**Paper 2016**

**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 obtained 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 used the LF – ENM to generate mutants with multiple mutations. We did as much mutations as it corresponds to the percent sequence identity of paired proteins of each family. We generated 2 sets of mutants. In one set, we mutated sites randomly and, on the other set, we accounted for natural selection mutating the 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 in 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.

**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 – 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.

To study more deeply the structural divergence, we followed a different approach than what it was done before. We selected different families of proteins with representatives of the major structural classes, we obtained their multiple structural alignments, and we simulated mutants with no single but multiple mutations. We did as much mutations as it corresponds to the percent sequence identity of paired proteins of each family. In the selection of which sites to mutate is where we included or not the natural selection. In one set of simulated mutants, we mutated sites randomly and, on the other set, we accounted for natural selection following the probabilities of mutation given by the “Stress Model”.

**Methods**

*Experimental dataset: families of proteins*

We selected 8 protein families 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 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.

Table 1: characterization of protein families

*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, first, we 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 the protein with the lower MSD as the reference protein.

*Alignments* *analysis*

For each family of proteins, the reference protein was aligned with each of the other proteins of the family, the aligned and nonaligned sites, their coordinates, and the average sequence identity for each pair 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 (Cα) connected by springs. 4–8 In general, the ENM potential is of the form

where, for a protein of N sites, r is a column vector with 3N elements: the x, y, z coordinates of the N Cα, r0 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 ≤ 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 not the most commonly used (refs.)

*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 r0mut is the value of r that minimizes Vmut. Using Eqs. (1) and (2) we find the structural variation due to the mutation:

This equation shows that the structural perturbation introduced by a mutation is related to the mutation (f) 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 more adequately the experimental data, we did not simulated proteins with a single mutation, we did simulate proteins with multiple mutantions. To do this, we considered additive mutations by the assumption that Kmut ≅ Kwt:

drtot = drmut1wt + drmut2wt + … + drmutNmutwt = Kwt-1 x (fmut1+fmut2+ …. + fmutNmut)

*ver formula*

Theoreticas datasets:

In order to study the effect of natural selection on structural divergence, for each family, we generated two datasets of simulated mutants. In one dataset, we applied natural selection at sequence level and, on the other dataset, we did not consider natural selection:

**Mutants with natural selection (NS = T):** we applied natural selection at the sequence level. To determine which sites to mutate accounting natural selection, we followed the “stress model”, by which probabilities of mutation are assigned for each site of the reference protein according to the eq:

prob.mut.i = 1 - (beta \* CN.i).

Were CNi is the number of contacts (sites with a distance ≤ 10 Å in this case) of site i.

To select the number of sites to mutate, we took information from the alignments. For each pair comprising the reference protein and one of the other proteins of the family, we obtained the average sequence identity (% ID) within the aligned region. Then, the number of mutated sites for this pair is

n.sites.mut = (100 - %id) x n.aa / 100

Being n.aa the number of aminoacids of the reference protein.

Then, for each of these pairs, we generated 10 mutants of n.sites.mut, where the mutated sites are the ones with more probabilities to mutate following the stress model.

In this way, if the family has M proteins (plus the reference protein), we obtained a simulated dataset of 10 x M members.

