_i m_i \Delta r_i^2 = \mathrm{MSD}^{\mathrm{mut}} + \mathrm{MSD}^{\mathrm{nomut}} + \mathrm{MSD}^{\mathrm{cross}} \tag{6}$$

$$\mathrm{MSD}^{\mathrm{mut}} = \sum_\alpha \left(\frac{F_\alpha'^{\mathrm{mut}}}{\omega_\alpha^2}\right)^2, \ \mathrm{MSD}^{\mathrm{nomut}} = \sum_\alpha \frac{1}{\omega_\alpha^2}, \ \mathrm{MSD}^{\mathrm{cross}} = 2\sigma_p \sqrt{\sum_\alpha \left(\frac{F_\alpha'^{\mathrm{mut}}}{\omega_\alpha^3}\right)^2}, \tag{7}$$

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, $F_\alpha'^{\mathrm{mut}}$ are the projections of the 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 MSD produced by the mutation force (mut), which depends on the mutation weights. (2) The null model fluctuations (nomut). (3) The cross term (cross) 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.(7) 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 fit 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. We acknowledge that fitting σ_p may overestimate the accuracy of the model. For this reason, in this work we carefully evaluated the mutation model based on evolutionary data.

An additional problem is that sometimes the observed RMSD derives from functionally distinct conformations rather than from 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 we recall it 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, adopting the same definition of contact as in the TNM. 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 $|i - j| \leq 2$, while in the TNM model we set $C_{ij} = 1$ 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 $G_{\rm nat}(C^{\rm nat}, A) \approx \sum_{i<j} C^{\rm nat}_{ij} U(A_i, A_j)$. For the unfolded state, we neglect the 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),

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

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 remains 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, which we pre-calculate when the program is initialized in order to accelerate the computations. This is important to achieve computational efficiency since the program must iteratively calculate averages over the site-specific distributions. In the initialization step, the program reads a file where we stored all the contact

matrices of a non-redundant subset of the PDB, with only one representative for each cluster at 25% sequence identity, and it computes the contact statistics as described below. This step takes 45 sec for the longest protein with $L = 1045$.

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 $\langle C_{ij} \rangle$. To reduce the number of parameters, we assume that $\langle C_{ij} \rangle$ depends only on the sequence distance $|i-j|$ and on chain length L, $\langle C_{ij} \rangle = a(L) f(|i-j|)$. We determine $f(|i-j|)$ as the probability that two residues with sequence distance $i-j$ are in contact, and 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. We sum over j in contact ij and over l in contact kl: $\sum_{ik} \left(\langle n_i n_k \rangle - \langle n_i \rangle \langle n_k \rangle \right) U_i U_k$, 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. We sum over kl: $\sum_{ij} \left(\langle C_{ij} N_c \rangle - \langle C_{ij} \rangle \langle N_c \rangle \right) U_{ij} U$, 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 estimates, we obtain the free energy difference between the native and the non-native states as

$$\Delta G(C^{\mathrm{nat}}, A) = G_{\mathrm{nat}} - kT \left(\mathrm{e}^{-G_{\mathrm{misf}}/kT} + \mathrm{e}^{-G_U/kT} \right) , \tag{8}$$

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 by $k_B T$ (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 Fisher's fixation probability. The latter distribution is proportional to the exponential of the logarithm of fitness $\varphi(A)$ multiplied times a selection parameter Λ conceptually 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, $\prod_i P^{\mathrm{mut}}(A_i)$.

Although the formulation in terms of a mutation process is elegant and is consistent with evolutionary theory, in practice it is difficult to cleanly separate selection at the amino acid level from mutation at the nucleotide level. The global (site-unspecific) amino acid frequencies reflect both mutational factors (number of codons and their stationary frequency) and unspecific selective factors (metabolic cost of the amino acids or their suitability to occurr in natural proteins) that are not taken into account by our model. To express this, we rename $P^{\mathrm{mut}}(a)$ as $P^{\mathrm{glob}}(a)$.

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 phylogenetic inference. If the fitness only depends on the properties of each individual protein site i, in other words if 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

$$P(A_1 \cdots A_L) = \prod_i P_i(A_i, \Lambda) \tag{9}$$

$$P_i(A_i, \Lambda) = \frac{1}{Z_i} P^{\mathrm{glob}}(A_i) \exp(\Lambda \varphi_i(A_i)) \tag{10}$$

with $Z_i = \sum_a P^{\mathrm{glob}}(a) \exp(\Lambda \varphi_i(a))$.

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:

$$e^{\Lambda \varphi_i^{\mathrm{MF}}(a)} = \sum_{\{A\}, A_i = a} \prod_k P^{k,\mathrm{MF}}(A_k, \Lambda) e^{-\Lambda \Delta G(\{A\})} \,. \tag{11}$$

The MF distribution can be computed through an iterative algorithm whose complexity scales as the square of the number of sites and that converges rapidly, taking advantage of the precomputation of the statistics of misfolded conformations. We perform this 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 ($\Delta\Delta G$) of the mutant protein $A_i^{\mathrm{WT}} ia$ in which amino acid A_i^{WT} at site i of the wild type sequence is changed to a:

$$\varphi_i^{\mathrm{WT}}(a) = -\Delta\Delta G(A_i^{\mathrm{WT}} ia) \tag{12}$$

The WT model is mathematically simpler than the MF model, since it avoids the self-consistent calculation of the selection factors, and it produces results that agree well with the amino acid frequencies of extant protein families (Arenas et al. 2019). In fact, 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:

$$\varphi_i^{\mathrm{RMSD}}(a) = -\mathrm{RMSD}(A_i^{\mathrm{WT}} ia)\,. \tag{13}$$

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:

$$\varphi_i^{\mathrm{DE}}(a) = -\Delta E^{\mathrm{harm}}(A_i^{\mathrm{WT}} ia)\ , \tag{14}$$

$$\Delta E^{\mathrm{harm}} = (\Delta r, H\Delta r) = \left(F^{\mathrm{mut}}, H^{-1}F^{\mathrm{mut}}\right) = \sum_\alpha \left(\frac{F_\alpha'^{\mathrm{mut}}}{\omega_\alpha}\right)^2\ , \tag{15}$$

where F^{mut} is the perturbation generated by any specific mutation and $F_\alpha'^{\mathrm{mut}}$ is its projection on normal mode α. The new version of the program tnm (Mendez and Bastolla 2010), 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.

Str-CPE for the wild-type residue

Note that the RMSD and DE deformations generated when the protein is mutated to the amino acid A_i^{WT} in the wild-type sequence in the PDB are zero. However, using this value as fitness would imply that the amino acid in the PDB has the maximal fitness under the structural point of view. Besides being unrealistic, this assumption produces very biased results under the point of view of phylogenetic inference that we inferred in this paper: The sequence represented in the PDB receives high likelihood, but the model infers phylogenetic trees that are very different from the reference tree, a clear sign of overfitting.

To address this problem, we compute it as the average deformation of the mutations to A_i^{WT} that start from all possible initial amino acids a at position i, assuming that they are distributed according to the stationary distribution $P_i(a)$ (with $P^{\mathrm{glob}}(a) = 1/20$ to avoid interferences with the estimate of this parameter) and that the deformation produced by the mutation $a \to A_i^{\mathrm{WT}}$ is the same as $A_i^{\mathrm{WT}} \to a$:

$$\varphi_i^{\mathrm{STR}}(A_i^{\mathrm{WT}}, \Lambda) = \frac{\sum_{a\neq A_i^{\mathrm{WT}}} \varphi_i^{\mathrm{STR}}(a) \mathrm{e}^{\Lambda\varphi_i^{\mathrm{STR}}(a)}}{\sum_{a\neq A_i^{\mathrm{WT}}} \mathrm{e}^{\Lambda\varphi_i^{\mathrm{STR}}(a)}} \tag{16}$$

where STR represents either the RMSD or the DE model. Note that in this approach the fitness factors depend on the selection strength Λ.

Combined Stab-CPE and Str-CPE models

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

$$P_i(a) = P^{\text{glob}}(a) \exp\left(\Lambda_{\text{stab}} \varphi_i^{\text{StabCPE}}(a) + \Lambda_{\text{str}} \varphi_i^{\text{StrCPE}}(a)\right) \tag{17}$$

The SSCPE models have two selection parameters Λ_{stab} and Λ_{str}.

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.

Global frequencies

The SSCPE models take as parameters a vector of global amino acid frequencies $P^{\text{glob}}(a)$ and the selection parameter Λ. Given the latter, we determine the global frequencies for each SSCPE model and each value of Λ so that the frequency of each amino acid summed over all sites is equal to its frequency in the MSA:

$$P^{\text{glob}}(a) \sum_i \frac{1}{Z_i} \mathrm{e}^{\Lambda \varphi_i(a)} = \sum_i f_i^{\text{MSA}}(a) \tag{18}$$

In this way, P^{glob} depends on Λ. Note that the normalization factor $Z_i = \sum_a P^{\text{glob}}(a) \mathrm{e}^{\Lambda \varphi_i(a)}$ depends on P^{glob}, so the equation must be solved iteratively. The above equation fits 19 parameters in such a way that the average frequencies are exactly equal in the model and in the MSA, and it is analogous to the +F option in phylogenetics.

Selection parameter and regularization

The most important parameter of the site-specific substitution processes is the global selection parameter Λ. We explored two criteria for optimizing Λ:

1. Maximizing the log-likelihood of the site-specific amino acid frequency observed in the input MSA $f_i^{\text{MSA}}(a)$ with respect to the model:

$$\text{LL}(\Lambda) = \sum_i \sum_a f_i^{\text{MSA}}(a) \log\left(P_i(a, \Lambda)\right) \tag{19}$$

2. Minimizing the symmetric Kullback-Leibler divergence between the input MSA and the model:

$$\text{KL}^{\text{symm}}(\Lambda) = \text{LL}(\Lambda) + \sum_i \sum_a P_i(a, \Lambda) \log\left(f_{\text{MSA}}(a)\right) \tag{20}$$

We apply a cut-off to the MSA so that its minimum value is $f_i^{\mathrm{MSA}}(a) \geq \epsilon$.

We obtained better results with the KL criterion, which produces higher values of Λ and stationary frequencies that on the average represent more stable proteins. Therefore, we only present results obtained with this criterion.

Since the Str-CPE fitness φ_i^{STR} and the global amino acid frequencies $P^{\mathrm{glob}}(a)$ depend on Λ, it is cumbersome to optimize the above scores analytically determining the value of Λ where the derivative vanish. We maximize them numerically through quadratic optimization, which is fast and accurate.

Tykhonov regularization with specific heat criterion

As with all regression problems, it is important to regularize the fitted parameters in order to limit overfitting. In our experience, regularization is always beneficial since it reduces the cross-validation error of the fits and often eliminates non physical values, such as negative spring constants in the context of the ENM.

