

No signature of natural selection in patterns of protein structural divergence

March 5, 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 variation of 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 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 lower 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 evolutionary models, such as the LF-ENM, cannot account for this fact. To explain such patterns of sequence variation, natural 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 diverse families of proteins from the Database of Multiple Structural Alignments of Homologous HOMSTRAD according to the criteria explained in Methods. We extracted from HOMSTRAD the multiple structural alignments and the superimposed coordinates of all of the proteins. Then, for each family, we selected the most structurally representative protein as the reference “ancestor” protein and we considered the other proteins different lineages of a “star tree” that begins with the ancestor. For each lineage, we extracted aligned aminoacids and calculated the “branch length”, which we considered as the number of mutated sites in the aligned region. After that, we simulated multiple mutants of the ancestor using the LF - ENM and selecting each single mutation according to the probability of fixation given by the stress model. To account for different selection regimens, we gave the average probability of acceptance 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. We compared these measures for experimental and theoretical profiles in order to find out whether there is any signal of natural selection.

Cartesian coordinates

We calculated $\langle zRSD_i \rangle$ profiles as explained in Methods. Then, for each family, we calculated the correlation coefficient (CC) between experimental profiles and theoretical profiles based on different selection regimens. The results are shown in Table 1.

Table 1 shows that, for all families studied, the CC between the experimental profile and the different theoretical profiles are very good (≈ 0.72) and

that, accounting for natural selection does not seem to improve the agreement. Figure 1 shows the $\langle zRSD_i \rangle$ profiles obtained for the reference protein of the Serine Proteases family and figure 2 shows the same protein colored according to the different $\langle zRSD_i \rangle$ profiles. The active site of the protein and its neighborhood are shown in both figures. It can be noticed in figure 1 and figure 2 that the qualitative similarity between experimental and theoretical profiles, with any selection regimen, is really high, even in the active site of the protein and on in the active site's neighborhood.

2.0.1 Active site distance

To focus on the active site and on its neighborhood, we calculated the CC between $\langle zRSD_i^{experimental} \rangle - \langle zRSD_i^{theoretical} \rangle$ and the distance of each site to its closest active site. If there were 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 all cases 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 on the conserved core of proteins (sites with no gaps on the whole structural alignment). We noticed 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.

Normal modes coordinates

We calculated $\langle P_n \rangle$ profiles as explained in Methods. Then, for each family, we calculated the CC between the experimental profile and theoretical profiles based on different selection regimens. 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 fit, another proof of the lack of natural selection on proteins structural evolution.

2.1 Stress model sequence variability prediction

As we used here a fitness function based on the stress model, we must verify that sequence evolutionary rates obtained for these families by this model correlate better with experimental sequence 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 on the subset of simulated mutants under each different selection regimens. Then, we correlated obtained profiles with site's evolutionary rate obtained from 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 signature of natural selection on protein structure evolution.

3 Methods

3.1 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 nodes connected by springs. In general, the ENM potential is of the form:

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

where k_{ij} is the force constant of the spring connecting nodes 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 calculated 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 $\langle \Delta_{ij} \rangle = 0$ and $Var(\Delta_{ij}) = \sigma^2$. Following this, the mutant’s potential is of the form:

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

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 nodes:

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

being \mathbf{r} a $3 \times N$ vector of coordinates and \mathbf{K} the stiffness matrix, which represents the network’s topology and 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. In turn, this time will

depend on mutational changes of the stability of the active conformation. The fixation probability of a mutant is modeled as:

$$P_{fix} \propto C_{mut}^F \rho_{mut}(\mathbf{r}_{active}) / C_{wt}^F \rho_{wt}(\mathbf{r}_{active}) \quad (4)$$

where wt stands for wild-type, mut for mutant, 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:

$$P_{fix} \propto e^{-\beta \Delta V^*} \quad (5)$$

where $\Delta V^* = V_{mut}(\mathbf{r}_{active}) - V_{wt}(\mathbf{r}_{active})$ is the energy difference between mutant and 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. From (4) and (5) we get

$$P_{fix} = e^{-\beta \frac{1}{2} \sum_{j \neq i} k_{ij} \Delta_{ij}^2} \quad (6)$$