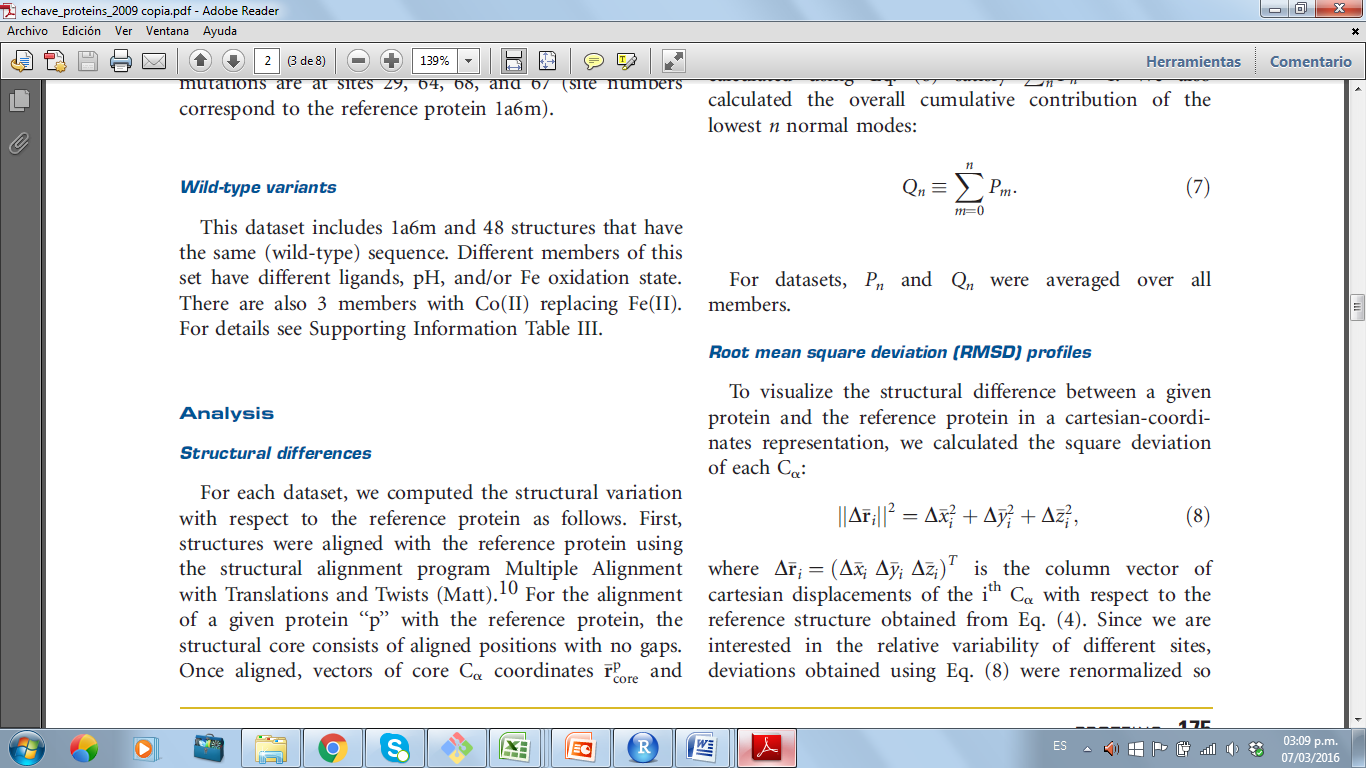
**Mutants without natural selection (NS = F):** For each pair comprising the reference protein and one of the other proteins of the family, we obtained the number of sites to mutate as mentioned before and we mutated randomly this number of sites. For each of these pairs, we generated 10 mutants. Hence, again, if the family has M proteins plus the reference protein, we obtained a simulated dataset of 10 x M members.

