Structural Variability

Maria Laura Marcos

May 9, 2016

1 Abstract

Proteins diverge during biological evolution, which is evident in the variation of the aminoacid sequences and the resulting structural, dynamical and functional changes. At the level of aminoacid sequences, there is a clear evidence of natural selection: different sites evolve at different speeds. These patterns are well reproduced by the recently proposed mechanistic model ("Stress Model"), in which a mutation is accepted at a rate proportional to the probability that the mutant adopts the "active structure". At the level of structure, empirical studies show 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 are structurally conserved cores in families of proteins. Applying a purely mutational model as the "Linearly Forced - Elastic Network Model" (LF - ENM) it has been shown that these structural divergence patterns can be well reproduced without resorting to natural selection. However, it is expected that the natural selection restricts, although very slightly, structural divergence. To study this, we deeply analyzed 8 structurally representative families of proteins. We obtained their multiple structural alignment and the structure of each protein of the family. Then, for each family, we used the LF - ENM to generate structures of multiple mutants of the reference "ancestor" protein. We did as much mutations as it corresponds to the sequence identity of this protein with each of the other proteins of the family. In the selection of which sites to mutate is where we included or not natural selection. In one case, we mutated sites randomly and, on the other case, we accounted for natural selection mutating sites more likely to mutate according to the "Stress Model". Finally, we calculated, for simulated and experimental sets of protein, profiles of structural variability in cartesian coordinates and projected on normal modes and compared them using the Pearson correlation coefficient and the "Mean Square Error" (MSE). We obtained that either at the level of cartesian coordinates or at the level of normal modes, there is no clear evidence of natural selection in protein evolution. These results give more evidence of the absence of natural selection in structural evolution.

2 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 and that the structural divergence occurs mainly along the lower energy vibrational modes of proteins and that there are structurally conserved "cores" in families of proteins [1]. These facts are difficult to interpret because most of the studies that have been made are purely empirical. To go forward in this sense, it has been developed the mechanistic model "Linearly Forced âĂŞ Elastic Network Model" (LF - ENM), which predicts the change in the equilibrium position of the sites as a result of random mutations, not subjected to natural selection. Applying this model, it was shown that the patterns of structural change (greater contribution of lower energy normal modes and the existence of a structurally conserved "core") can be predicted without resorting to natural selection [3,4]. With regard to the dynamics, this model reproduces well the observed pattern that the normal modes of lower energy are more evolutionarily conserved, even in the absence of natural selection [5]. This implies that such modes are more robust to random mutations, so that they would conserve more even in the absence of natural selection. All these results 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 and dynamical divergence patterns, at the level of aminoacid sequences, different sites evolve at different speeds. 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. Recently, we have proposed a mechanistic model ("Stress Model") in which a mutation is accepted at a rate proportional to the probability that the mutant adopts the "active structure". This model has been used to account for the average evolutionary variation from site to site [6]. As we said, the LF-ENM model has been successfully used to explain the observed patterns of structural and dynamical divergence in the absence of natural selection. However, it is expected that the selection restricts the structural divergence and / or dynamics. For example if the structure and fluctuations of an enzyme‘s active site are important for enzymatic activity, one would expect that the structure and movements are evolutionarily conserved significantly higher than expected for purely mutational patterns. Therefore, no matter how weak, some evidence of natural selection at the level of structural and dynamical divergence would be expected. The aim of this work is to find evidence of natural selection at the structural divergence level.

3 Methods

ENM of a reference protein:

We considered 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:

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

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^0_{ij} is the equilibrium (native) 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: $d^0_{ij} \rightarrow d^0_{ij} + \Delta_{ij}$, 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 $Var(\Delta_i j) = \sigma^2$. Following this, the mutant's potential is of the form:

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

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:

$$f_{ij} = k_{ij}\Delta_{ij} \quad (3)$$

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:

$$d\mathbf{r}^0 \equiv \mathbf{r}^0_{mut} - \mathbf{r}^0_{wt} = \mathbf{K}^{-1}_{wt}\mathbf{f} \quad (4)$$

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. This equation shows that the structural perturbation introduced by a mutation is related to the mutation and to the network of oscillators, via $\mathbf{K}^{-1}_{wt}$. We should note here that $\mathbf{K}^{-1}_{wt}$ is actually the pseudo inverse, because $\mathbf{K}_{wt}$ has six zero eigenvalues, corresponding to translations and rotations, so that it is not invertible.

Stress Model of protein evolution

We have described how to model mutations, but we have not described yet the way we select them. The stress model of protein evolution predicts the probability of acceptance of mutations. As the LF - ENM, it is based on an Elastic Netowork Model in which mutations of an ancestor protein are modeled as perturbations to the spring lengths that connect the mutated site with its neighbors.

