No signature of natural selection in patterns of protein structural divergence

March 20, 2017

Abstract

Proteins diverge during biological evolution. At sequence level, different sites evolve at different rates, mainly due natural selection. In contrast, it has been suggested that observed patterns of structural divergence are not a signature of natural selection but, rather, of the response of protein structure to random mutations. Here, we have systematically studied whether there is any signal of natural selection in patterns of protein structural evolution. We model evolution as follows: (1) proteins are Elastic Networks of amino acids, (2) a mutation at a site perturbs the springs that connect it to its neighbors, (3) selection is either not considered (by fixing all mutations) or included by fixing mutants according to a stability-based fitness function. We analyzed the structural divergence among sites and among normal modes. We compared predicted and observed patterns for several protein families. We found very good agreement between predicted and empirical structural divergence patterns whether natural selection is considered or not. For all cases studied, including selection does not improve model fit. Therefore, observed patterns can be explained in terms of mutational robustness and sensitivity of the structure. In a word, we found no evidence of natural selection in patterns of structural divergence.

1 Introduction

Proteins diverge during biological evolution, which is evident in the variation of the aminoacid sequences and the resulting structural, dynamical and functional changes. It is known that the structure diverges much more slowly than the sequence, that the structural divergence occurs mainly along the low energy vibrational modes and that there is a structurally conserved core. These facts are difficult to interpret because most of the studies made so far are purely empirical. To go forward in this sense, the mechanistic model "Linearly Forced – Elastic Network Model" (LF - ENM) was developed , which predicts the change in the equilibrium position of protein sites as the result of random mutations, not subjected to natural selection [1]. Applying this model, it was shown that the experimental patterns of structural change can be reproduced without resorting to natural selection [3,4]. This result call into question interpretations based on the assumption that everything that is conserved or that varies is related to the biological function.

While natural selection apparently little affects structural divergence patterns, at the level of aminoacid sequences different sites evolve at different speeds mainly due to natural selection. Purely mutational evolutional models, such as the LF-ENM, cannot account for this fact. To explain such patterns of sequence variation, natural

Family	no selection	strong selection
Serin Proteases	0.67	0.70
Azurin - Plastocyanins	0.61	0.65
Phospholipases	0.66	0.67
Fatty acid binding proteins	0.74	0.79
Globins	0.69	0.67
RNA recognition motif	0.75	0.75
Snake toxins	0.82	0.78
SH3 homology domain	0.78	0.74
Mean	0.72	0.72

Table 1: Correlation coefficient (CC) between experimental ¡$zRSD_i$¿ profiles and theoretical ¡$zRSD_i$¿ profiles. Only no selection and strong selection theoretical profiles are shown.

selection must be modeled. We have recently proposed a mechanistic stress model of evolution [5], which is based on the idea that a mutant is viable to the extent that it spends time in the active conformation. This model has been successfully used to account for the average evolutionary variation from site to site [6].

Considering this scenario, we set out to study the role of natural selection at the structural divergence level with a different approach: (1) using the LF - ENM to simulate mutations and (2) either not selecting them or fixing them according to the stress model fitness function. We will show that the agreement between experimental and simulated structural divergence profiles is high either considering or not natural selection and that, including natural selection, does not improve the model fit.

2 Results and discussion

We aim to study the role of natural selection on the structural divergence of proteins by comparing experimental proteins with simulated mutants obtained either considering or not natural selection. To do this, we first selected several families of experimental proteins. Then, for each family, we selected the most structurally representative protein as the reference "ancestor" and we simulated multiple mutants of this protein using the LF - ENM and accepting each single mutation according to its stress model probability of fixation. To account for different selection regimens, we gave the average probability of fixation of mutations different values: ≈ 1 (no selection, all mutants are accepted), ≈ 0.9 (weak selection), ≈ 0.5 (medium selection) and ≈ 0.1 (strong selection). Finally, we calculated measures of structural variability on Cartesian coordinates and projected on the normal modes of the reference protein.