We apply the Tykhonov regularization, also called ridge regression (Hoerl and Kennard 1970), and we minimize the sum of the KL divergence plus a term that limits the value of the fitted selection parameter Λ, i.e. we minimize

$$\mathrm{KL}(\Lambda) - R\Lambda^2, \tag{21}$$

where R is the regularization parameter. Note that R is not a new fitted parameter, on the contrary it worsens the score achieved through the fit. 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. Given R, we determine Λ by minimizing the score either numerically or analytically, as the value at which the derivative vanishes.

In (Dehouck and Bastolla 2017) one of us noted the analogy between regularization and thermodynamics, in the sense that the error of the fit (here KL) can be interpreted as an energy, which increases when increasing the regularization parameter R, interpreted as a temperature. Therefore, the derivative of the error with respect to R, which can be shown to be positive, can be interpreted as a specific heat C_V:

$$C_V \equiv \frac{\partial \mathrm{KL}(\Lambda, R)}{\partial R} > 0 \tag{22}$$

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.Our program samples several values of R, computing the optimal value of Λ and the C_V numerically for each of them. Finally, it chooses the value of R with maximum C_v and refines its value through quadratic optimization.

Site-specific sequence entropy and substitution rate

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

$$S_i = -\sum_{a=1}^{20} P_i(a) \log(P_i(a)). \tag{23}$$

The site-specific substitution rates of the evolutionary models are computed as the weighted average of the substitution rate matrix $Q_i(a,b) = E_i(a,b)P_i(b)$,

$$R_i = \sum_{a \neq b} P_i(a)P_i(b)E_i(a,b) \tag{24}$$

Exchangeability matrix

Up to now, we did not describe how we define the exchangeability matrices $E_i(a,b)$ at each site i that are necessary for computing the site-specific substitution process $Q_i(a,b)$:

$$Q_i(a,b) = E_i(a,b)P_i(b) \tag{25}$$

$$P_i(a) = \frac{1}{Z_i} P^{\text{glob}}(a) e^{\Lambda \varphi_i(a)} \tag{26}$$

For computing the exchangeability matrices, we consider the following alternatives.

Halpern-Bruno model

The simplest choice consists in choosing the same global exchangeability matrix for all the sites, $E_i(a,b) \equiv E^{\text{glob}}(a,b)$. We call this option "noHB". The alternative option, which we call "HB", consists in applying the Halpern-Bruno model (Halpern and Bruno 1998) that expresses $E_i(a,b)$ as the product of a global substitution process $E^{\text{glob}}(a,b)$ that is the same at all sites times a site-specific fixation probability. To this end, following Halpern and Bruno we interpret $\varphi_i(a)$ as the logarithm of fitness and compute the fixation probability through the Fisher's formula, which results in

$$E_i(a,b) = E^{\text{glob}}(a,b) \left(\frac{\Lambda\left(\varphi_i(b) - \varphi_i(a)\right)}{\frac{1}{Z_i}\left(e^{\Lambda\varphi_i(b)} - e^{\Lambda\varphi_i(a)}\right)} \right) \tag{27}$$

Flux model

To determine the global exchangeability matrix E^{glob}_{ab}, we also have two alternatives. The simplest one, which we call "noFL", consists in choosing E^{glob}_{ab} equal to one of the empirical exchangeability matrices such as JTT (Jones et al. 1992), LG (Le and Gascuel 2004) or WAG (Whelan and Goldman 2001): $E^{\text{glob}}_{ab} = E^{\text{emp}}_{ab}$.

The alternative option, which we call "FL", consists in assuming that the empirical substitution matrices have to be trusted more in terms of the flux between any amino acid pairs that they determine, because this is the quantity that they fitted from empirical data. Thus, in the FL option we impose that the flux of any two amino acids a and b determined by the model, $R_i(a,b) = P_i(a)P_i(b)E_i(a,b)$, and averaged over all sites i equals the flux measured in the empirical model for each pair of amino acids a and b, $\sum_i R_i(a,b)/L = (P^{\text{emp}}(a)P^{\text{emp}}(b)E^{\text{emp}}(a,b))$. We get

$$P^{\text{glob}}(a)P^{\text{glob}}(b)E^{\text{glob}}(a,b) = \frac{P^{\text{emp}}(a)P^{\text{emp}}(b)E^{\text{emp}}(a,b)}{\frac{1}{L}\sum_i \frac{1}{Z_i} e^{\Lambda(\varphi_i(a)+\varphi_i(b))} \frac{\Lambda(\varphi_i(b)-\varphi_i(a))}{e^{\Lambda(\varphi_i(b)} - e^{\Lambda(\varphi_i(a)}}} \tag{28}$$

for the HB model, and

$$P^{\text{glob}}(a)P^{\text{glob}}(b)E^{\text{glob}}(a,b) = \frac{P^{\text{emp}}(a)P^{\text{emp}}(b)E^{\text{emp}}(a,b)}{\frac{1}{L}\sum_i \frac{1}{Z_i} e^{\Lambda(\varphi_i(a)+\varphi_i(b))}} \tag{29}$$

for the noHB model.

Normalization of the substitution rate

Finally, another alternative arises from a technical feature of the program RAxML-NG, which internally normalizes the exchanggeability matrix at each site so that its substitution rate is always one: $\sum_{ab} P_i(a)P_i(b)E_i(a,b) \equiv 1$. We explored two alternatives: either we allow the program to internally normalize the exchangeability (rate0) or we input to RAxML-NG a scale factor for each site so that, after the normalization, the exchangeability matrix is multiplied times this scaling factors that restores the substitution rate predicted by the SSCPE model (rate1).

As a result, we have eight different exchangeability models, corresponding to the combinations of HB/noHB, FL/noFL and rate1/rate0. We explored all these combinations, and we found the best results on the average with the options HB, FL and rate1, although for some score, model or data other combinations may yield better results. The presented results correspond to these options, unless otherwise state.

Choice of the empirical substitution model

The empirical substitution model can either be chosen by the user or be internally optimized by the program, maximizing the following approximate likelihood function that does not consider the tree topology and assumes short branch lengths. The likelihood of the substitution between amino acid a and b in a pair of sequences that evolved through a time t is $P(a)\left(\mathrm{e}^{tQ}\right)_{ab}$. At first order in t, this is approximately $tP(a)E_{ab}P(b)$ if $b \neq a$ and $P(a)(1+tQ_{a,a})$ otherwise, and its logarithm is approximately $\log(t)+\log(P(a))+\log(P(a))+\log(E_{ab})$ or $\log(P(a))+tQ_{a,a}$. Summing over all pairs of aligned amino acids, whose number is $N(a,b)$ and denoting $N(a)$ the number of occurrences of amino acid a, we get:

$$\begin{aligned} \text{loglik} \;\approx\; & \sum_a \sum_{b\neq a} N(a,b)\log(E_{ab}) \\ + \; & \sum_a N(a)\log(P(a)) + \log(t)\sum_a (N(a)-N(a,a)) + t\sum_a N(a,a)P(a)E_{aa} \end{aligned} \tag{30}$$

The value of t that maximizes Eq.(31) is readily determined where the derivative vanishes, $t = -\sum_a (N(a)-N(a,a))/\sum_a N(a,a)P_aE_{aa}$ with $E_{aa} = -\sum_{b\neq a} E_{ab}P(b)$. Plugging it into Eq.(31) we obtain the approximate value of the log likelihood for each of the three empirical substitution matrices implemented in our program (JTT, LG and WAG), from which we choose the optimal one.

Assessment of the SSCPE exchangeability matrix

For given SSCPE model (8 types) and for the chosen empirical exchangeability matrix, the SSCPE model can adopt 8 combinations of 3 binary options of the site-specific exchangeability matrices (HB/noHB, FL/noFL, rate0/rate1). The program considers all of them and computes for each site the approximate likelihood given by Eq.(31). In this way, it assesses which combination provides the highest value of the approximate likelihood. This information is printed in the output file with extension SSCPE.summary.dat. The best combination is almost always HB+rate1, often in

combination with FL and sometimes in combination with noFL. Through this score, we can also explore which SSCPE model presents the highest likelihood. However, not always the predicted model receives the highest phylogenetic likelihood in computations by RAxML-NG.

Implementation

We implemented the computer program Prot_evol that computes eight SSCPE models and prints their site-specific amino acid frequencies, summary statistics and site-specific substitution rates and sequence entropies.

We also implemented the framework SSCPE.pl that runs Prot_evol or reads its results if they are present in the current folder and infers the maximum likelihood phylogenetic tree with the program RAxML-NG (Kozlov et al. 2019) for the required SSCPE model (possible choices: MF, WT, DE, RMSD, DEMF, DEWT, RMSDMF, RMSDWT, default: the model that obtains the best non-phylogenetic likelihood according to Prot_evol) and the required options of the exchangeability matrix (8 combinations of the binary options HB/noHB, FL/noFL, rate0/rate1, default: HB+FL+rate1).

We implemented the perl program K2_mat.pl, using as starting point the program K_mat.pl (Soria-Carrasco et al. 2007), for computing the similarity between the inferred trees and a reference tree. Besides the K score based on one-parameter branch length fit and the topological score by Robinson and Foulds (1981) computed by K_mat.pl, K2_mat.pl also computes the K2 score based on two-parameter branch length fit that is more appropriate for quantifying the similarity between sequence-based and structure-based trees (see Results).

Finally, we implemented the master script script_run_SSCPE.pl that, for an input MSA, runs a combination of several SSCPE models with several options of exchangeability matrix and several empirical substitution models with or without the options +F and +G. The script submits the computations to a queue system, which must be also modified by the user. After all computations are done, the master script must be called again with the option -examine and with a reference tree in Newick format (-tree) for comparing the trees inferred with each model to the reference tree with the program K2_mat. The results of these comparison, the log-likelihood computed by RAxML-NG, the sum of branch lengths and the REGMLAME scores are printed in a table called Results.[].txt.

Further details on the usage of the framework are presented in GitHub (see Data Avaibility).

Concatenated alignments and multiple structures

The program Prot_evol receives as input a protein MSA (-ali), and either an individual pdb file (-pdb) or a list of pdb codes (-pdblist) together with the folder where the pdb files are stored (-pdbdir). The corresponding structures may align to different regions of the MSA, which may be obtained by concatenating multiple proteins.

