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 to create lineages. 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 other cases, we accounted for natural selection acepting mutations according to the ``Stress Model'' with different selection regimens. 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.

The main ideas of the Darwin - Wallace theory of evolution of species are that organisms are the product of an evolutionary history that implies hereditary changes of a common ancestor and the subsequent selection of more favorable changes, procces called ``natural selection". The role of natural selection is widely accepted regarding morphological characters. However, the importance of natural selection at the molecular level is a matter of ongoing debate .

Proteins are the molecules that decodify hereditary information and they are known to diverge during biological evolution. At the level of aminoacid sequences, there is a clear evidence of natural selection: different aminoacids evolve at different speeds. However, at the level of structure, it has been rouhgly shown that structural divergence patterns can be well reproduced without resorting to natural selection.

Here we give convincing evidence that there is no evidence of natural selecion on the structural variation of proteins neither in cartesian coordinates nor projected onto the vibrational modes of proteins.

belief that evolutionary theory applied to molecules

belief that evolutionary theory applied to molecules

The Darwin - Wallace theory of evolution of species relates that organisms are the product of an evolutionary history that implies hereditary changes of a common ancestor and the subsequent selection of more favorable changes, procces called ``natural selection". Although the role of natural selection is widely accepted regarding to morphological characters, its importance at the molecular level is matter of debate.

Proteins are the molecules that decodify hereditary information and they are known to diverge during biological evolution. At the level of aminoacid sequences, there is a clear evidence of natural selection: different aminoacids evolve at different speeds. However, at the level of structure, it has been rouhgly shown that structural divergence patterns can be well reproduced without resorting to natural selection.

Here we give convincing evidence that there is no effect of natural selecion on the structural variation of proteins neither in cartesian coordinates nor projected onto the vibrational modes of proteins. This results go against the belief that the evolutionary theory of species applies to molecules.

The Darwin - Wallace theory of evolution of species relates that organisms are the product of an evolutionary history that implies hereditary changes of a common ancestor and selection of more favorable changes, procces called ``natural selection". Although the role of natural selection is widely accepted regarding morphological characters, its importance at the molecular level is matter of debate.

Proteins decodify hereditary information and thus diverge during biological evolution. At the level of aminoacid sequences, there is a clear evidence of natural selection: different aminoacids evolve at different speeds. However, at the level of structure, it has been rouhgly shown that structural divergence patterns can be reproduced without resorting to natural selection.

Here, we give convincing evidence that there is no effect of natural selecion on the structural variation of proteins.

This results go against the belief that the evolutionary theory of species applies to molecules.

The Darwin - Wallace theory of evolution of species relates that organisms are the product of an evolutionary history that implies hereditary changes of a common ancestor and selection of favorable changes, procces called ``natural selection". Although the role of natural selection is widely accepted regarding morphological characters, its importance at the molecular level is matter of debate.

Proteins decodify hereditary information and thus diverge during biological evolution. At the level of aminoacid sequences, there is a clear evidence of natural selection: different aminoacids evolve at different speeds. At the level of strucute, the are not works that elucidate the rol of natural selection .

Here, we give convincing evidence that there is no effect of natural selecion on the evolutionary structural variation of proteins. This results go against the belief that the evolutionary theory of species applies to the evolution of molecules.

The Darwin - Wallace theory of evolution of species relates that organisms are the product of an evolutionary history that implies hereditary changes of a common ancestor and selection of favorable changes, procces called ``natural selection". Although the role of natural selection is widely accepted regarding morphological characters, its importance at the molecular level is matter of debate.

Proteins decodify hereditary information and thus diverge during biological evolution. At the level of aminoacid sequences, there is a clear evidence of natural selection: different aminoacids evolve at different speeds. However, at the level of strucute, the rol of natural selection has not been elucidated yet.

Here, we give convincing evidence that there is no effect of natural selecion on the evolutionary structural variation of proteins. This result go against the common belief that the evolutionary theory of species applies to molecules and is a step forward in understandig evolution.

{\subsection\* {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 ($C\_{\alpha}$) connected by springs. In general, the ENM potential is of the form:

\begin{equation} \label{eq:Vwt}

V\_{wt} = \frac{1}{2} \sum\_{i < j} k\_{ij} (d\_{ij} - d ^ 0\_{ij}) ^ 2

\end{equation}

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.

{\subsection\* {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 $\Delta\_{ij}$ 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:

\begin{equation} \label{eq:Vmut}

V\_{mut} = \frac{1}{2} \sum\_{i<j} k\_{ij} [d\_{ij}-(d^0\_{ij}+\Delta\_{ij})]^2

\end{equation}

Then, the LF - ENM is obtained from expanding Eq. \ref{eq:Vmut} 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:

\begin{equation} \label{eq:fij}

f\_{ij} = k\_{ij} \Delta\_{ij}

\end{equation}

Finally, the equilibrium structure of the mutant $\mathbf{r}^{0}\_{mut}$ is the value of $\mathbf{r}$ that minimizes $V\_{mut}$. Using Eqs. \ref{eq:Vwt} and \ref{eq:Vmut} and after some algebra we find the structural variation due to the mutation of a reference protein of $N$ sites:

\begin{equation} \label{eq:dr}

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

\end{equation}

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.

\subsection\*{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:

\begin{equation} \label{eq:Paccept}

P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{i<j} k\_{ij} \Delta\_{ij}^2}

\end{equation}

Therefore, from Eq. \ref{eq:Paccept} and \ref{eq:fij} we find:

\begin{equation} \label{eq:PacceptFij}

P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{i<j} f\_{ij}^2/k\_{ij}}

\end{equation}

For the special case of ENM ``Anisotropic Network Model'' (ANM), $k\_{ij}$ are either 0 or 1, as we explain later, and we get:

\begin{equation} \label{eq:PacceptANM}

P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{j\sim i} f\_{ij}^2}

\end{equation}

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

\subsection\*{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.