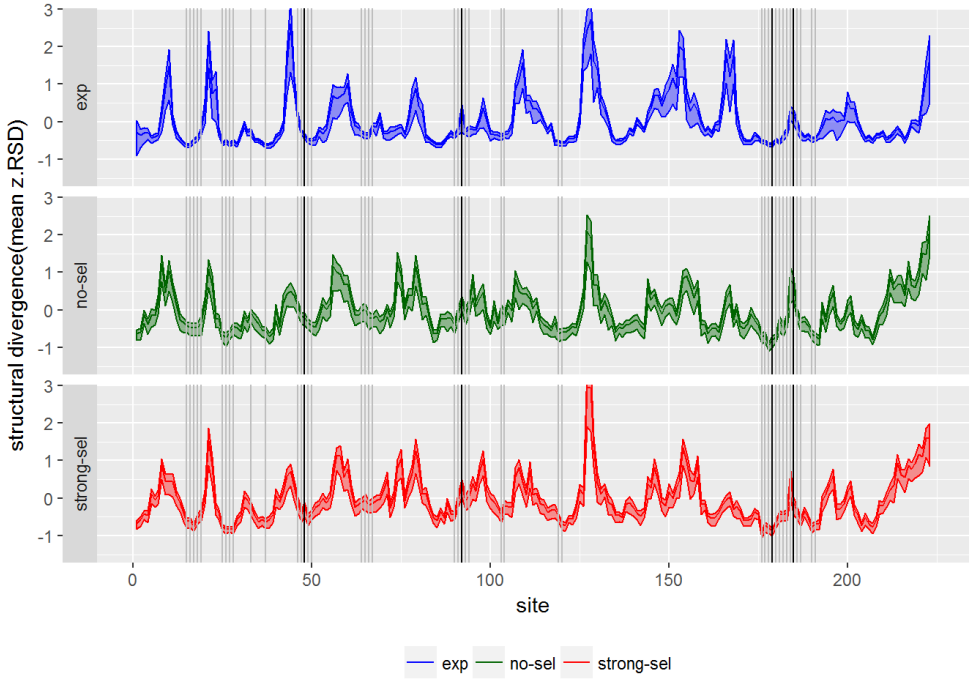
Cartesian coordinates

For each family we calculated experimental and theoretical $< zRSD_i >$ profiles as explained in Methods and we calculated the correlation coefficient (CC) between these profiles. The results are shown in Table 1.

Table 1 shows that, for all families studied, the CC between the experimental profile and each theoretical profiles is high (≈ 0.72) and that, accounting for natural selection, does not seem to improve the agreement. Figure 1 shows $< zRSD_i >$ profiles obtained for 1MCT chain A, the reference protein of the Serine Proteases

family, and figure 2 shows the structure of the same protein colored according to the $< zRSD_i >$ profiles. The active site of the protein and its neighborhood are highlighted in both figures.

Figure 1: Experimental and theoretical $< zRSD_i >$ profiles obtained for the Serine Proteases family. Only no selection and strong selection theoretical profiles are shown. The active site of the reference protein corresponds to the vertical black lines and its neighborhood corresponds to the vertical gray lines.



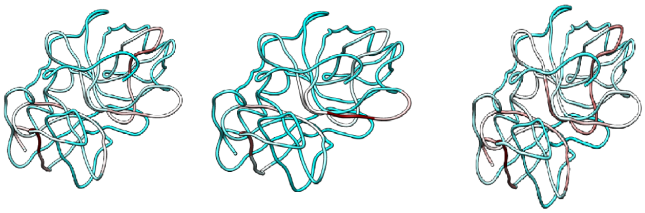
It can be noticed in figure 1 and figure 2 that the qualitative similarity between experimental and theoretical profiles, with or without considering natural selection, is high even in the active site of the protein and in its neighborhood. A high qualitative similarity is observed for the other families too (data not shown).

2.0.1 Active site closeness

To study this profiles regarding the active site closeness of the sites, we calculated the CC between $< zRSD_i^{experimental} > - < zRSD_i^{theoretical} >$ and the distance of each site to its closest active site. If there was any signal of natural selection we would expect that the difference between these profiles would be negative for sites near the active site and close to 0, positive or negative, for distant sites. Thus, we would expect a positive correlation coefficient. Table 3 shows the results obtained for all enzymatic families.