The program reads the MSA and the pdb files. Then it aligns each pdb sequence against each sequence of the MSA. If the best match has < 0.5 sequence identity the pdb structure is discarded, otherwise the program adopts the corresponding alignment ali and computes the folding free energy $\Delta G(A_s, C_p, \text{ali})$ and the fitness $f_{ps} = 1/(1 + \exp(\Delta G))$ for each sequence A_s in the MSA aligned to the structure C_p. Typically, the sequence of the pdb protein $s(p)$ has $\Delta G(A_{s(p)}, C_p, \text{ali}) \ll -1$ and $f_{ps(p)} \approx 1$. We discard the structures that are predicted to be unstable ($\Delta G(A_{s(p)}, C_p, \text{ali}) \geq 0$)

or stabilize too few sequences ($\sum_s f_{ps}$ is lower than the minimum beltween 4.5 and 0.45 times the number of sequences in the MSA). Otherwise, the structure is used for computing the fitness factors $\varphi_i^p(a)$ of the models MF, WT, DE and RMSD at the sites i of the MSA that are aligned with structure p and we compute the weight $w_p \propto \sum_s f_{sp}$ for obtaining the weighted average of $\varphi_i^p(a)$ at sites with multiple structures.

Subsequently, we discard from the MSA the sequences for which the stability score $\sum_p f_{sp}$ is less than 0.9, we remove the columns of the MSA that do not have any residue and relabel the columns of the MSA. Each such column i can either have a structural model and the corresponding fitness $\varphi_i(a) = \sum_p w_p \varphi_i^p(a)$ or it has $\varphi_i(a) \equiv 0$ and it is described by the global amino acid frequencies $P^{\text{glob}}(a)$.

For each SSCPE model, the program considers several values of the regularization parameter R, optimizes the selection coefficients Λ numerically by minizing $\text{KL}(\Lambda) + R\Lambda^2$, computes numerically the derivative $C_V = \partial\text{KL}/\partial\Lambda$, finds the value of R that maximizes C_V and recomputes the optimal Λ and the site-specific amino acid frequency $P_i^{\text{SSCPE}}(a)$, which it prints in a file called [ ].SSCPE.[MODEL].AA_profiles.txt for the MSA columns that are structured, together with the global frequencies $P^{\text{glob}}(a)$ corresponding to that model.

Finally, for each model the program computes the site-specific exchangeability matrices according to the four models HB/noHB and FL/noFL, which depend on the site-specific frequencies, the global frequencies $P_{\text{glob}}(a)$ and the empirical substitution model, computes the corresponding approximate likelihood, and prints the four site-specific rates in a file called [ ].SSCPE.[MODEL].rate_profile.dat. Since they can be recomputed rapidly, the exchangeability matrices are not printed to avoid to write large files. They are recomputed by SSCPE.pl for the required model and exchangeability options, then they are input to RAxML-NG as partitions and the correponding files are deleted upon completion of the phylogenetic computation.

Results

Predicting the amount of structure change caused by a mutation

We determined the weights W_{size}, W_{stab} and W_{dist} in Eq.(5) by numerically fitting Eq.(7) over a training set of mutants with known structure. We omitted structures that are almost identical to the wild-type (WT), requiring that the root mean square deviation (RMSD) is at least 0.15Å. This condition is rather tolerant, but more restrictive conditions reduce the data set too much. 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 with RMSD> 0.15Å and standard deviation 0.28Å. We randomly split the corresponding protein families into training set (156 pairs) and test set (39 pairs, i.e. 20%).

To avoid the influence of the binary parameter σ_p, which we fit 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 the value of σ_p does not influence the predicted RMSD produced by the mutation, which we need for modelling the substitution process. The parameters that result from the optimization are $W_{\text{size}} = 3.05$, $W_{\text{dist}} = 15.6$ and $W_{\text{stab}} = 2.39$, yielding the correlation coefficient between observed and predicted RMSD $r_{\text{WT}} = 0.86$ and root mean square error $\text{RMSE}_{\text{WT}} = 0.20$Å over the training set from wild-type to mutant, and $r_{\text{mut}} = 0.75$, $\text{RMSE}_{\text{mut}} = 0.22$Å from mutant to wild-type (the two sets 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 $r_{\mathrm{WT}} = 0.95$ and $r_{\mathrm{mut}} = 0.97$ and $\mathrm{RMSE}_{\mathrm{WT}} = 0.16$Å and $RMSE_{\mathrm{mut}} = 0.11$Å (see Supplementary Figure S1). These correlations were high because the test set contains only one pair with large RMSD, which is more difficult to predict. Considering only the mutation model we obtained poorer correlations ($r_{\mathrm{WT}} = 0.54$, $r_{\mathrm{mut}} = 0.28$ for the training set and $r_{\mathrm{WT}} = 0.68$, $r_{\mathrm{WT}} = 0.65$ for 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 structural effect model, independent of the particular values of the parameters, are important. They are shared with the random force model proposed by 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}.

Assessment of the predicted amino acid frequencies

In this work, we compare our two previous Stab-CPE models MF (in which stability of amino acid a at site i is estimated self-consistently considering the amino acid frequencies at all other sites) and WT (in which stability is estimated considering the amino acids at all other sites in the wild-type sequence) against the new Str-CPE models RMSD and DE (in which we compute the RMSD and DE produced by amino acid a at site i adopting the ENM of the wild-type protein) and their four combinations RMSDMF, RMSDWT, DEMF and DEWT. They are summarized in table 1. We also consider the global (site-unspecific) amino acid frequencies determined by setting the selection parameters to zero (GLOB).

In Fig.2, we compare the results obtained with the site-specific amino acid frequencies computed with the 8 different SSCPE models on the MSA of 203 monomeric proteins. For all models, the global frequencies and the Λ parameter were determined minimizing the average of the symmetric Kullback-Leibler (KL) divergence between the amino acid frequencies of the model and the observed MSA at the regularization value determined with the "maximum specific heat criterion" (see Methods). However, setting a constant regularization value does not modify the qualitative

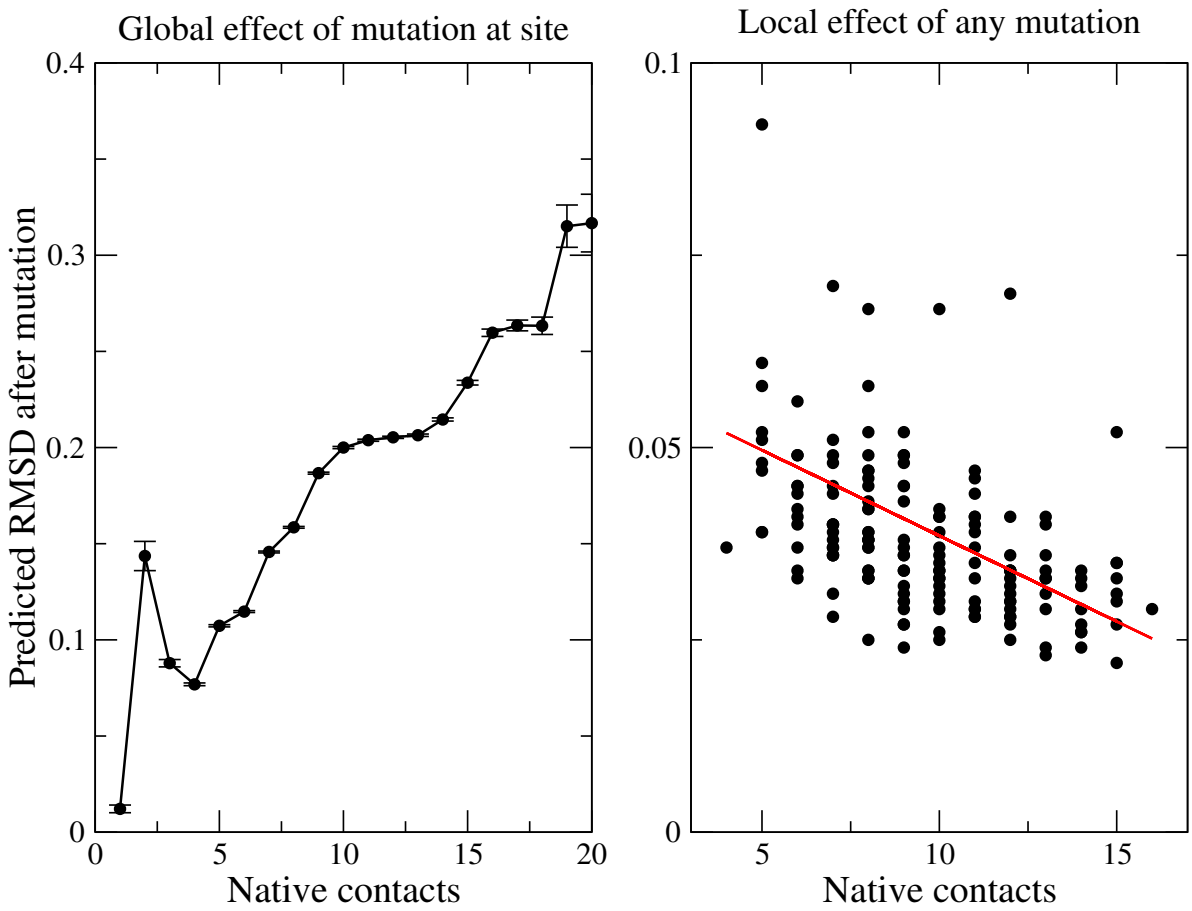

Figure 1: Effect of the mutation model as a function of the number of contacts of the mutated site (A) and the number of contacts at the position where the structure change is monitored (B).

results. We also consider the model GLOB obtained with the global amino acid frequencies setting $\Lambda = 0$.

One can see from Fig.2A that the average likelihood of the MSA can be ranked as GLOB $<$ DE $<$ RMSD $<$ MF $<$ WT $<$ DEMF $\approx$ RMSDMF $<$ DEWT $\approx$ RMSDWT, i.e. Stab-CPE models generally present higher likelihood than the Str-CPE models and the combined models present higher likelihood, the highest being for the RMSDWT model. The same ranking is obtained with the KL divergence (Fig.2B), where the signs of the inequalities are inverted. Not surprisingly, Str-CPE models represent unstable proteins, since their average estimated folding free energy ΔG is positive, however the combined models are more stable than Stab-CPE models and their average ΔG is almost exactly the same as for the structures in the PDB (Fig.2C). Str-CPE models present lower entropy than Stab-CPE models (Fig.2D), which indicates that they are less tolerant to mutations, and the combined models show the lowest entropies. A similar message is obtained from Fig.2E, where one can see that Str-CPE models tend to deviate more than Stab-CPE models from the site-unspecific model GLOB, and from Fig.2F, where one can see that the selecion parameter associated to structure conservation (orange) tends to be larger than the one associated with

Model	Selection on stability against misfolding and unfolding	Selection on structure conservation	Selection parameters
GLOB	No	No	None
MF	Yes	No	Λ_{stab}
WT	Yes	No	Λ_{stab}
DE	No	Yes	Λ_{str}
RMSD	No	Yes	Λ_{str}
DEMF	Yes	Yes	Λ_{stab},Λ_{str}
DEWT	Yes	Yes	Λ_{stab},Λ_{str}
RMSDMF	Yes	Yes	Λ_{stab},Λ_{str}
RMSDWT	Yes	Yes	Λ_{stab},Λ_{str}

Table 1: Definition of the SSCPE models studied in this work

folding stability (black). All these results suport the view that structure conservation is a stronger selective constraint than conservation of protein folding stability.