To carry on with this formula, we must specify parameters for \mathbf{K}_{wt} . In this work, we used the ‘‘Anisotropic Network Model’’ (ANM). Following this model, we gave a spring force constant of 1 to nodes at a distance $\leq 10 \text{ \AA}$ and of 0 to nodes at a distance $> 10 \text{ \AA}$. Using this information, we can write $\langle P_{fix} \rangle$ as follows:

$$\langle P_{fix} \rangle = e^{-\beta \frac{1}{2} \langle \Delta_{ij}^2 \rangle CN_i} \quad (7)$$

being CN_i the number of contacts of site i .

3.2 Two-nodes per site model

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

$$V_{wt} = \frac{1}{2} \sum_i \sum_{i < j} k_{\alpha_i \alpha_j} (d_{\alpha_i \alpha_j} - d_{\alpha_i \alpha_j}^0)^2 + \frac{1}{2} \sum_i \sum_{i < j} k_{\alpha_i \rho_j} (d_{\alpha_i \rho_j} - d_{\alpha_i \rho_j}^0)^2 + \frac{1}{2} \sum_i \sum_{i < j} k_{\rho_i \alpha_j} (d_{\rho_i \alpha_j} - d_{\rho_i \alpha_j}^0)^2 + \frac{1}{2} \sum_i \sum_{i < j} k_{\rho_i \rho_j} (d_{\rho_i \rho_j} - d_{\rho_i \rho_j}^0)^2 \quad (8)$$

where $d_{n_i n_j}$ is the distance between nodes n_i and n_j (n is α or ρ) $k_{n_i n_j}$ is the force constant of the spring connecting these nodes, and $d_{n_i n_j}^0$ 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. We emphasize: while the mutation may well induce global structural changes involving the backbone and other side chains, the only parameters that will change are those of the mutated side chain. Following [ref], we model a mutation at i by adding random perturbations to the lengths of the springs connected to ρ_i : $d_{\rho_i \rho_j}^0 \rightarrow d_{\rho_i \rho_j}^0 + \Delta_{\rho_i \rho_j}$ and $d_{\rho_i \alpha_j}^0 \rightarrow d_{\rho_i \alpha_j}^0 + \Delta_{\rho_i \alpha_j}$. We can again express the potential in terms of ‘‘forces’’ directed along the α and ρ contacts of the mutated site with lengths of the form $f_{\rho_i \rho_j} = k_{\rho_i \rho_j} \Delta_{\rho_i \rho_j}^2$ and $f_{\rho_i \alpha_j} = k_{\rho_i \alpha_j} \Delta_{\rho_i \alpha_j}^2$. Then, we use the same criteria explained

for a one-node per site model and we get the structure of the mutant using Eq. 3. Then, using () and () we get:

$$\Delta V^* = \frac{1}{2} \sum_{i \neq j} (k_{\rho_i \alpha_j} \Delta_{\rho_i \alpha_j}^2 + k_{\rho_i \rho_j} \Delta_{\rho_i \rho_j}^2) \quad (9)$$

hence, for this model we get:

$$P_{fix} = e^{-\beta \frac{1}{2} \sum_{i \neq j} (k_{\rho_i \alpha_j} \Delta_{\rho_i \alpha_j}^2 + k_{\rho_i \rho_j} \Delta_{\rho_i \rho_j}^2)} \quad (10)$$

For the special case of ANM, from () and () we can rewrite this equation as the two nodes per site analogous of ()

$$\langle P_{fix} \rangle = e^{-\beta \frac{1}{2} \langle \Delta_{ij}^2 \rangle (CN_{i\alpha\rho} + CN_{i\rho\rho})} \quad (11)$$

being $CN_{i\alpha\rho}$ and $CN_{i\rho\rho}$ the number of α and ρ contacts of ρ_i respectively.