It can be noticed in Table 3 that the CC obtained for some families have the predicted sign but are very low. Moreover, as we suspected that this slight CC might be due to the fact that more divergent sites have less information and that their variability tend to be underestimated, we repeated the analysis only considering the

Figure 2: Reference protein of the Serine Proteases family colored according to experimental and theoretical $< zRSD_i >$ profiles. Only no selection and strong selection theoretical profiles are shown. The active site of the reference protein corresponds to the gold sites and its neighborhood corresponds to the orange sites.



Family	no selection	strong selection
Serin Proteases	0.15	0.16
Azurin - Plastocyanins	-0.04	-0.06
Phospholipases	-0.05	-0.05
Fatty acid binding proteins	0.13	0.1
Globins	0.01	0.01

Table 2: Correlation coefficient (CC) between $< zRSD_i^{experimental} > - < zRSD_i^{theoretical} >$ and the distance of each site to its closest active site. Only no selection and strong selection theoretical profiles are shown.

conserved core of proteins (sites of the ancestor with no gaps on the whole structural alignment). We obtained that, for all cases, taking out divergent and noisy sites diminished the CC to a very low value (data not shown). These results are more proof that there is no evidence of natural selection on structural divergence of proteins even in the active site neighborhood.

2.0.2 $zRSD_i$ vs. CN_i

To study whether structural variation profiles are related to the degree of packaging of the site, as it has been suggested in previous studies, we calculated the CC between $zRSD_i$ profiles and the contact number (CN_i) of the sites. Table 3 shows the results.

It can be noticed in Table 3 that the CC obtained for all families is high. This result implies that the structure varies more along less packed sites and that there is a core of packed sites structurally conserved.

Family	experimental	no selection	strong selection
Serin Proteases	-0.62	-0.78	-0.78
Azurin - Plastocyanins	-0.55	- 0.68	-0.72
Phospholipases	-0.61	-0.68	-0.75
Fatty acid binding proteins	-0.67	-0.65	-0.68
Globins	-0.5	-0.64	-0.56
RNA recognition motif	-0.64	-0.52	-0.54
Snake toxins	-0.61	-0.69	-0.68
SH3 homology domain	-0.58	-0.66	-0.59

Table 3: Correlation coefficient (CC) between $< zRSD_i >$ profiles and $< CN_i >$. Only no selection and strong selection theoretical profiles are shown.

Family	no selection	strong selection
Serin Proteases	0.83	0.82
Azurin - Plastocyanins	0.69	0.71
Phospholipases	0.59	0.56
Fatty acid binding proteins	0.85	0.88
Globins	0.77	0.73
RNA recognition motif	0.56	0.46
Snake toxins	0.66	0.69
SH3 homology domain	0.94	0.95
Mean	0.74	0.73

Table 4: Correlation coefficient (CC) between experimental ¡P_n¿ profiles and simulated mutants ¡P_n¿ profiles. Only no selection and strong selection theoretical profiles are shown

Normal modes coordinates

For each family, we calculated theoretical and experimental $< P_n >$ profiles as explained in Methods and we calculated the CC between these profiles. The results are shown in Table 2.

Table 2 shows that the CC between the experimental profiles and theoretical profiles is high (≈ 0.73) and that natural selection does not improve the model fit, another proof of the lack of natural selection on the structural evolution of proteins.