Site-specific entropy and substitution rates

Fig.3 shows the average site-specific sequence entropy (A), substitution rate (B) and hydrophobicity (C) as a function of the number of native contacts for the different models and for real MSA, as reported by Jimenez-Santos et al. (2018). As previously observed by Jimenez-Santos et al. (2018b), the substitution rate is strongly influenced by the model of the exchangeability matrix. We tested the Halpern-Bruno exchangeability model (HB) versus a model without fixation probability (noHB), and the FLUX model in which, for each pairs of amino acids, their flux averaged across all sites is the same as the flux of the empirical substitution model versus the model noFL that adopts the exchangeability matrix of the empirical substitution model (Supplementary figure S2). The noHB option and the FLUX option tend to increase the rate at sites with many contacts, which are occupied by hydrophobic amino acids. In the main text, we show the substitution rates for the option HB noFL.

The entropy and the substitution rate of the Stab-CPE models MF and WT present a maximum for intermediate number of contacts (Fig.3A and B), indicating that these models are 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. Also the entropy and the substitution rate of the combined models present a maximum for intermediate number of contacts, although this maximum tends to occur before the one of Sta-CPE models. The combined models present lower entropies and substitution rates, i.e. they are less tolerant, although all models are quantitatively more tolerant than real MSA data.

Figure 3C represents the mean hydrophobicity of each site as a function of the number of contacts in the native state, for SSCPE 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 since this trend favors folding stability. This is even true for the Str-CPE models that do not select for folding

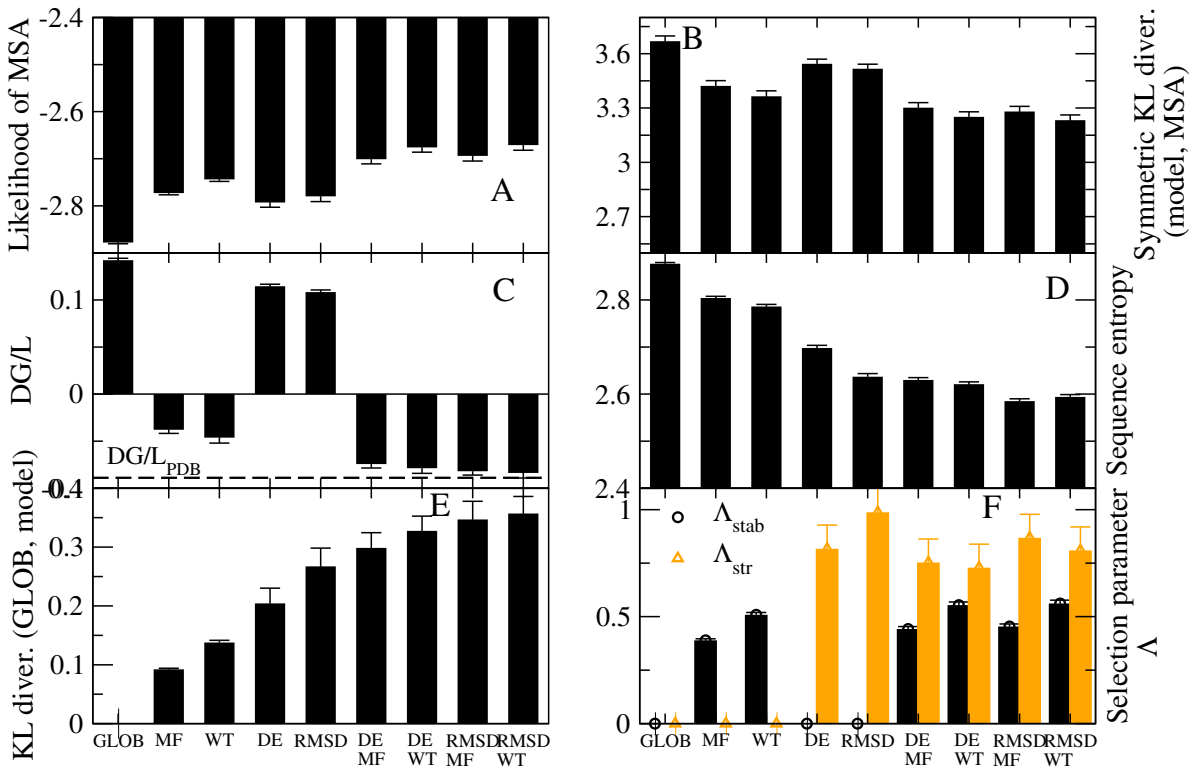

Figure 2: **Assessment of the SSCPE profiles.** Results obtained with eight SSCPE site-specific substitution models, listed in the horizontal axis, on a test set of 203 MSA with 10 to 300 sequences, at least one with known structure. A: likelihood of the observed MSA. B: Symmetric Kullback-Leibler divergence between each model and the observed MSA. C: Average folding free energy per residue of the model profile. D: Site-specific sequence entropy of the model profile. E: Kullback-Leibler divergence between the SSCPE model and the global model. F: Selection parameters. All shown quantities represent the average across all sites.

stability, although in these cases hydrophobicity responds very weakly to the number of contacts. Interestingly the maximum hydrophobicity is attained before the maximum number of contacts (11 or 12 instead of 15) both for the MSA and for the RMSDWT model, which is the one that agrees best with the observed data. The hydrophobicity of the MSA 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.

Assessment of the SSCPE models for phylogenetic tree inference

We have shown that the new combined SSCPE models are more realistic than the site-unspecific model and our previous Stab-CPE models in terms of the dependence of site-specific evolutionary properties (sequence entropy, substitution rate, hydrophobicity) on the structural properties of the site, such as the number of native contacts, as well as in terms of the likelihod of the observed amino acids in the MSA and the average estimated folding free energy. We assess here the question whether

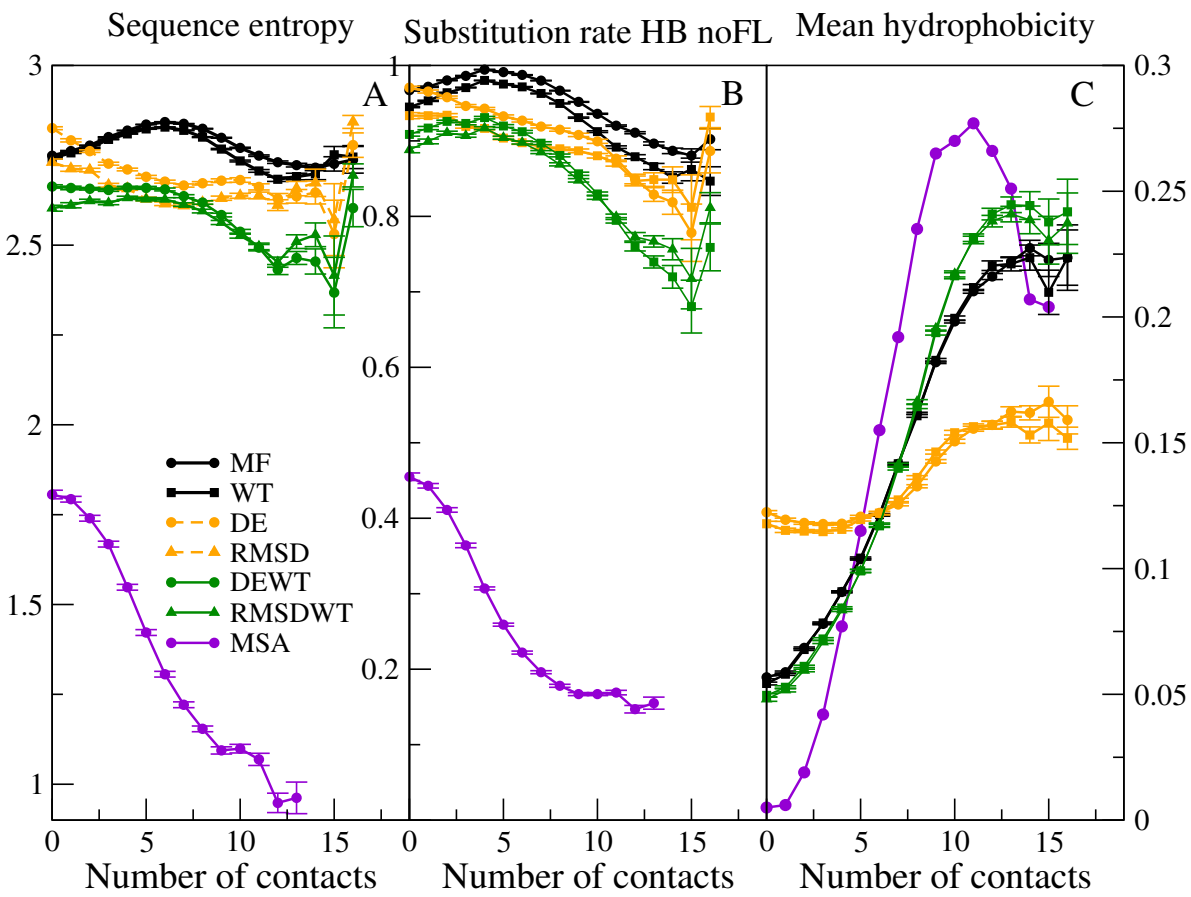

Figure 3: **Sequence variability and hydrophobicity as a function of the number of contacts.** Average site-specific sequence entropy (A), substitution rate (B) and hydrophobicity (C) as a function of the number of contacts of the site in the native state of the wild-type sequence (horizontal axis) for various SSCPE models. Data are averaged across 203 MSA with 10 to 300 sequences.

these models are also useful for inferring more accurate phylogenetic trees. This is a difficult test, for which we cannot use simulations since, for distinguishing Stab-CPE and Str-CPE models we should simulate protein evolution with selection on the native structure, and we do not know any model that implements this. Therefore, we decide to compare the trees inferred with SSCPE and empirical substitution models to the trees derived with structural information with the program PC_ali (Bastolla et al. 2023). This program adopts an integrated measure of protein similarity in sequence (sequence identity) and structure (TM score, i.e. fraction of superimposed residues (Zhang and Scolnick 2004) and contact overlap, which does not require spatial superimposition of the structures), uses it for constructing a hybrid multiple alignment, and uses the corresponding evolutionary divergences for constructing a Neighbor-Joining tree (NJ, Saitou and Nei 1987). Note that PC_ali does not adopt the models of folding stability and structural changes on which the SSCPE model is based, and it infers phylogenetic trees through the NJ algorithm while we infer trees with SSCPE adopting the ML approach. Therefore, approximating the structure-based trees

Superfamily	Sequences	Mean branch length NJ	Mean branch length ME	Corr (loglik, K2(NJ))	Corr (reglik, K2(NJ))	Corr (loglik, K2(ME))	Corr (reglik, K2(ME))
Aldolase	21	0.327	0.195	−0.91	0.87	0.33	−0.31
Globins	64	0.137	0.072	−0.61	0.54	−0.39	0.46
NADP	84	0.246	0.131	−0.22	0.19	0.53	−0.30
Ploop_C1	48	0.210	0.117	−0.08	−0.05	0.49	0.00
Ploop_C2	14	0.339	0.199	−0.43	0.49	0.31	−0.27

Table 2: **Properties of the five studied superfamilies**: Number of sequences, mean branch length of the structure-based NJ tree and ME tree, correlation between log-likelihood and K2 score from the NJ tree (negative means that higher likelihood trees are more similar to the reference), correlation between regularized likelihood and K2 score from the NJ tree (positive means that trees with better regularized score are more similar to the reference), correlation between log-likelihood and K2 score from the ME tree, correlation between regularized likelihood and K2 score from the ME tree.

with sequence-based trees is a difficult test.

