# Structural Variability

Maria Laura Marcos

April 11, 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 "wilt type" 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

**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 proteins we selected a reference "wilt type" 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.

**ENM of the reference protein:**

We considered the backbone fluctuations of the reference 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 the alpha carbons ($C_\alpha$) connected by springs. In general, the ENM potential is of the form:

$$V(r) = 1/2 * (r - r_0)^T K (r - r_0) \tag{1}$$

where, for a protein of $N$ sites, $r$ is a column vector with $3N$ elements: the $x$, $y$, $z$ coordinates of the $N$ $C_\alpha$, $r_0$ is the equilibrium structure, and $K$ is the "stiffness" matrix, which represents the network, s topology and spring force constants. Specifically, 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 Åand a spring force constant of 0 to sites at a distance $>$ 10 Å. We selected 10 Åas 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.

**LF-ENM model of mutant proteins:**

To simulate mutants of each reference protein and thus to generate datasets of theoretical proteins, we used the "Linearly Forced Elastic Network Model" (LF - ENM). This model predicts the effect of a single mutation adding a linear perturbative term to the reference potential:

Agregar formula

where $f$ is a column vector with $3N$ elements that models the mutation. The equilibrium structure of the mutant $r^0_{mut}$ is the value of $r$ that minimizes $V_{mut}$. Using Eqs. (1) and (2) we find the structural variation due to the mutation:

Agregar formula

This equation shows that the structural perturbation introduced by a mutation is related to the mutation and to the network of oscillators, via $K^1$. We should note here that $K^1$ is actually the pseudo inverse, because $K$ has six

zero eigenvalues, corresponding to translations and rotations, so that it is not invertible.

**Multiple-sites mutants:** In order to simulate adequately the experimental data, we did not simulated proteins with a single mutation, we did simulate proteins with multiple mutations. To do this, we considered additive mutations by the assumption that $K^{mut} \cong \mathrm{K}^{wt}$:

$$dr^{tot} = dr_{wt}^{mut1} + dr_{wt}^{mut2} + ... + dr_{wt}^{mutNmut} = K_{wt}^{-1} \times (fmut1 + fmut2 + ... + fmutNmut) \tag{2}$$

Where $dr^{tot}$ is the total structural variation due the M mutations, $\mathrm{dr}_{wt}^{mutm} is the structural variation produ$ and fmutn is the vector of force of $n$.

# 4 Results and discussion

# 5 Conclusion