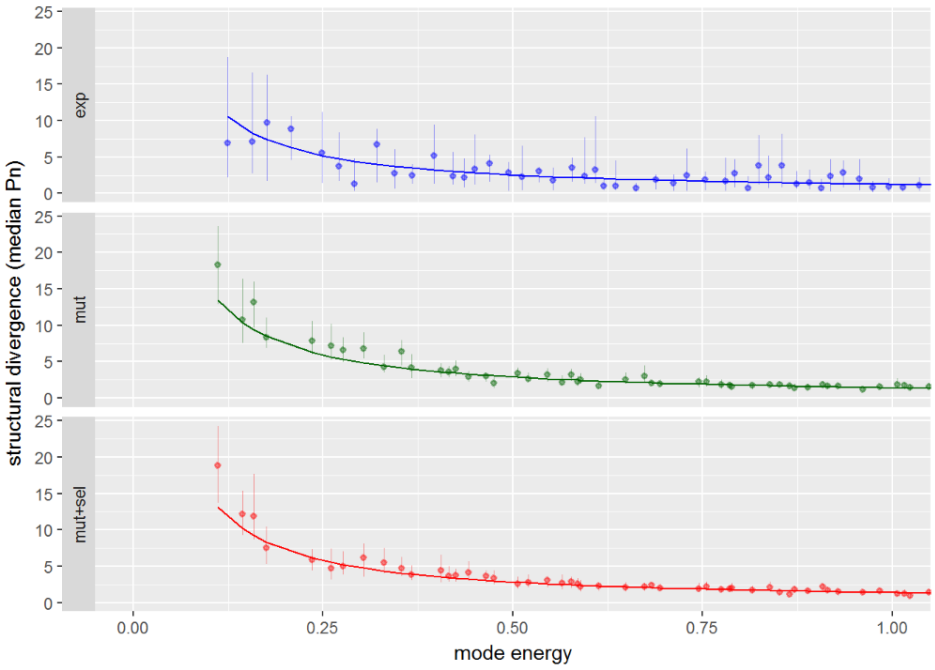
Figure 3 shows $< P_n >$ profiles obtained for the reference protein of the Serine Proteases family.

It can be observed in Figure 3 that, in agreement with previous studies, the structure varies mores along the low energy normal modes. Moreover, the qualitative similitude between these profiles is high.

2.1 Stress model verification

As we used a fitness function based on the stress model to generate theoretical mutants, we verified that sequence evolutionary rates obtained for these families by this model correlate better with experimental sequence divergence profiles than evolutionary rates of a purely mutational model. To do this, for each family we calculated the number of times we had mutated each site under each different selection regimen. Then, we correlated obtained profiles with site ´ s evolutionary rates obtained from

Figure 3: Experimental and theoretical $< P_n >$ profiles obtained for the Serine Proteases family. Only no selection and strong selection theoretical profiles are shown.



ConsurfDB. Results are shown in Table 4.

It can be observed in Table 4 that the stress model indeed predicted the site´s evolutionary rate at a great extent. Thus, we proved that our selection function is suitable and that not finding differences between different regimens profiles means that there is no evidence of natural selection on protein structure evolution.

3 Conclusions

We studied the role of natural selection on the structural divergence of proteins by comparing experimental proteins with simulated mutants obtained either considering or not natural selection. To do this, we selected diverse families of proteins and we simulated multiple mutants of a reference protein of each family using the LF - ENM and selecting each single mutation according to the fitness function given by the stress model. Varying the main parameter of this function, we generated theoretical mutants with different selection regimens: no selection, weak selection, medium selection and strong selection.

We found that a purely mutational model (no selection) already predicts patterns of protein structural evolution in excellent agreement with observations. Moreover, adding selection does not improve the model-data fit. Therefore, observed patterns are just the response of protein structure to random mutations and can be explained in terms of mutational robustness and sensitivity. Specifically, the structure is more sensitive in the direction of low-energy normal modes, which explains their higher contribution to evolutionary divergence. Also, the protein core is more structurally

conserved than the surface because it is more robust with respect to random mutations.

To summarize, we found no evidence of natural selection in patterns of structural divergence. This finding challenges our fundamental understanding of protein evolution since it goes against the common view that conservation patterns are generally signatures of selection.