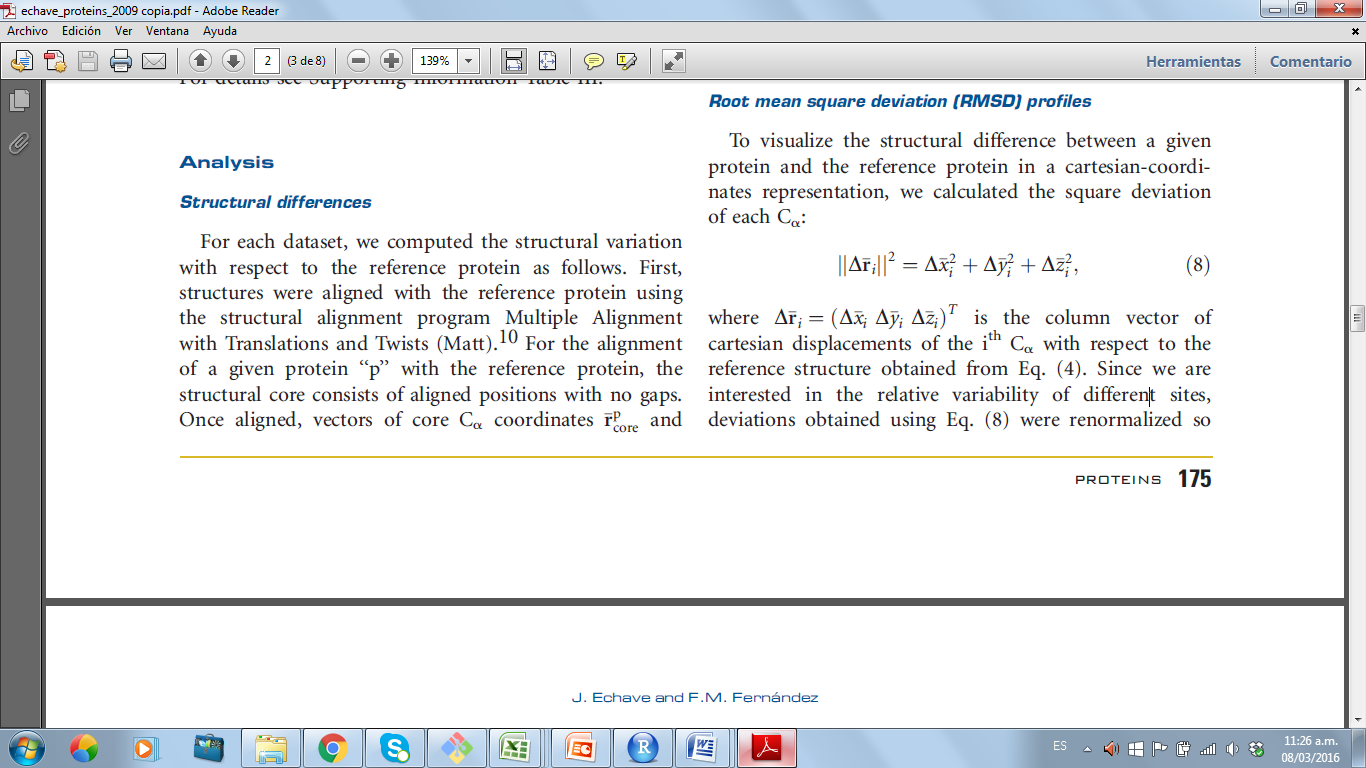
*Structural variability measures*

For all families and for all sets, experimental and theoretical (with or without natural selection), 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:



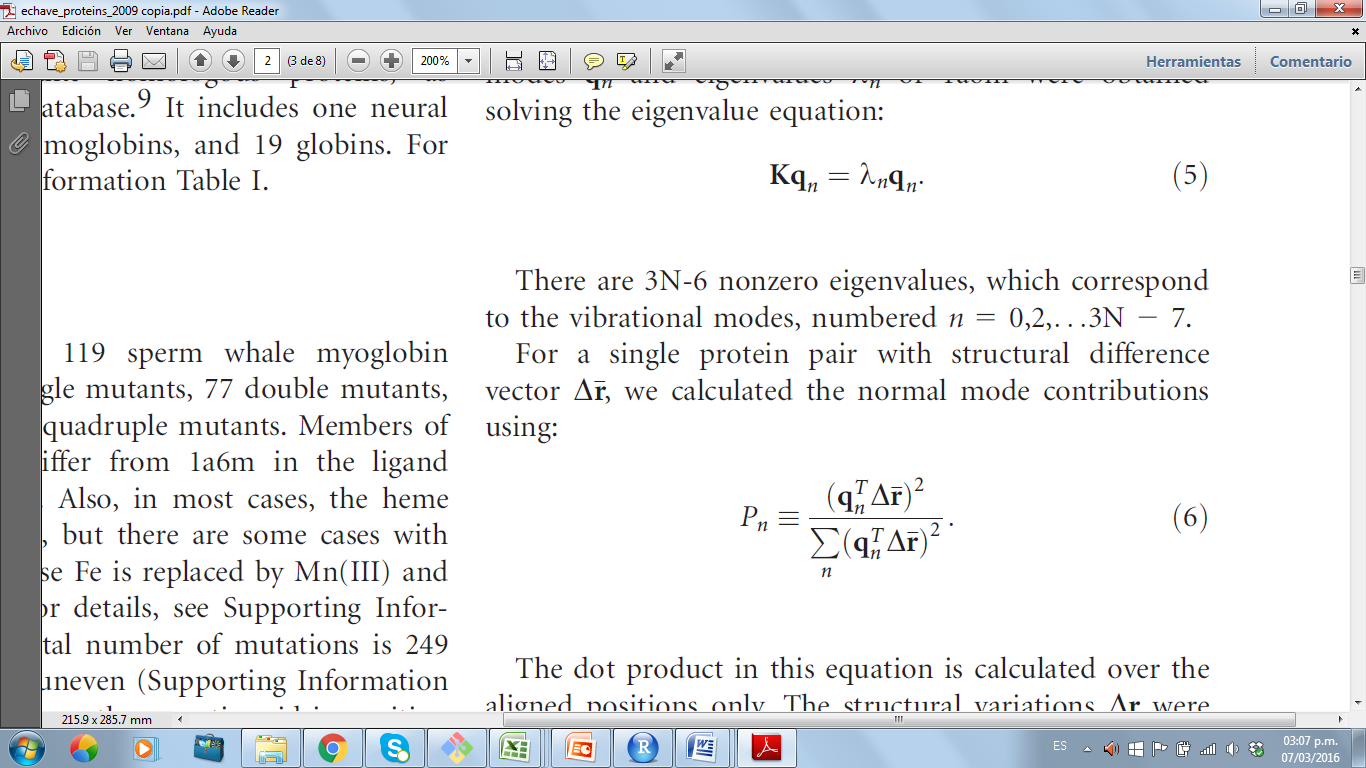
Were  is the column vector of cartesian displacements of the ith Ca with respect to the reference structure obtained from Eq. (4). To diminish noisy information, we smoothed these profiles as shown:

Formula

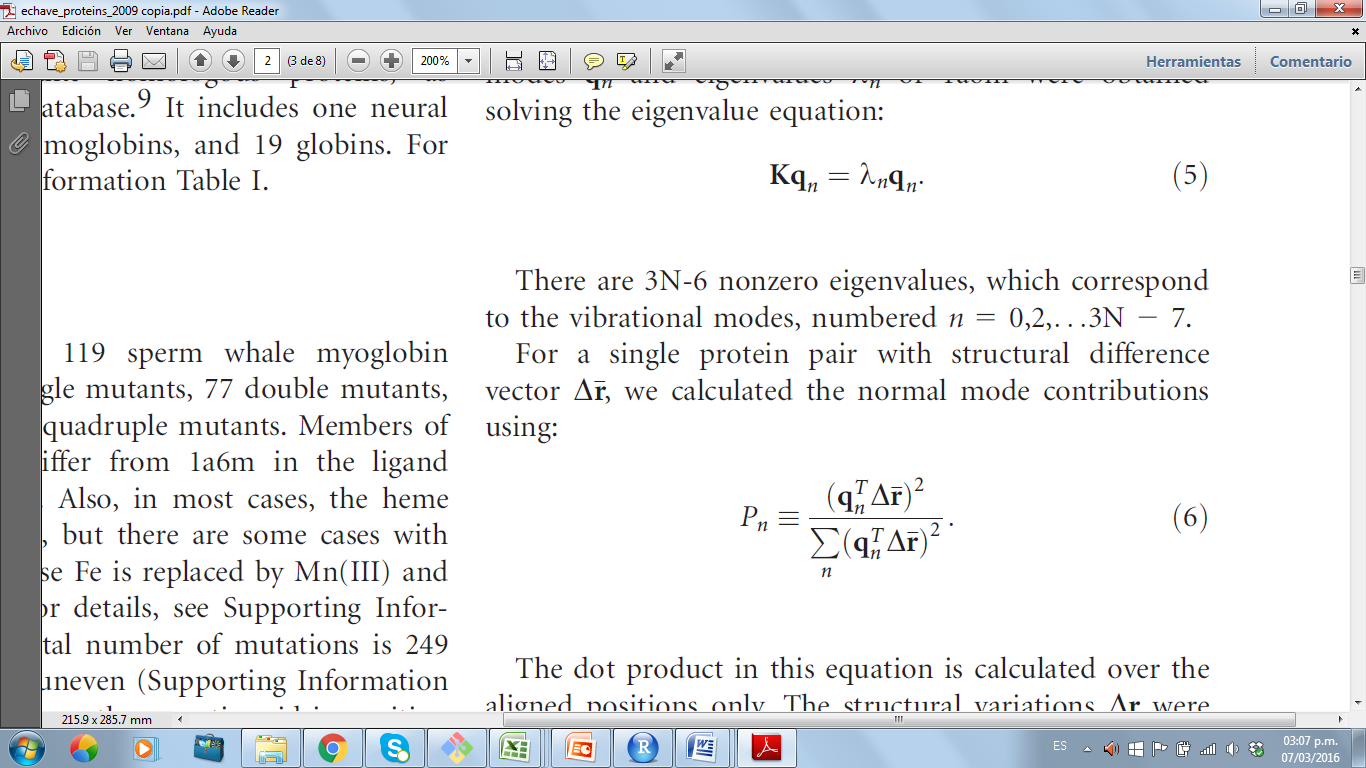
Then, since we are only interested in the relative variability of different sites, deviations obtained using Eq. () were normalize so that they add up to 1.