The stress model predicts that the probability of accepting such a mutant is:

$$P_{accept} = e^{-\beta\frac{1}{2}\sum_{i<j} k_{ij}\Delta^2_{ij}} \quad (5)$$

Therefore, from Eq. 5 and 3 we find:

$$P_{accept} = e^{-\beta \frac{1}{2} \sum_{i<j} f_{ij}^2 / k_{ij}} \quad (6)$$

For the special case of ENM "Anisotropic Network Model" (ANM), k_{ij} are either 0 or 1, as we explain later, and we get:

$$P_{accept} = e^{-\beta \frac{1}{2} \sum_{j \sim i} f_{ij}^2} \quad (7)$$

Where the sum $j \sim i$ is over all j that are in contact with the mutated site i (i.e. for which $k_{ij} \neq 0$.

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. The selected families and their characteristics are shown in Table 1.

Selection of the reference protein:

For each family of the dataset we selected a reference "ancestor" protein, which we consider is the most structurally representative protein of the family. To get this protein, we first calculated the average structure of each multiple alignment. Then, we calculated the "Mean Square Deviation" (MSD) between the structure of each protein of the family and the obtained average structure. Finally, we selected as the reference protein the one with the lower value of MSD.

Alignments analysis:

For each family of proteins, the reference protein was aligned with each of the other proteins of the family. Then, the aligned and nonaligned sites, their coordinates, and the sequence identity for each pair of proteins were obtained.

Theoretical dataset:

To generate a theoretical dataset that is comparable with the experimental dataset, for each family, we aligned the ancestor with the other proteins, we calculated the number of mutated sites n and we created a lineage of 100 simulated mutants following a path of substitutions until we had mutated n sites. This path is composed of various evolutionary steps each of them comprising a single substitution.

One evolutionary step:

Let us define a time-step such the time-step when there is a single substitution event (an accepted mutation) for the whole protein. To simulate such an event we did the following:

1. We picked one random site l of the reference protein.

2. We introduced a "trial" mutation by obtaining a set of forces f_{lj} for each of the j contacts of site l and the reaction force over site l.

3. We calculated the probability of accepting this trial mutation: $P_{accept} = e^{-\beta \frac{1}{2} \sum_{j \sim i} f_{lj}^2}$

4. We calculated the logical variable $Accept = P_{accept} \geq runif(1,0,1)$; Accept will be TRUE with probability P_{accept}.

5. If Accept was TRUE, we accepted the mutation (i.e. the new ancestor was the generated mutant) and the evolutionary step was finished (i.e. we found an accepted mutation: one substitution). Else, we rejected the trial mutation and tried again (i.e. went back to Step 1).

An evolutionary path of substitution at different sites:

We wanted to simulate a lineage (an evolutionary path) such that it started at a known ancestor (the reference protein) and it ended when n sites had accepted mutations. ($n \leq N$, where N is the sequence length). n corresponds to the number of mutated sites of the ancestor compared with one of the other proteins of the dataset. To do this, for each pair of proteins, we repeated the single evolutionary step of the previous paragraph n times, making sure that the set of sites where we tried mutations did not include sites that had previously accepted mutations. We calculated 100 proteins of each lineage.

Regimens of selection:

To account for different regimens of selection we gave β different values. This differente regimens corresponds to:

No selection: Paccept ≈ 1

Weak selection: Paccept ≈ 0.9

Medium selection: Paccept ≈ 0.5

Strong selection: Paccept ≈ 0.1

To gel the β values for each regimen we did as follows:

$$\beta^{reg} = \frac{-log(Paccept^{reg})}{(< f_{ij}^2 > \times < CN >)} \quad (8)$$

being $< CN >$ the average number of contacts of the reference protein.

Star tree

Our experimental data are sets of proteins, one of which we consider as the reference "ancestor" protein. In principle, we should infer the phylogenetic tree and try to simulate structures with our model following a tree with the same topology. However, we assume that the results are not too sensitive with respect to tree topology so that we can approximate the tree by a "star tree". The common ancestor of all lineages of our star tree is our "reference" protein. Then, each lineage corresponds to a pair alignment of each protein with this protein. Thus, different lineages have different "branch lengths", were we assume the branch length is the number of sites in which the sequence of the ancestor and the tip of the lineage differ.