4 Methods

4.1 ENM

We consider the backbone fluctuations of a protein around its equilibrium conformation to be described by a coarse - grained "Elastic Network Model" (ENM), which represents a protein as a network of sites connected by springs. In general, the ENM potential is of the form:

$$V_{wt} = \frac{1}{2} \sum_{i} \sum_{i<j} k_{ij} (d_{ij} - d_{ij}^0)^2 \tag{1}$$

where k_{ij} is the force constant of the spring connecting sites i and j, d_{ij} is the distance between sites i and j and d_{ij}^0 is the equilibrium distance between these sites. These distances are calculates as the modules of $\mathbf{d}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{d}_{ij}^0 = \mathbf{r}_i^0 - \mathbf{r}_j^0$ respectively, being $\mathbf{r}$ the position of a given site and $\mathbf{r}_0$ the equilibrium position of the site.

LF - ENM

To simulate mutants of a protein we used the "Linearly Forced - Elastic Network Model" (LF - ENM). This model simulates the effect of a single mutation by perturbing the equilibrium lengths of the ENM springs: $d_{ij}^0 \rightarrow d_{ij}^0 + \Delta_{ij}$, where Δ_{ij} are picked independently from the same uniform distribution, which satisfies $< \Delta_{ij} >= 0$ and $Var(\Delta_{ij}) = \sigma^2$. Following this, the mutant's potential is of the form:

$$V_{mut} = \frac{1}{2} \sum_{i} \sum_{i<j} k_{ij} [d_{ij} - (d_{ij}^0 + \Delta_{ij})]^2 \tag{2}$$

Then, the LF - ENM is obtained from expanding Eq. 2 up to second order. The potential is expressed in terms of "forces" directed along the contacts of the mutated site with lengths of the form $f_{ij} = k_{ij}\Delta_{ij}$. Finally, the equilibrium structure of the mutant $\mathbf{r}_{mut}^0$ is the value of $\mathbf{r}$ that minimizes V_{mut}. Using Eqs. 1 and 2 and after some algebra we find the structural variation due to the mutation of a protein of N sites:

$$d\mathbf{r}^0 \equiv \mathbf{r}_{mut}^0 - \mathbf{r}_{wt}^0 = \mathbf{K}_{wt}^{-1}\mathbf{f} \tag{3}$$

being $\mathbf{r}$ a $3N$ vector of coordinates, $\mathbf{f}$ a $3N$ vector of forces and $\mathbf{K}$ a $3N$ x $3N$ stiffness matrix, which represents the network's topology and the spring force constants.

Stress Model of protein evolution

The stress model of protein evolution predicts the acceptance probability of single mutations. The model is based on the idea that a mutant is viable to the extent that it spends time in the active conformation, which depends on mutational changes of the stability of the active conformation. The fixation probability of a mutant is modeled as:

$$P_{fix} \propto C^F_{mut}\rho_{mut}(\mathbf{r}_{active})/C^F_{wt}\rho_{wt}(\mathbf{r}_{active}) \tag{4}$$

where C^F is the concentration of folded protein and $\rho(\mathbf{r}_{active})$ its probability of adopting the active conformation. Assuming that C_{mut}/C_{wt} is equal to the ratio of partition functions, from basic statistical physics it follows that:

$$P_{fix} \propto e^{-\beta\Delta V^*} \tag{5}$$

where $\Delta V^* = V_{mut}(\mathbf{r}_{active}) - V_{wt}(\mathbf{r}_{active})$ is the energy difference between the mutant and the wild-type in the active conformation and β represents the selective pressure. Lower values of β imply weaker selective pressure and higher values of β imply stronger selective pressure.

4.2 Two-nodes per site evolution model

As we previously found that sequence evolutionary rates are better reproduced using a model that considers both alpha carbons (C_α) and geometric centers (ρ) of aminoacids [] than using a one-node per site model, in this work we represented proteins by means of the two-nodes per site model. The ENM potential of Eq. 1 can be rewritten in terms of C_α and ρ distances:

$$\begin{aligned} V_{wt} = &\frac{1}{2}\sum_i\sum_{i<j} k_{C_{\alpha_i}C_{\alpha_j}}(d_{C_{\alpha_i}C_{\alpha_j}} - d^0_{C_{\alpha_i}C_{\alpha_j}})^2 + \frac{1}{2}\sum_i\sum_{i<j} k_{C_{\alpha_i}\rho_j}(d_{C_{\alpha_i}\rho_j} - d^0_{C_{\alpha_i}\rho_j})^2 \\ &+ \frac{1}{2}\sum_i\sum_{i<j} k_{\rho_i C_{\alpha_j}}(d_{\rho_i C_{\alpha_j}} - d^0_{\rho_i C_{\alpha_j}})^2 + \frac{1}{2}\sum_i\sum_{i<j} k_{\rho_i\rho_j}(d_{\rho_i\rho_j} - d^0_{\rho_i\rho_j})^2 \end{aligned} \tag{6}$$

where $d_{n_in_j}$ is the distance between nodes n_i and n_j (n is either C_α or ρ), $k_{n_in_j}$is the force constant of the spring connecting these nodes, and $d^0_{n_in_j}$ is the equilibrium spring length.

A mutation at site i will replace ρ_i, affecting only the parameters of the energy function related to this node. Following [], we model a mutation at site i by adding random perturbations to the lengths of the springs connected to ρ_i: $d^0_{\rho_i\rho_j} \rightarrow d^0_{\rho_i\rho_j} + \Delta_{\rho_i\rho_j}$ and $d^0_{\rho_i C_{\alpha_j}} \rightarrow d^0_{\rho_i C_{\alpha_j}} + \Delta_{\rho_i C_{\alpha_j}}$. We can again express the potential of the mutant in terms of "forces" directed along the C_α and ρ contacts of the mutated site with lengths of the form $f_{\rho_i\rho_j} = k_{\rho_i\rho_j}\Delta_{\rho_i\rho_j}$ and $f_{\rho_i C_{\alpha_j}} = k_{\rho_i C_{\alpha_j}}\Delta_{\rho_i C_{\alpha_j}}$. We use the same criteria explained for the one-node per site model to get the structure of the mutant using Eq. 3, being now $\mathbf{K}$ a $6N$ x $6N$matrix and $\mathbf{f}$ a column vector of length $6N$.

To calculate the fixation probability of the mutant, we use Eqs. 6 and 5 to get:

$$P_{fix} = e^{-\beta\frac{1}{2}\sum_{i\neq j}(k_{\rho_i C_{\alpha_j}}\Delta^2_{\rho_i C_{\alpha_j}} + k_{\rho_i\rho_j}\Delta^2_{\rho_i\rho_j})} \tag{7}$$

For the special case of the "Anisotropic Network Model" (ANM), which gives a spring force constant of 1 to nodes at a distance $< R_0$ (the contacts) and of 0 to nodes at a distance $> R_0$, we get the average probability of mutation of site i:

$$< P^i_{fix} >= e^{-\beta \frac{1}{2} < \Delta^2_{ij} > (CN_{i_{C_\alpha \rho}} + CN_{i_{\rho\rho}})} \quad (8)$$

being $CN_{i_{c_\alpha \rho}}$ and $CN_{i_{\rho\rho}}$the number of C_α and ρ contacts of ρ_i respectively.

4.3 Experimental dataset

We selected 8 families of proteins from the Database of Multiple Structural Alignments of Homologous HOMSTRAD. In the dataset, there are representatives of the major structural classes: all alpha, all beta, alpha and beta and small proteins. We looked for families that possess multiple structural alignments with more than 12 proteins and with an alignment length greater than 50 sites. For each family, we obtained the multiple structural alignment and the superimposed coordinates of the proteins. Then, we selected a reference "ancestor" protein. To do this, we calculated the average structure of the family and selected the protein with the lower "Mean Square Deviation" (MSD) between its structure and the average structure. We approximated each family tree topology by a "star tree" that begins with the ancestor protein. Each lineage corresponds to a pair alignment of each of the other proteins with the ancestor protein and the corresponding"branch length" is the number of mutated sites.

