No signature of natural selection in patterns of protein structural divergence

February 28, 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 of proteins and that there is a structurally conserved core. These facts are difficult to interpret because most of the studies that have been made so far are purely empirical. To go forward in this sense, it has been developed the mechanistic model "Linearly Forced – Elastic Network Model" (LF - ENM), which predicts the change in the equilibrium position of proteins 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 selection must be modeled. We have recently proposed a mechanistic stress model [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 rol of natural selection at the structural divergence level with a different approach: (1) using the LF - ENM to simulate single 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 data is high either considering or not natural selection and that including the selective preasure does not improve model fit.

2 Results and discussion

We aim to study the role of natural selection on the structural divergence of proteins. To do this, we first selected diverse families of proteins from the Database of Multiple Structural Alignments of Homologous HOMSTRAD and we obtained their multiple structural alignments. For each family, we selected one 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 the aligned and not aligned aminoacids and calculated the “branch length”, which we estimated as the number of mutated sites on the reference protein. Then, we simulated multiple mutants of the ancestor using the LF - ENM and selecting each single mutation according to its probability of fixation given by the stress model. To account for differnt 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 (stron selection). Finally, we calculated measures of structural variability on cartesian coordinates and projeted on the normal modes of the reference protein. We compared these measures for experimental and theoretical profiles in order to find out wheter there is any signal of natural selection.

Cartesian coordinates

We calculated mean $zRSD_i$ profiles as explained in Methods. Then, for each family, we calculated the correlation coefficient (CC) between experimental profiles and the 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 mean $zRSD_i$ profiles obtanied for the reference protein of the Serine Proteases family and figure 2 shows the same protein colored according to the different mean $zRSD_i$ 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 cualitative similarity between experimental and theoretical profiles, with any regimen of selection, is realy 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 its neighborhood, we calculated the CC between the difference of the experimental mean $zRSD_i$ profile and theoretical mean $zRSD_i$ profiles and (1) the distance of the sites to their closest active site and (2) minus the sum of the inverse of the distances to the sites to each 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 obtanied 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 noteced that, for all cases, thaking 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 average P_n 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 control

As we used here a fitness function based on the stress model, we must confirm that sequence evolutionary rates obtained by this model correlate better with experimental sequence profiles than evolutionary rates of a purely random mutational model. To do this, we obtained the number of times we had mutated each site on the subset of simulated mutants under different selection regimens. Then, we correlated the 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 signatute of natural selection on protein structure evolution.

3 Methods

3.1 ENM of the reference protein

We consider the backbone fluctuations of the reference “ancestor” 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 placed at its alpha carbons (C_α) connected by springs. In general, the ENM potential is of the form:

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

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^0_{ij} 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 $\mathbf{d}^0_{ij} = \mathbf{r}^0_i - \mathbf{r}^0_j$ 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 the reference 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: $d^0_{ij} \to d^0_{ij} + \Delta_{ij}$, where Δ_{ij} are picked independently for each of the contacts of the mutated site 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:

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

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 $\mathbf{r}^0_{mut}$ 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 reference protein of N sites:

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

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

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

where wt stands for wild-type, mut for mutant, C^F is the concentration of folded protein and $\rho(r_{active})$ 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 β represents the selection pressure and $\Delta V^* = V_{mut}(r_{active}) - V_{wt}(r_{active})$ is the energy difference between mutant and wild-type in the active conformation. Lower values of β imply weaker selective preasure and higher values of β imply stronger selective preasure. From () and () we get

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

To carry on with this formula, we must specify parameters for $\mathbf{K}_{wt}$. In this case, we used the "Anisotropic Network Model" (ANM). Following this model, we gave a spring force constant of 1 to sites at a distance ≤ 10 A and of 0 to sites at a distance ¿ 10 A. We can rewrite P_{accept} as follows:

$$P_{fix} = e^{-\beta \frac{1}{2} \Delta_{ij}^2 CN_i} \tag{7}$$

being CN_i the number of contacts of site i.

3.2 Two nodes per site model

As we previously found in [ref] that site-specific substitution rates are better reproduced using a model that considers not only C_α of aminoacids but also their geometric centers ρ, in this work we used this two nodes per site model to represent proteins. Acording to this model, the ENM potential is of the form:

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

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 $d^0_{n_i n_j}$ 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. 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: $d^0_{\rho_i \rho_j} \to d^0_{\rho_i \rho_j} + \Delta_{\rho_i \rho_j}$ and $d^0_{\rho_i \alpha_j} \to d^0_{\rho_i \alpha_j} + \Delta_{\rho_i \alpha_j}$, to find, using () and ():

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

then, for this model we get:

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

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

$$P_{fix} = e^{-\beta \frac{1}{2} \Delta^2_{ij\alpha} CN_{i\alpha} + \Delta^2_{ij\rho} CN_{i\rho}} \tag{11}$$

3.3 Experimental dataset

We selected 8 families of proteins from the Database of Multiple Structural Alignments of Homologous HOMSTRAD (http://mizuguchilab.org/homstrad/). In this 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 their multiple structural alignment and the superimposed coordinates of the proteins. Then, for each family we selected a reference "ancestor" protein. To get this protein, we calculated the average structure of the multiple alignment and selected the protein with the lower "Mean Square Deviation" (MSD) between its structure and the average structure. In principle, we should infer the phylogenetic tree and simulate structures following a tree with the same topology. However, we assume that the results are not too sensitive with respect to tree topology so we approximated the tree by a "star tree" that begins with the ancestor. Then, each lineage corresponds to a pair alignment of each of the other proteins with the ancestor and the"branch length" is the number of mutated sites of the ancestor compared to the other protein.

Theoretical dataset

For each family and for each lineage we simulated 50 mutants following a path of substitutions according to the branch length of the lineage. This path is 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 contacts of site l and the reaction force over site l, calculating the structure and the probability of accepting this trial mutation from Eqs. 3 and 7 and calculating the logical variable Accept = $P_{accept} \geq$ 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 os mutated sites that corresponds to the lineage. We simulated mutants with different regimens of selection; No selection $P_{accept} \approx 1$, weak selection $P_{accept} \approx 0.9$, medium selection $P_{accept} \approx 0.5$ and strong selection $P_{accept} \approx 0.1$. To get these P_{accept} we gave β different values:

$$\beta^{regimen} = -log(P_{accept}^{regimen})/(-\beta\frac{1}{2} < \Delta^2_{ij\alpha} >< CN_{i\alpha} > + < \Delta^2_{ij\rho} >< CN_{i\rho} >) \tag{12}$$

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 alignemnt obtanied 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 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 aligned sites of each protein relative to the reference protein on the normal modes of the last one. To do this, the K matrix of the reference protein was replaced by the K_{eff} matrix, which is calculated like in [ref] and which considers only the effective movements of the C_α of aligned sites. Both not aligned C_α and all C_ρ where thaken away. Then, we calculated normal modes as follows:

$$K_{eff} q_n = \Lambda_n q_n \tag{13}$$

There are 3N-6 non-zero eigenvalues, which correspond to the vibrational modes, numbered n = 0, 2,...,3N–7 and which were discarted. 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. Finally, 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 i 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 physicochemically similar amino acids ($f \approx 0$) up to mutations between very different amino acids ($f \leq f_{max}$).

3.4 Figures and Tables

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: Corelation 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: Corelation 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