We determined PC_ali trees for five subsets of distantly related proteins, related at the superfamily level, whose properties are summarized in Table 2. The reference phylogenetic trees are iluustrated in Supplementary Figures S3-S7. One can see that they are highly consistent with the functional classification of proteins in terms of Gene Ontology and Enzyme Commission numbers, which supports their good quality.

We compared the reference NJ trees determined through structural information with the ML trees determined with 12 empirical models: JTT (Jones et al. 1992); LG (Le and Gascuel 2008); WAG (Whelan and Goldman 2001), with all four combinations of the options +F and +G) and the eight SSCPE models with the options HB, FLUX and rate (we also studied other options, but present here only these ones for the sake of simplicity because they produced on the average better results). We adopted three metrics: (1) The normalized topological distance (RF; Robinson and Foulds 1981). (2) The normalized K-score dissimilarity (Soria-Carrasco et al. 2007), which assumes that the branch lengths of the comparison tree are given by the corresponding branches of the reference tree multiplied times a fitted scaling factor. The K score consists of the mean square error of the fit. With respect to the RF distance, it has the advantage that it penalizes little the discrepancies that occur at short branches. Since protein structure is known to diverge more slowly than protein sequence (Illergard et al. 2009, Pascual-Garcia et al. 2010), we expect that the scaling factor of the K score is less than one. (3) Previous work suggested that protein structures diverge very slowly when the protein function is conserved, but they diverge faster when protein function changes, because of positive selection that targets the protein structure (Pascual-Garcia et al. 2019). Therefore, we extended the K score to the K2 score that adopts two scaling factors, which are interpreted as taking care of branches where protein function is conserved and where protein function changes. For this sake, we developed the perl program K2_mat, using as starting point the program K_mat (Soria-Carrasco et al. 2007).

The results are presented in Fig.4. Fig.4A shows that the average K score is in general lower for the SSCPE models (black bars) than for the empirical models (orange bars). This difference is

more accentuated when we examine the K2 score (Fig.4B), whose value is much smaller than the K score, supporting its better quality. Indeed, in the figures we rank the models according to their K2 score. The two-parameter fit is supported by at least two of the criteria AIC, BIC or corrected AIC for almost all models and superfamilies (Supplementary Figure S8). Interestingly, the best model in terms of average K2 score is the SSCPE model that maximizes the likelihood for each superfamily (ML(SSCPE)), followed by the simple Stab-CPE model MF and then the combined models RMSDMF, DEMF, DEWT. The Str-CPE models RMSD and DE score the worst among the SSCPE models, but still they perform better than the empirical models. The empirical model showing the maximum likelihood for each superfamily (ML(emp)) is one of the worst in terms of tree distance, which seems in contradiction with the good score of the ML(SSCPE) models.

As expected, the scaling factor of the one-parameter fit, which represents the ratio between structure divergence and sequence divergence, is less than one for all models and all superfamilies (Fig.4D). The two-parameter fit evidences slow branches in which the scaling factor is even smaller (Fig.4E) and fast branches in which the scaling factor approaches one (Fig.4F). Inspection of the phylogenetic trees supports the view that the latter branches represent events in which the protein function, represented by its GO term or by the EC numbers, changed. We will present a more quantitative analysis of this important property in another publication.

Regularized maximal likelihood and minimum evolution

We then tested whether the presented results are robust when, instead of the NJ trees adopted by our program PC_ali, we use minimum evolution (ME) trees inferred with the program FastME (Lefort et al. 2015) applied to the structure-based divergence matrices generated by PC_ali. By visual inspection of the trees (Supplementary Fig.S3-S7) we realized that in some cases the GO terms or EC numbers support the NJ tree rather than the ME tree, despite it is frequently claimed that ME trees are more accurate (Lefort et al. 2015). Therefore, we sought an objective score for selecting between the two types of tree.

In the ML inference methods, the branch lengths t_b are obtained through a fit. It is well known that fits are threatened by overfitting, and their performances (for instance in terms of cross-validation) improve considerably when they are regularized by penalizing large values of the fitted parameters, for instance through the Tykhonov regularization that we adopt here for regularizing the fit of the selection parameter Λ. In the context of phylogenetics, the parameters are the branch lengths and, by the mathematical definition of the likelihood function, there is a clear trade-off that longer branches tend to be associated with higher likelihood. In this case, the L1 regularization that constrains the absolute values of the parameter is equivalent to minimizing the following quantity:

$$\text{RegMLaME} = -\text{loglik}(T, \{t_b\}) + \mu \sum_b t_b. \tag{31}$$

In other words, the regularized log likelihood also integrates the minimal evolution principle that is adopted for instance by the FastME program. Furthermore, the regularized log likelihood score can be seen also as the maximum posterior probability (MPP) of a tree T with an exponential prior distribution of branch lengths, while the ML tree can be seen as the MPP tree with uniform prior distribution, which seems to be less justified than the exponential distribution. Therefore, the regularized ML and ME tree unifies the three criteria most frequently adopted for tree inference: ML, ME and Bayesian. Finally, in (Dehouck and Bastolla 2017) one of us noted the analogy

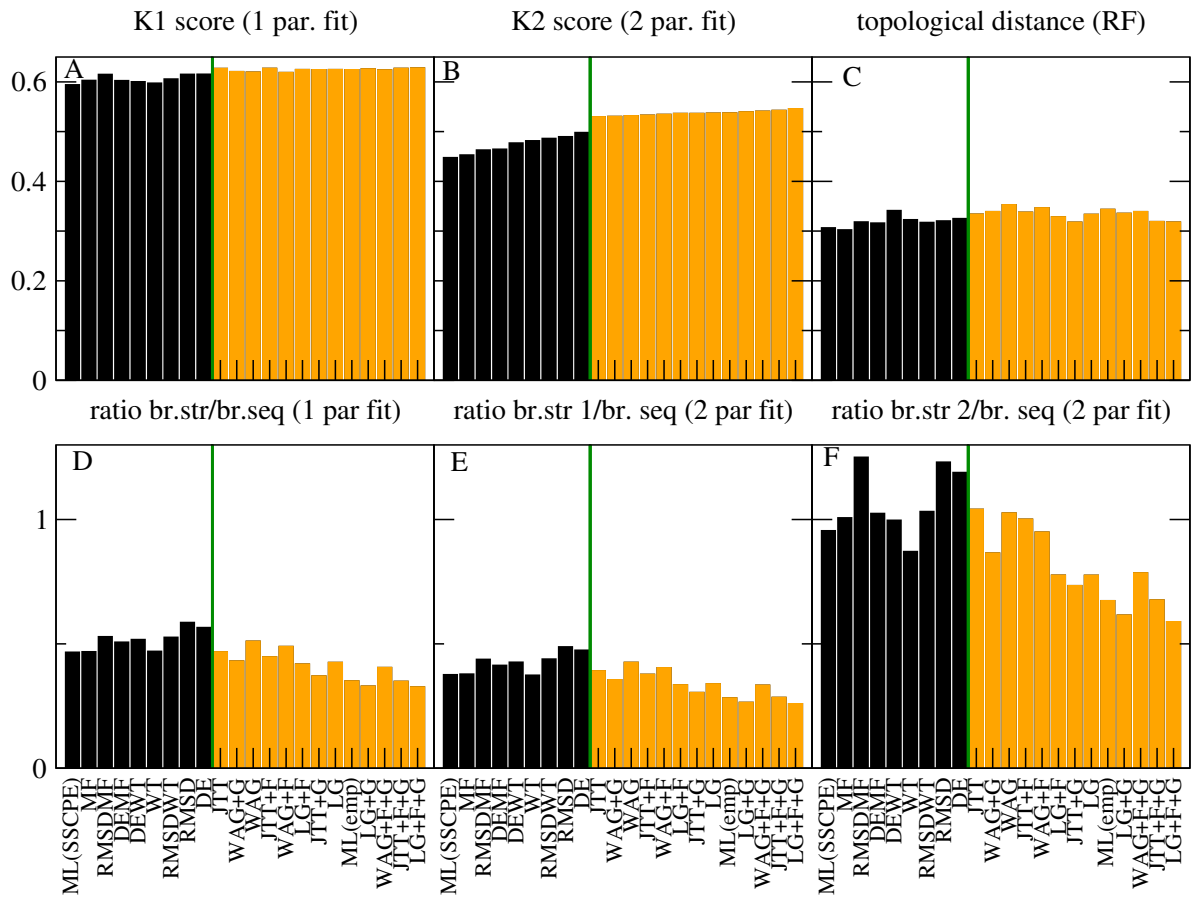

Figure 4: **Assessment of SSCPE and empirical models with NJ structure-based phylogenetic trees.** Comparisons between the NJ structure-based trees of five superfamilies and the corresponding ML trees obtained with SSCPE models (black bars) or empirical models (orange bars). A: Normalized K score, B: Normalized K2 score, C: Normalized topological distance (Robinson-Foulds) between structure based and sequence-based branches. Fitted ratios between the structure-based and sequence-based branches for 1 parameter fit (D), two parameter fit and short structure branches (E) or long structure branches (F). The models are ranked by smaller K2 scores. Each value is averaged over five superfamilies. The SSCPE models are run with the options HB, FLUX and rate all active.

between regularization and thermodynamics. In the present context, $-log.lik$ can be interpreted as an energy, which increases when increasing the regularization parameter μ, interpreted as a temperature, so that the quantity represented in Eq.(31) can be interpreted as a free energy.