Model parameters

To completely specify the model, we must specify parameters for $\mathbf{K}_{wt}$ and $\mathbf{f}$. As we mentioned before, we used the "Anisotropic Network Model" (ANM). Following this model, we gave a spring force constant of 1 to sites at a distance $\leq$ 10 Å and a spring force constant of 0 to sites at a distance $>$ 10 Å. We selected 10 Å as the cut off value after optimization using a range of values form 8 to 12 (data not shown). This cut off value is also the most commonly used. To calculate $\mathbf{f}_{lj}$, given a mutation at a

site l, each site j in contact with l is assigned a force directed along the $l-j$ contact and site i is assigned a reaction force. To simulate random mutations, the magnitudes of $\mathbf{f}_{lj}$ were randomly picked from a uniform distribution of Δ_{lj} 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 ($f \leq f_{max}$).

Structural variability measures:

We obtained the coordinates of the aligned sites of each protein. For the theoretical dataset we considered that there are not nonaligned sites. Then, structural variability measures were calculated in cartesian coordinates and projected onto the normal modes coordinates:

Cartesian coordinates:

Structural variation of each protein was calculated relative to the reference protein into the aligned sites. To do this, for each aligned site, we calculated the square deviation:

$$\left\|d\mathbf{r}_i^0\right\|^2 = dx_i^{0,2} + dy_i^{0,2} + dz_i^{0,2} \quad (9)$$

Were $d\mathbf{r}_i^0 = (d\mathbf{x}_i^0 d\mathbf{y}_i^0 d\mathbf{z}_i^0)^T$ is the column vector of cartesian displacements of the i^{th} C_α with respect to the reference structure. To diminish noisy information, we smoothed these profiles as shown:

$$\left\|d\mathbf{r}_i^0\right\|^2_{smooth} = \frac{(mathbfC \times \left\|d\mathbf{r}_i^0\right\|^2)}{\sum_i \mathbf{C}} \quad (10)$$

Being $\mathbf{C}$ a $N \times N$ matrix with 1 in the diagonal and in the contacts and 0 elsewhere. Finally, since we are only interested in the relative variability of different sites, deviations obtained using Eq. 10 were normalize so that they add up to 1.

Normal modes coordinates:

Analysis of structural change of normal modes was calculated by projecting structural differences of aligned sites of each protein onto the normal modes of the reference protein. The normal modes were obtained by solving the equation:

$$\mathbf{K}\mathbf{q}_n = \lambda_n \mathbf{q}_n \quad (11)$$

Being $\mathbf{q}_n$ the eigenvectors and λ_n their eigenvalues. For proteins that do not align with all the sites of the reference protein, instead of $\mathbf{K}$, we used $\mathbf{K}_{eff}$, whose normal modes describe the motions of the aligned sites only (see reference [x]). Then, for a protein with structural variation of the aligned sites $d\mathbf{r}^0$, the projection onto the normal modes was calculated as follows:

$$P_n \equiv \frac{(\mathbf{q}_n^T d\mathbf{r}^0)^2}{\sum_n (\mathbf{q}_n^T d\mathbf{r}^0)^2} \quad (12)$$

Profile comparisons:

Cartesian coordinates:

For the theoretical datasets and for the experimental dataset we averaged

$\left\|d\mathbf{r}_i^0\right\|^2$ over each site i of the reference protein to obtain profiles of MSD_i. Then, the theoretical average profiles were fitted with the experimental profiles. Lastly, for each dataset, we concatenated the profiles obtained for all of the families to obtain a sole profile.

Normal modes coordinates:

For the theoretical datasets, the 100 x P profiles of $P_n$were split in 100 groups so that in all of them there was a member of each lineage. Then we calculated, for these 100 groups, the average and 0.05 and 0.95 quantiles. Finally, we calculated the average of averages and the average of quantiles. For the experimental dataset, the average and 0.05 and 0.95 quantile profiles of each family were calculated.

4 Results and discussion

Cartesian coordinates:

In order to analyze the effect of natural selection on the structural variability in cartesian coordinates we calculated unique per dataset MSD_i profiles as explained in Methods. These profiles contain information of all of the families. Then, we calculated the R^2 and the MSE between the experimental dataset compared with theoretical datasets (with and without natural selection). The results are shown in figure 1. Figue 1 shows that the R^2 between the experimental dataset and theoretical datasets, with or without natural selection, is high, ≈ 0.5, and that accounting for natural selection does not seem to improve the agreement.

Normal modes coordinates:

In order to analyze the effect of natural selection on the structural variability projected onto normal modes coordinates of the reference protein, we obtained average P_n profiles and 0.05 and 0.95 quantiles profiles as explained in Methods. To compare the experimental dataset with the theoretical datasets we calculated the R^2 and the MSE between them. The results are shown in Table 2 and figure 2. Table 1 and figure 2 show that, again, the R^2 between experimental data and theoretical data, with or without natural selection, is high, ≈ 0.7 and that natural selection does not seem to improve the agreement with experimental data.

5 Conclusion