3.3 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’s 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 following a path of substitutions. This path was composed of various evolutionary steps, each of them comprising a single substitution. The steps were simulated by picking one random site l of the reference protein, obtaining a set of forces f_{lj} for each of the j α or ρ contacts of ρ_l and the reaction force over l , and calculating the mutant’s structure from Eq. 3. Then, we calculated the probability of fixation of the mutation from Eq. eq:PacceptANM and the logical variable Accept = $P_{fix} \geq \text{runif}(1,0,1)$. 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 $\langle P_{fix} \rangle \approx 1$, weak selection $\langle P_{fix} \rangle \approx 0.9$, medium selection $\langle P_{fix} \rangle \approx 0.5$ and strong selection $\langle P_{fix} \rangle \approx 0.1$. To obtain these $\langle P_{fix} \rangle$ we gave β different values according to the following equation:

$$\beta^{regimen} = -\ln(\langle P_{fix}^{regimen} \rangle) / (-\beta \frac{1}{2} \langle \Delta_{ij}^2 \rangle (\langle CN_{\alpha\rho} \rangle + \langle CN_{\rho\rho} \rangle)) \quad (12)$$

being the average over the ρ s of the proteins.

Structural variability measures

For each family, we obtained the coordinates of aligned and nonaligned sites of each protein relative to the reference protein. For experimental proteins, aligned and nonaligned indexes were taken from the multiple structure alignment obtained from HOMSTRAD. For theoretical proteins, we considered that there were not naligned sites (no gaps or insertions). Then, we obtained experimental proteins structures from de pdb file of superimposed coordinates also provided by HOMSTRAD and theoretical proteins structures from the simulation output. With this information we calculated measures of structural variability.

3.3.1 Cartesian coordinates

For each family, we calculated structural variation in the cartesian coordinates of the aligned sites of each protein relative to the reference protein. To do this, for each aligned site, we calculated its root square deviation by adding up the square difference of each cartesian coordinate x , y and z and taking the root of it. Then, we calculate z-normalized profiles. Finally, we promediated these profiles to obtain mean - $zRSDi$ (z Root Square Deviation of site i) profiles.

3.3.2 Normal modes coordinates

For each family, we calculated structural variation on normal modes coordinates by projecting the structural difference of the aligned sites of each protein relative to the reference protein on the normal modes of the last one. To do this, the \mathbf{K} matrix of the reference protein was replaced by the \mathbf{K}_{eff} matrix, which was calculated like in [ref] and which considers the effective movements of the C_α of the aligned sites. Both not aligned C_α and all ρ s were thaken away to calculate \mathbf{K}_{eff} . Then, we calculated normal modes as follows:

$$\mathbf{K}_{eff}\mathbf{q}_n = \Lambda_n\mathbf{q}_n \quad (13)$$

There are $3N - 6$ non-zero eigenvalues, which correspond to the vibrational modes, numbered $n = 1, 2, \dots, 3N-7$ and which were discarded. Then, we calculated P_n , the relative contribution of each normal mode to the total structural variation, by projecting structural differences on each normal mode n . After that, we normalized profiles so that they added up to 1. Finally, we promediated these profiles to obtain mean P_n profiles.

Model parameters

To completely specify the model, we must specify parameters for \mathbf{f} . To calculate \mathbf{f}_{lj} , given a mutation at a site l , each site j in contact with l was assigned a force directed along the $l - j$ contact and site l is assigned a reaction force. The magnitudes of each \mathbf{f}_{lj} , which depends on 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 $f_{max} = 2$. We can think of the range $[-f_{max}, f_{max}]$ as a continuous approximation of the perturbations introduced by the mutations, covering from mutations between

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

Table 1: Correlation coefficient (CC) between experimental mean $zRSD_i$ profiles and simulated mutants mean $zRSD_i$ profiles selected under different selection regimens: no selection, weak selection, medium selection and strong selection

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

Table 2: Correlation coefficient (CC) between experimental mean P_n profiles and simulated mutants mean P_n profiles selected under different selection regimens: no selection, weak selection, medium selection and strong selection

physicochemically similar amino acids ($f \approx 0$) up to mutations between very different amino acids ($f \leq f_{max}$).

3.4 Figures and Tables