If we fit branch lengths with ML for different trees using the same model, we find that they are positively related with the sum of branch lengths, loglik $\approx \mu \sum_b t_b + c$. This fitted parameter μ represents a natural choice for the regularization parameter μ. This choice amounts at not rewarding the likelihood in Eq.(31) unless it exceeds what is expected given the branch lengths of the tree.

For ME fits, the situation is the opposite. They are biased to favour short trees, but the trade-

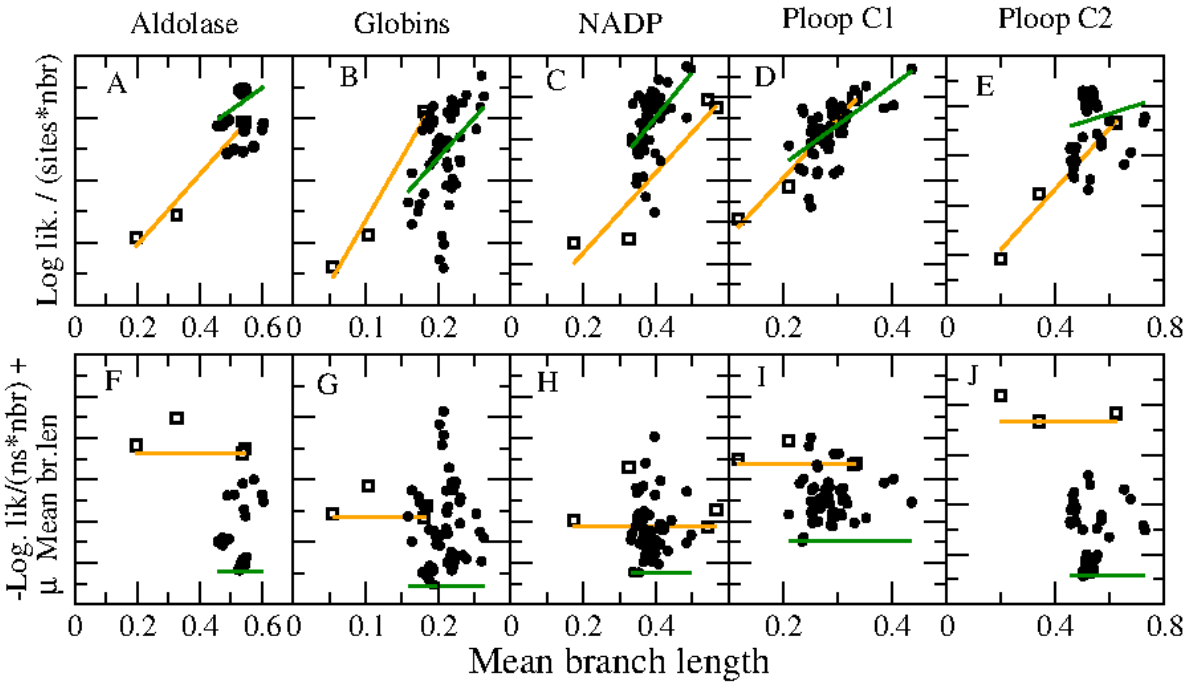

Figure 5: **Selection of the reference phylogenetic tree.** A-E: For the neighbor joining (NJ) and minimal evolution (ME) structure-based trees, we compute the log-likehood adopting the WAG+F+G model and recompute the branch lengths through maximum likelihood, obtaining four trees (squares) that evidence the trade-off between branch length and log-likelihood. A similar relationship holds for log-likelihoods of the ML trees computed with each SSCPE and empirical model (back symbols) versus their branch lengths. The fits of log-likelihood versus branch lengths (orange lines) estimates the regularization parameter μ with which we compute the REGMLAME score (figures F-J). For each superfamily, we choose as reference the structure-based tree with minimum REGMLAME score.

off is that they present low likelihood. Also in this case, Eq.(31) provides a less biased score than the length of the tree. We adopted this score for comparing the NJ and ME trees. We proceeded as follows: first, we computed the likelihood of the trees with the model WAG+F+G, which is the ML model for most superfamilies. In this way, we verified that the ME trees have much lower likelihood than the NJ trees, as expected. Second, we recomputed the branch lengths maximizing the likelihood while maintaining the topology of the tree. In this way, we obtained four trees with their branch lengths and likelihood scores that are positively correlated, as expected (see Fig.5A to E, squares). Third, we computed the parameter μ for each superfamily from the regularized fit of the log likelihood versus the branch lengths and, with this value, we computed Eq.(31) for all trees, choosing the one with minimum value. For three superfamilies (Globins, NADP, Ploop_C1) this is the ME tree, but for the remaining two it is the NJ tree. We did not modify its branch lengths, since they contain important structural information.

We then compared the ML trees inferred with the empirical models and with the SSCPE models to the reference trees obtained with the best between the NJ and ME approach. Despite the reference tree changed in three out of five cases, the qualitative results are robust and they

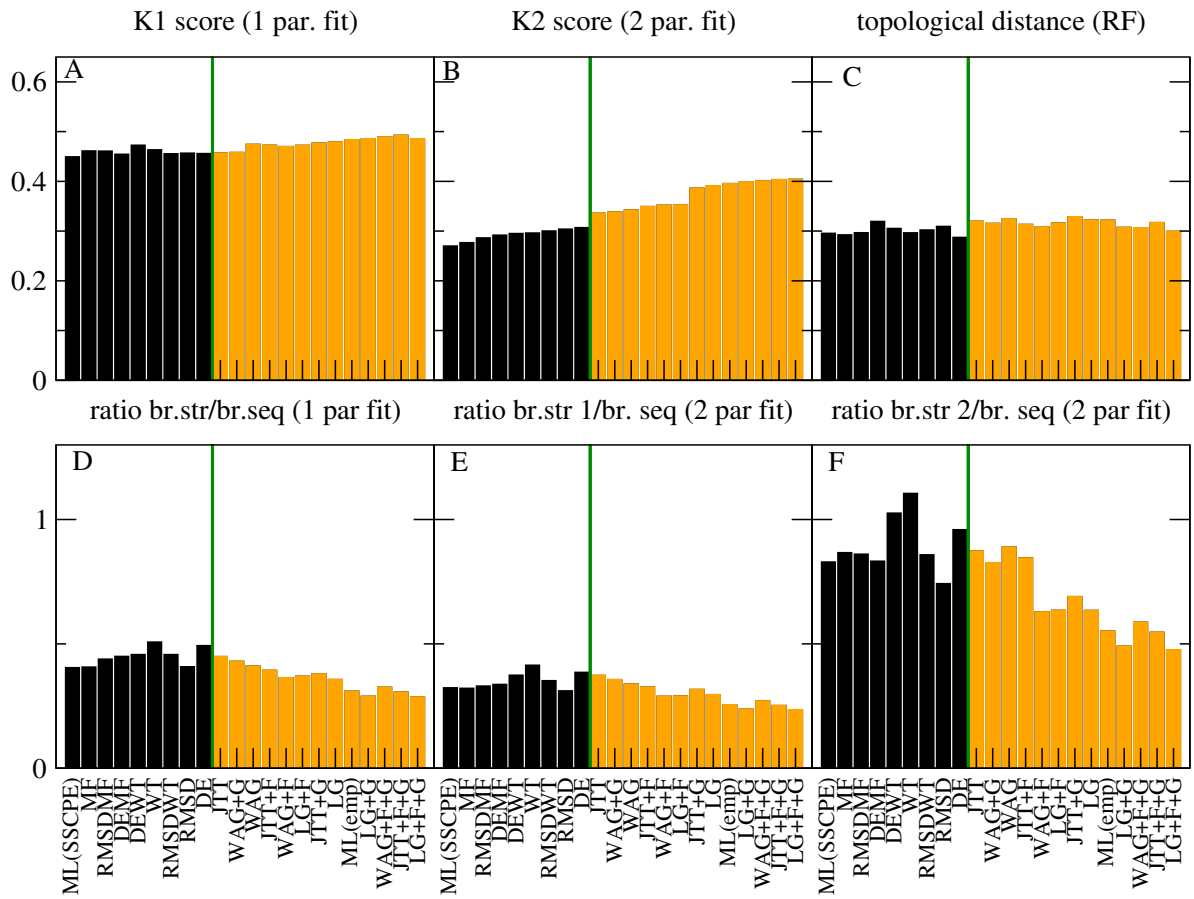

Figure 6: **Assessment of SSCPE and empirical models with selected structure-based phylogenetic trees.** Comparisons between five reference structural trees (best between NJ and ME tree) and the corresponding ML trees obtained with SSCPE models (black bars) or empirical models (orange bars). A: Normalized K score, B: Normalized K2 score, C: Normalized topological distance (Robinson-Foulds) between structure based and sequence-based branches. Fitted ratios between the structure-based and sequence-based branches for 1 parameter fit (D), two parameter fit and short structure branches (E) or long structure branches (F). The models are ranked by smaller K2 scores. Each value is averaged over five superfamilies. The SSCPE models are run with the options HB, FLUX and rate all active.

generally improve, since all the dissimilarities K, K2 and RF decreased (compare Fig.6 with Fig.4). The K2 scores are better supported with the new reference trees for the SSCPE models (see Supplementary Figure S8B), although they are slightly worse supported for the empirical models.

We also analyzed the relationships between the dissimilarity scores and the log likelihood and the REGMLAME scores of all ML trees built with different models (see Table 2 and Supplementary Figures S9 and S10). For NJ trees, as expected the log likelihood scores are negatively correlated and the REGMLAME scores are positively correlated with the dissimilarity scores for all superfamilies except Ploop_C1 for which the correlations are not significant. The absolute value of the correlations is larger for the log likelihood score than for the regularized score, probably because the latter was

not computed in a consistent manner (the branch lengths were not obtained by minimizing Eq.(31). However, for the ME trees the correlations have the opposite sign, positive between dissimilarity and log likelihood and negative between dissimilarity and Eq.(31). This supports the view that the branch lengths in ME trees are strongly biased.

Comparison of the phylogenetic likelihood of the models

We compared the empirical models and the SSCPE models under the point of view of the likelihood scores. Fig.7B shows that the average likelihood of the SSCPE models (black bars) is always larger than the one of the empirical models (orange bars) without the +G option, but it is smaller than some of the empirical models with the +G option. However, this higher likelihood of the +G empirical models is attained at the price of longer branch lengths (Fig.7C) and when this is taken into account the SSCPE models (except DE and RMSD) always have lower REGMLAME score than the empirical models (Fig.7A. In all figures, the models are ranked according to the REGMLAME score).