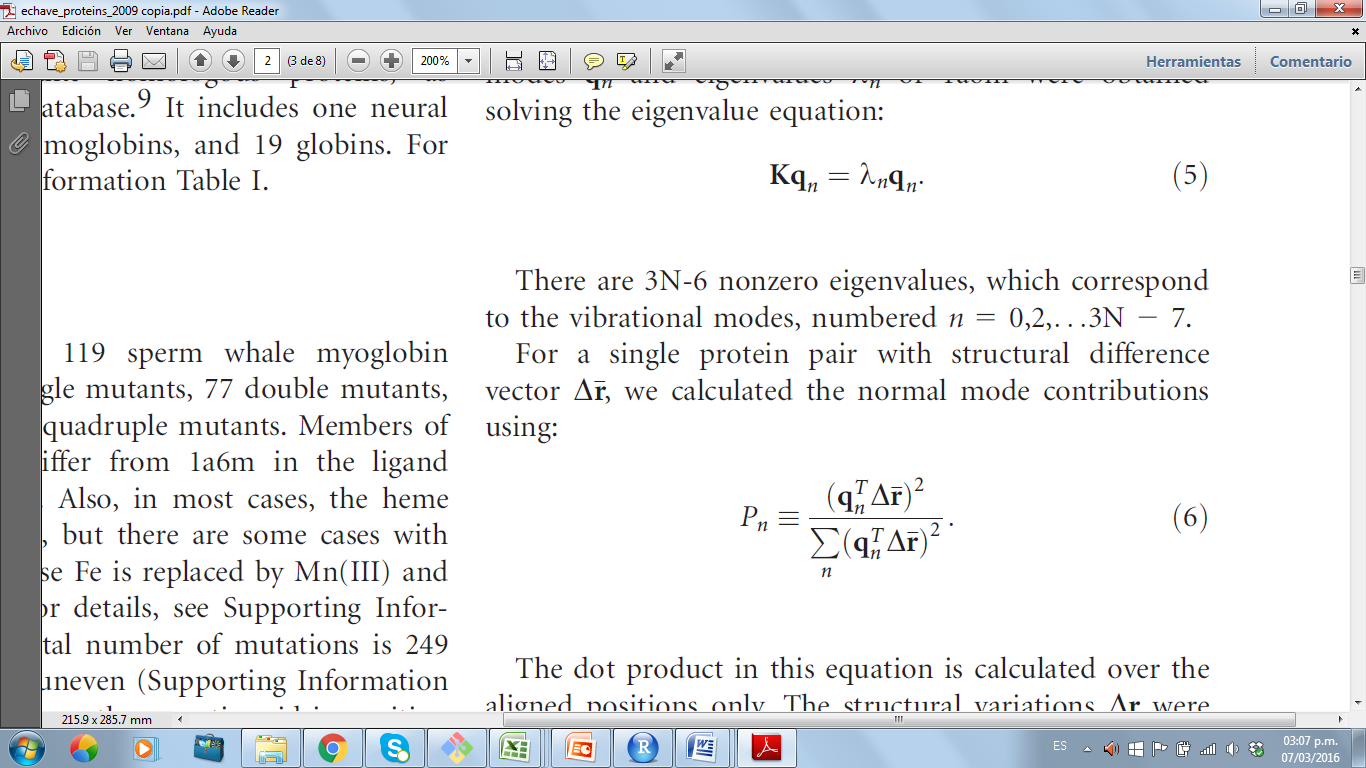
**Normal modes:**

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



For proteins that do not align in all reference sites instead of K, Keff was used, which allows for normal modes describing the motions of the aligned sites (ref).

Then, for a protein with structural variation of the aligned sites , the projection onto eigenvectors was calculated as follows:



**Profile comparisons:**

To make comparisons between theoretical and experimental structural divergence measures, was preceded as follows:

• theoretical (with or without natural selection) measures: the (10 x M) Pn and MSDi profiles of each family were regrouped in 10 sub sets so that in all of them there is a simulated mutant that corresponds to an experimental pair of proteins in the % ID. Then we calculated for these 10 sets, the average and 0.05 and 0.95 quantiles. Finally, we calculated the average of averages and the average of quantiles.

• Experimental measures: the average and 0.05 and 0.95 quantile profiles MSDI Pn and each family was calculated.

Subsequently, the theoretical average profile of each measure was fitted with the experimental profile and R2 correlation between these profiles were calculared. In the case of MSDi, we concatenated profiles obtained for each family and then we calculated de R2.

**Resultados**

MUESTRO GRAFICO DE MSDI CONCATENADO

MUESTRO IMÁGENES DE JMOL DE TODAS LAS FAMILIAS CON DR EN B-FACTOR.

MUESTRO GRAFICOS PN DE TOTAS LAS FAMILIAS Y TABLAS CON CORRELACIONES Y MSE