Theoretical dataset

For each family and for each lineage we simulated 100 mutants of the reference protein following a path of substitutions composed of various evolutionary steps, each of them comprising a single substitution. The steps were simulated by first picking one random site l of the reference protein, obtaining a set of forces f_{lj} for each of the j (C_α and ρ)contacts of ρ_l and the reaction force over ρ_l, and calculating the mutant´s structure from Eq. 3. The $\mathbf{K}$ matrix used is $6N$ x $6N$ and is based on a two-nodes per site model that considers both C_α and ρ interactions. After we obtained the structure of the mutant, we calculated the probability of fixation of the mutation from Eq. 8 and, with this value, we calculated the logical variable Accept = $P_{fix} \geq$ runif(0,1). If Accept was TRUE, we accepted the mutation and the evolutionary step was finished. Else, we rejected the trial mutation and tried again until we had mutated the number of mutated sites that corresponds to the lineage.

We simulated mutants with different selection regimens; No selection $< P_{fix} > \approx 1$, weak selection $< P_{fix} > \approx 0.9$, medium selection $< P_{fix} > \approx 0.5$ and strong selection $< P_{fix} > \approx 0.1$. To obtain these $< P_{fix} >$ we gave β different values according to the following equation:

$$\beta^{regimen} = -ln(< P^{regimen}_{fix} >)/(-\beta \frac{1}{2} < \Delta^2_{ij} > (< CN_{C_\alpha \rho} > + < CN_{\rho\rho} >) \quad (9)$$

being the average over all ρ_i of the corresponding reference protein. It is remarkable that each family has a different value of β.

Structural variability measures

For each family and for each subset, experimental or theoretical with different selection regimens, we obtained the coordinates of each C_α and the aligned and nonaligned sites. For the experimental subset of each family, aligned and nonaligned sites of the reference protein relative to each of the other proteins were extracted from the

multiple structure alignment provided by HOMSTRAD. For theoretical subsets, we considered that there were not nonaligned sites (no gaps or insertions). With this data, we calculated the following measures of structural variability.

4.3.1 Cartesian coordinates

For each family and for each subset we calculated the average structural variation on Cartesian coordinates. To do this, for each comparison between a protein and the reference protein of the corresponding family, we calculated the Root Square Deviation (RSD_i) of each aligned C_α as follows:

$$RSD_i = ((\Delta x_i)^2 + (\Delta y_i)^2 + (\Delta z_i)^2)^{1/2} \tag{10}$$

being Δx_i, Δy_i and Δz_i the difference in the x_i, y_i and z_i Cartesian coordinates of the C_{α_i} of the proteins. Then, we calculate $z-$normalized profiles. Finally, we averaged these profiles to obtain $< z\mathbf{RSD}_i >$ profiles.

4.3.2 Normal modes coordinates

For each family and for each subset we calculated the median of the structural variation on normal modes coordinates. To do this, we first obtained the $\mathbf{K}_{eff}$ matrix of the reference protein of each family like in []. Both not aligned C_α and all ρ where taken away to calculate $\mathbf{K}_{eff}$ so that the matrix accounts for the effective movements of the aligned C_α. Then, we calculated the normal modes as follows:

$$\mathbf{K}_{eff}\mathbf{q}_n = \Lambda_n \mathbf{q}_n \tag{11}$$

where Λ_n and $\mathbf{q}_n$ are the n eigenvalue and eigenvector respectively. There are $3N_{aligned} - 6$ non-zero eigenvalues, which correspond to the vibrational modes. Then we calculated the contribution of each normal mode to the total structural variation (P_n) by projecting structural differences of the aligned C_α on each normal mode n of the reference protein. We normalized profiles so that they added up to 1. Finally, we calculated the median of these profiles to obtain $< \mathbf{P}_n >$ profiles.

Model parameters

To completely specify the model, we must specify parameters for $\mathbf{K}$ and $\mathbf{f}$. To calculate $\mathbf{K}$, as we mentioned before, we used the ANM. We chose $R_0 = 7.5$ after a step of optimization using different values (data not shown). To calculate $\mathbf{f}$, the magnitude of each $\mathbf{f}_{lj}$, which depend on the value of Δ_{lj}, were randomly picked from a uniform distribution in the interval $[-f_{max}, f_{max}]$. The forces for different contacts were picked independently. Since f_{max} does not affect the results, we set $f_{max} = 2$ like in [].

5 References