We caution that, in this and the following section, the computation of the REGMLAME score is only approximate because the branch lengths should be computed minimizing Eq.(31) and the parameter μ should be fitted separately for each model. However, these results are consistent with the finding that the SSCPE model perform better than empirical models, as assessed through the comparisons with the structure-based trees.

Finally, we used the site-specific substitution processes to infer phylogenetic trees of 10 randomly chosen protein alignments of our database (Supplementary Table S1) with the ML method implemented in the program RAxML-NG (Kozlov et al. 2019). The results are presented in Fig.8. In this case all of the +G empirical models present higher likelihood than the SSCPE models (Fig.8B), which is particularly so for large trees with more than 100 sequences (Supplementary Fig.S12B and D). The increased likelihood of the option +G is partly compensated by the biased increase of the branch length (Fig.8C), so that the REGMLAME score is more uniform (Fig.8A). The WAG+G model has still lower regularized score than any SSCPE model for large alignments (100 residues or more), except for the smallest one (1iu4, 10 sequences) that presents an advantage for SSCPE models so large that the minimum value of the average REGMLAME score is attained by the models RMSDWT and RMSDMF when we consider 1iu4. We removed this special case from Fig.8 in order to avoid to bias the results. Again, we note that the REGMLAME score that we compute here is only approximate.

Discussion

Although the selective forces that act in protein evolution are complex and difficult to model, many years of modelling and observations have revealed 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. Here we show that introducing biophysical realism in the substitution models of protein evolution can improve phylogenetic inferences.

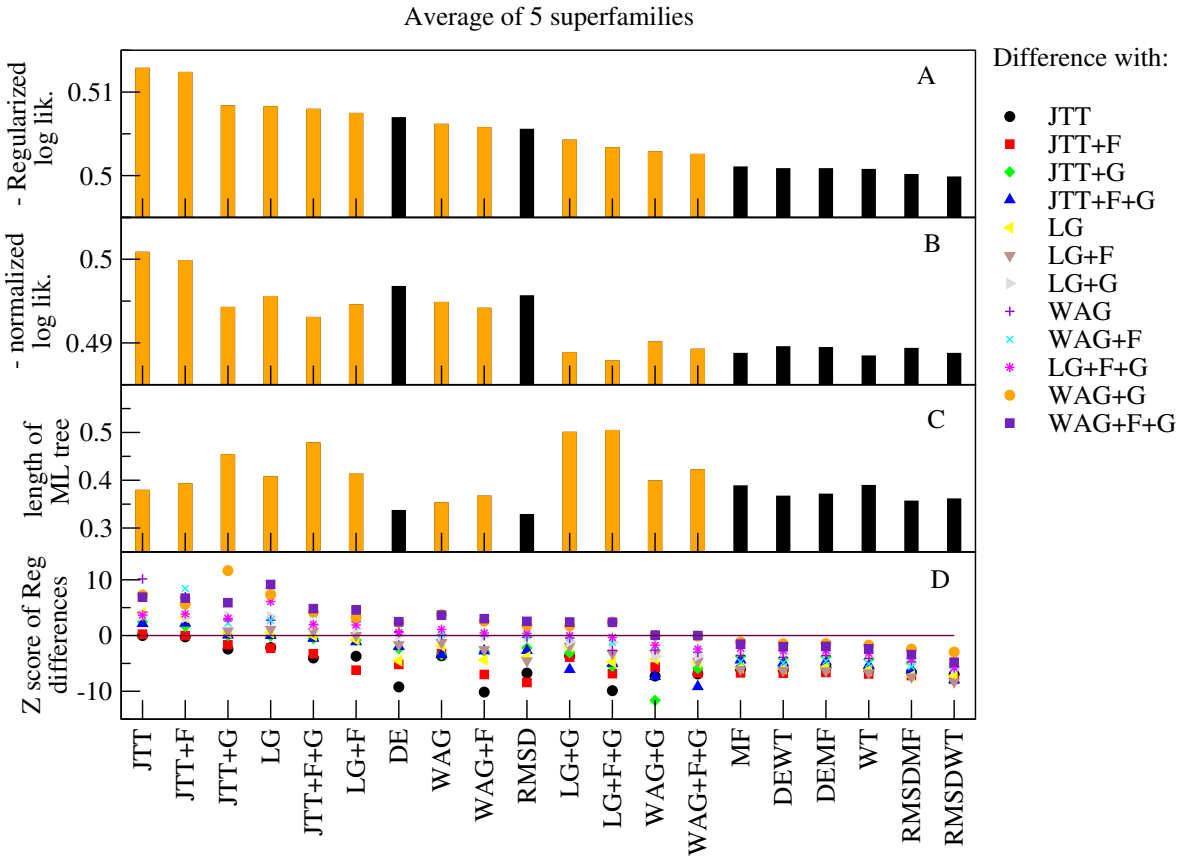

Figure 7: **Regularized likelihood scores of SSCPE and empirical models for the superfamily MSAs.** Substitution models, listed in the horizontal axis, are ranked according to their regularized log-likelihood (REGMLAME) score averaged over the five superfamilies. Empirical substitution models are orange, SSCPE substitution models are black. A: REGMLAME score. B: Minus normalized log-likelihood. C: Average branch length of ML trees. One can see that low REGMLAME scores trade-off high likelihood and short branch length. D: Z score of the difference of the REGMLAME score between the named model and each of the empirical substitution models.

Stability-constrained versus structure-constrained protein evolution

The folding stability of the native state is a biophysical property that received much attention in the literature (see for review Goldstein 2011, Serohijos and Shakhnovich 2014, Bastolla et al. 2017). Previous work produced site-specific stability-constrained substitution models that can be used for phylogenetic inference (Arenas et al. 2015). However, these Stab-CPE models assume for simplicity that the native structure is strictly conserved in evolution, and they cannot represent the selective process for maintaining it.

The native structure strongly influences native protein dynamics (Tirion 1996, Atilgan et al. 2001) and it is thought to be under strong negative selection when the protein function is conserved (Illergard et al. 2009, Pascual-Garcia et al. 2010) and under strong positive positive selection when the protein function changes (Pascual-Garcia et al. 2019). Therefore, it is crucial to consider it for

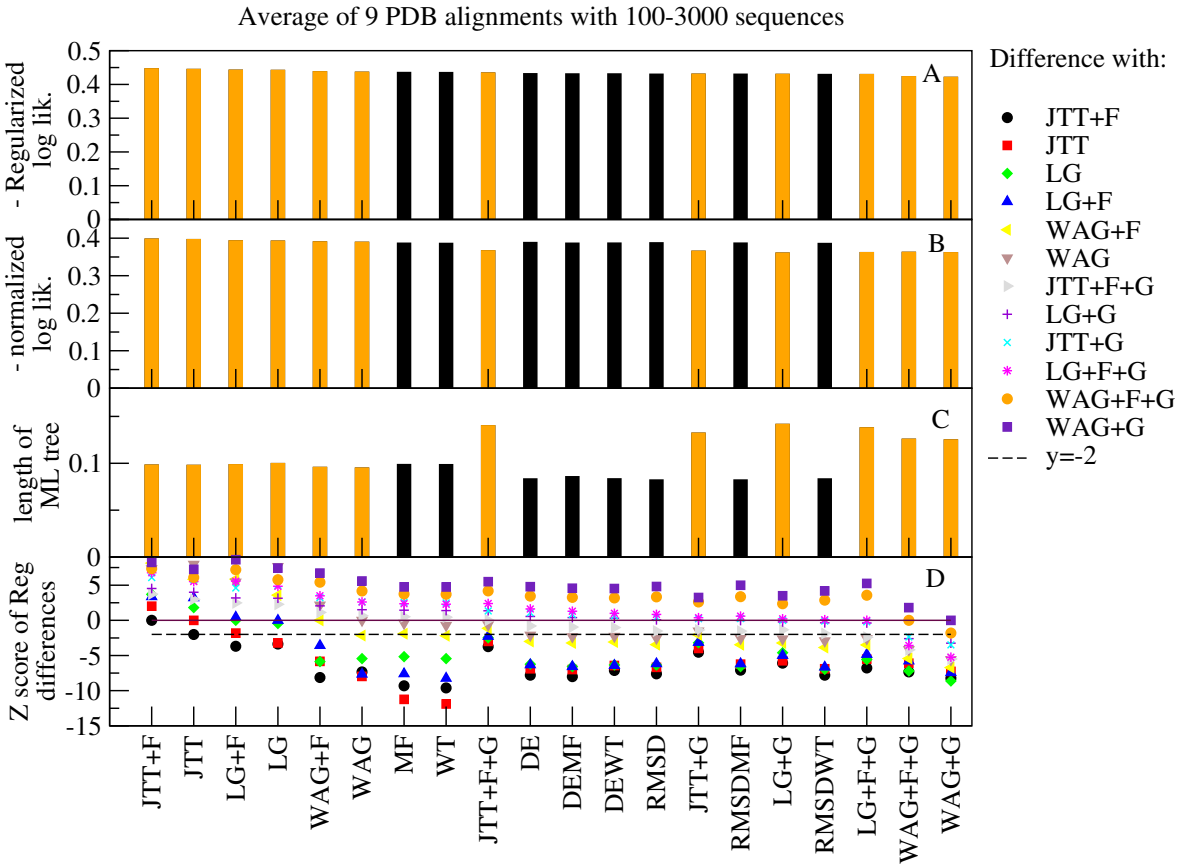

Figure 8: **Regularized likelihood scores of SSCPE and empirical models for 9 MSAs with known native structure.** We excluded 1iu4 because in this small MSA (10 very similar proteins) the empirical models score much worse than the SSCPE models, biasing the averages. Substitution models, listed in the horizontal axis, are ranked according to their regularized log-likelihood (REGMLAME) score averaged over nine multiple alignments of proteins with known structure. Empirical substitution models are shown in orange, SSCPE substitution models are shown in black. A: REGMLAME score. B: Minus normalized log-likelihood. C: Average branch length of ML tree. One can see that low REGMLAME scores trade-off high likelihood and short branch length. D: Z score of the difference of the regularized log-likelihood between the named model and each of the empirical substitution models.

modelling natural selection. In this work, we introduced two structurally constrained fitness functions based on the model by Echave (2008) that predicts the structural changes due to mutations as the linear response of the elastic network model that represents the protein dynamics to the perturbation caused by the mutation. Advancing with respect to Echave's proposal, we developed an explicit model of the specific amino-acid mutations in terms of changes of size, stability and optimal distance of the native contacts formed by the mutated residue. Irrespective of the parameters, and similar to Echave's original model, our 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).

We introduce here two Str-CPE models that represent the logarithm of the fitness as minus the predicted RMSD due to the mutation (RMSD model) or as the predicted change in the harmonic energy of the elastic network (DE model, similar to the original proposal by Echave). We compared these models with two variants of our previous Stab-CPE models that model the logarithm of the fitness as minus the folding free energy ΔG evaluated with contact interactions, and consider as non-native states both the unfolded state and compact but incorrectly folded states (Minning et al. 2013).

These structure and stability constrained (SSCPE) evolutionary models depend on the global amino acid frequencies (as in empirical models with +F), which we determine imposing that the mean amino acid frequencies across all sites equal the global amino acid frequencies, and on the selection parameter Λ. We fit Λ by minimizing the symmetric Kullback-Leibler (KL) divergence between the site-specific amino acid frequencies of the MSA and the equilibrium frequencies of the models. To reduce overfitting, which favour large values of Λ and attributes very small probabilities to amino acids that are not observed, we regularized the fit by the Tykhonov regularization scheme (Hoerl and Kennard 1970), also known as ridge regression, choosing the regularization parameter R that maximizes the derivative of the score with respect to R (Dehouck and Bastolla 2017), inspired by the analogy between regularization problems and statistical physics.

The resulting site-specific SSCPE models provide lower KL divergence, lower sequence entropy and more stable proteins than the global model with optimized frequencies and $\Lambda = 0$ (Fig.2A-D). We found that Str-CPE models are characterized by lower entropy, diverge more from the site-unspecific global model and tend to have larger selection parameter Λ than the Stab-CPE models (Fig.2D-F). These results support the view that selection on protein structure is stronger than selection on protein stability. Nevertheless, the sequences described by the Str-CPE models are not stable on the average (Fig.2C) and they describe worse the amino acid distributions in the MSA (Fig.2A-B). Next, we combined the Str-CPE and Stab-CPE models. These combined models present better agreement with empirical MSAs, lower entropy, higher folding stability and diverge more from the site-unspecific model than both Stab-CPE and Str-CPE models (Fig.2A-E), showing consistent improvement under all criteria.

We examined site-specific evolutionary properties as a function of the number of contacts of the site in the native structure. 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 and agree better with observed data (Fig.3A, B). Moreover, SSCPE models 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 be also in part attributed to insufficient sampling of the protein sequences, which underestimates the site-specific variability.

In conclusion, the combined SSCPE models represent site-specific selection on protein sequences more realistically than Stab-CPE and Str-CPE models, and much more than the global model that does not consider site-specific selection.

Phylogenetic inference

We then tested whether the SSCPE models improve phylogenetic inference. We obtained the structure-based reference trees of five data sets of proteins distantly related at the superfamily level with the program PC_ali (Bastolla et al. 2023), and we inferred the maximum likelihood (ML) trees of the same proteins with several combinations of SSCPE and empirical substitution models with the program RAxML-NG (Kozlov et al. 2019).

We compared the structure-based and the sequence-based trees by fitting the lengths of the corresponding branches either with only one scale parameter (K score, Soria-Carrasco et al. 2007) or with two scale parameters (K2 score, our own extension of the Soria-Carrasco et al. program). For all models and all superfamilies, we found that the branches of the structure-based trees tend to be shorter than for sequence-based trees, in agreement with the notion that protein structures tend to evolve more slowly than sequences, but some of the branches are almost equally long in structure and in sequence, in agreement with the notion that structure evolution accelerates when the protein function changes (Pascual-Garcia et al. 2019). This feature may provide a mean of detecting changes of protein function, which we plan to explore in the future.

We found that the site-specific SSCPE models tend to infer trees that are more similar to the reference trees in terms of the K score, in terms of the topological distance by Robinson and Foulds (1981), and above all in terms of the K2 score, which is more relevant in this context since it takes into account the acceleration of protein structures upon function change.

Regularized maximum likelihood and minimal evolution (REGMLAME) score

We first considered structure-based trees inferred with the neighbor joining method (NJ, Saitou and Nei 1987), and later tested whether these results are robust under structure-based trees inferred with the minimal evolution method (ME, Lefort et al. 2015). However, by direct inspection of the ME trees, it was apparent that sometimes their quality was not as good as those of the NJ trees. We devised a function for scoring them in an objective way.

For doing so, we used the program RAxML-NG for computing the likelihood of the structure-based trees under the WAG+G empirical model. We observed that the log likelihood is strongly posiively correlated with the sum of branch lengths, as expected from the mathematical definition of the likelihood. In fact, the likelihood is computed from the elements of the matrix e^{tQ}, where t is the branch length, and for different amino acids the logarithm of the likelihood is positively related with t. Thus, there is a trade-off between log likelihood and branch length. A natural way to determine the optimal balance is by regularizing the fit of branch lengths, penalizing large values of the parameters (i.e. long branches) or, equivalently, penalizing short branches with low likelihood. This regularization leads to minimize the REGMLAME score, Eq.(31). It is well known that regularizing the fitted parameters improves their ability to generalize and reduces the occurrence of unphysical values (see for instance (Hoerl and Kennard 1970, Dehouck and Bastolla 2017, Bastolla and Dehouck 2019)), and we expect that this holds true also for the regressions used for phylogenetic inference. The REGMLAME score can also be interpreted as a maximum posterior probability under a prior distribution of trees that is exponential of the branch lengths. Therefore, the REGMLAME score unifies the three most used tree inference principles: ML, ME and MPP.

Applying an approximate version of the REGMLAME score, we found that the NJ trees have better score for two superfamilies, and the ME trees have better score for the remaining three.

When we compared the SSCPE and empirical trees to the selected reference trees, we found that the comparison scores consistently improved and once again the phylogenetic trees inferred under the SSCPE models were more similar to the reference phylogenetic trees than the phylogenetic trees inferred under the empirical substitution models.

Finally, we computed the likelihood scores and the branch lengths of the ML trees obtained with 12 empirical substitution models and 8 SSCPE models. Empirical models with the option +G sometimes present higher likelihood than the SSCPE models. However, this improvement is achieved at the expense of biased longer branches. When this is taken into account with the REGMLAME score, the SSCPE models score better.

Software implementation

Since the computations of the site-specific substitution matrices are fast 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 or can be predicted by AlphaFold2 (Jumper et al. 2021). The program tnm is the computational bottleneck, because of the computation of the normal modes and the $2 \times 19L$ deformations produced by all possible mutants, which scales as L^3. The tnm results are stored and reused if needed. The computer time is 9.7 sec for a small protein with $L = 129$ amino acids (PDB 132l), 51.9 sec for an intermediate one of 217 residues (1zio) 153.9 sec for one with 314 residues (1hqc) and 3598 sec for a large protein with 907 amino acids (1vlb). The program Prot_evol computes the 8 SSCPE models in 7.1 sec for 132l, 9.3 sec for 1zio, 11.7 for 1hqc and 109.9 sec for 1vlb. To reduce the computation time, users may run only Stab-CPE models for large proteins and Str-CPE and combined models for small ones.

Data Availability

The framework SSCPE.pl that performs phylogenetic tree inference under the SSCPE models with the program RAxML-NG is freely available at https://github.com/ugobas/SSCPE. Given an alignment of protein sequences and a list of representative protein structures (i.e., PDB files), SSCPE.pl uses the output of Prot_evol or runs the program if the output does not exist, and performs phylogenetic tree inference with RAxML-NG under the specified SSCPE model.

SSCPE.pl can be installed by downloading the package SSCPE.zip at https://github.com/ugobas/SSCPE/blob/main/SSCPE.zip that provides all the needed codes and compiles and installs the programs tnm, Prot_evol and SSCPE.pl. Instructions for the installation and the use are present at https://github.com/ugobas/SSCPE. The program RAxML-NG is included for completeness, but we recommend to install the latest version from https://github.com/amkozlov/raxml-ng.

The package SSCPE.zip also contains the program K2_mat, which compares two phylogenetic trees by performing two-parameter fit of the corresponding branch lengths under a modification of the program K_mat.pl (Soria-Carrasco et al. 2007) and the master script script_run_SSCPE.pl that, for given MSA, runs a combination of several SSCPE models and empirical substitution models, submits the computations to a computing cluster, and compares the resulting trees with a reference tree computing several other scores, including an approximated REGMLAME score.

Supplementary figures and data are stored in the Dryad data repository at http://datadryad.org. They include: 1. Supplementary Figures S1-S12 described in the main text. 2. Mutation parameters used for tnm predictions. 3. MSAs of the 203 studied proteins. 4. RMSD and 5. DE predicted by the tnm program for all possible mutations of the studied proteins. 6. Data generated for testing the models: Site-specific substitution rates, sequence entropy, mean hydrophobicity and number of contacts. 7. Summary data for all tested models and all proteins. 8. For each superfamily, structure-based phylogenetic trees and ML trees inferred with the different empirical models and SSCPE models. Table with the comparisons between the reference structure-based trees and the ML trees of all models. 9. For each of the 10 protein families studied for phylogenetic inference, ML trees inferred with the different empirical models and SSCPE models. Table with the likelihood, sum of branch length and regularized likelihood of the ML trees of all models.

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 Julián Echave for interesting discussions, and Fernando Otero de Navascués for collaborating to a previous version of this work. We thank Centro de Supercomputación de Galicia (CESGA) for part of 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, Piette O, Abia D 2023. PC_ali: a tool for improved multiple alignments and evolutionary inference based on a hybrid protein sequence and structure similarity score. Bioinformatics 39:btad630.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] Dehouck Y, Mikhailov A. 2013. Effective harmonic potentials: Insights into the internal cooperativity and sequence-specificity of protein dynamics. PLoS Comp Biol 9:e1003209.

[17] 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.

[18] Dehouck Y, Bastolla U. 2021. Why are large conformational changes well described by harmonic normal modes? Biophys J. 2021 120:5343-5354.

[19] 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.

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

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

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

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

[24] Fisher R. 1930. The distribution of gene ratios for rare mutations. Proc R Soc Edin 50: 205-220.

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

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

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

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

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

[30] 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.

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

[32] 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.

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

[34] 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.

[35] Jimnez-Santos MJ, Arenas M, Bastolla U. 2018b. Influence of mutation bias and hydrophobicity on the substitution rates and sequence entropies of protein evolution. PeerJ 6:e5549

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

[37] Jumper, J., Evans, R., Pritzel, A. et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589.

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

[39] 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.

[40] Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. Mol Biol Evol 25:1307-1320.

[41] Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol Biol Evol 32:2798-800.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math. Biosci. 53: 131-147.

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

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

[52] 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.

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

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

[55] Soria-Carrasco V, Talavera G, Igea J, Castresana J. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23:2954-6.

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

[57] Uversky VN. 2011. Intrinsically disordered proteins from A to Z. Int J Biochem Cell Biol. 43:1090-103.

[58] 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.

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

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

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

[62] Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15: 1600-1611.

[63] Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25: 568-579.

[64] 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.

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

[66] Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57:702-10.