\subsubsection\*{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.

\subsubsection\*{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.

\subsection\*{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, for each of the pairs, 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.

\subsubsection\*{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:

\begin{enumerate}

\item We picked one random site $l$ of the reference protein.

\item 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 $l$.

\item We calculated the probability of accepting this trial mutation: $P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{j\sim i} f\_{lj}^2}$

\item We calculated the logical variable $\text{Accept} = P\_{accept} \ge \text{runif(1,0,1)}$; Accept will be TRUE with probability $P\_{accept}$.

\item 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).

\end{enumerate}

\subsubsection\*{An evolutionary path of substitution at different sites:}

As we mentioned before 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. ($n \le N$). $n$ 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.

\subsubsection\*{Regimens of selection:}

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

No selection: Paccept $\approx 1$

Weak selection: Paccept $\approx 0.9$

Medium selection: Paccept $\approx 0.5$

Strong selection: Paccept $\approx 0.1$

Since $P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{j\sim i} f\_{lj}^2}$ and because $<f^{2}\_{li}>$ is independent of the site, each $\beta$ only depends on this magnitud, on the average contact number of the sites of the reference protein $<CN>$ and on the selected $P\_{accept}$. Then, to gel the $\beta$ values for each regimen we did as follows:

\begin{equation} \label{eq:PacceptFij}

\beta^{reg} = \frac{- log(Paccept^{reg})}{(1/2 \times <f\_{ij}^2> \times <CN>)}

\end{equation}

\subsubsection\*{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.

\section\*{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 $\AA$ and a spring force constant of 0 to sites at a distance > 10 $\AA$. We selected 10 $\AA$ 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 $\Delta\_{lj}$ 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 ($f \leq f\_{max}$).

\subsection\*{ 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:

\subsubsection\*{ 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:

\begin{equation} \label{eq:drSquarei}

\left\|d\mathbf{r}^{0}\_{i}\right\|^{2}=dx^{0,2}\_{i}+dy^{0,2}\_{i}+dz^{0,2}\_{i}

\end{equation}

Were $d\mathbf{r}^{0}\_{i}=(d\mathbf{x}^{0}\_{i}d\mathbf{y}^{0}\_{i}d\mathbf{z}^{0}\_{i})^{T}$ is the column vector of cartesian displacements of the $i^{th}$ $C\_{\alpha}$ with respect to the reference structure.

To diminish noisy information, we smoothed these profiles as shown:

\begin{equation} \label{eq:SmoothdrSquarei}

\left\|d\mathbf{r}^{0}\_{i}\right\|^{2}\_{smooth} = \frac{(\mathbf{C} \times \left\|d\mathbf{r}^{0}\_{i}\right\|^{2})}{\sum\_{i} \mathbf{C}}

\end{equation}

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. \ref{eq:SmoothdrSquarei} were normalize so that they add up to 1.

\subsubsection\*{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:

\begin{equation}

\mathbf{K}\mathbf{q}\_{n} = \lambda\_{n}\mathbf{q}\_{n}

\end{equation}

Being $\mathbf{q}\_{n}$ the eigenvectors and $\lambda\_{n}$ 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:

\begin{equation}

P\_{n} \equiv \frac{(\mathbf{q}^{T}\_{n}d\mathbf{r}^{0})^2}{\sum\_{n}(\mathbf{q}^{T}\_{n}d\mathbf{r}^{0})^2}

\end{equation}

\subsection\*{ Profile comparisons:}

\subsubsection\*{Cartesian coordinates:} For the theoretical datasets and for the experimental dataset we averaged $\left\|d\mathbf{r}^{0}\_{i}\right\|^{2}$ 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.

\subsubsection\*{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.

\end{document}

{\subsection\* {ENM of a 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 ($C\_{\alpha}$) connected by springs. In general, the ENM potential is of the form:

\begin{equation} \label{eq:Vwt}

V\_{wt} = \frac{1}{2} \sum\_{i < j} k\_{ij} (d\_{ij} - d ^ 0\_{ij}) ^ 2

\end{equation}

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.

{\subsection\* {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 $\Delta\_{ij}$ 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:

\begin{equation} \label{eq:Vmut}

V\_{mut} = \frac{1}{2} \sum\_{i<j} k\_{ij} [d\_{ij}-(d^0\_{ij}+\Delta\_{ij})]^2

\end{equation}

Then, the LF - ENM is obtained from expanding Eq. \ref{eq:Vmut} 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:

\begin{equation} \label{eq:fij}

f\_{ij} = k\_{ij} \Delta\_{ij}

\end{equation}

Finally, the equilibrium structure of the mutant $\mathbf{r}^{0}\_{mut}$ is the value of $\mathbf{r}$ that minimizes $V\_{mut}$. Using Eqs. \ref{eq:Vwt} and \ref{eq:Vmut} and after some algebra we find the structural variation due to the mutation of a reference protein of $N$ sites:

\begin{equation} \label{eq:dr}

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

\end{equation}

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.

\subsection\*{Stress Model of protein evolution}

The stress model of protein evolution predicts the probability of acceptance of mutations. 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:

\begin{equation} \label{eq:Paccept}

P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{i<j} k\_{ij} \Delta\_{ij}^2}

\end{equation}

Therefore, from Eq. \ref{eq:Paccept} and \ref{eq:fij} we find:

\begin{equation} \label{eq:PacceptFij}

P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{i<j} f\_{ij}^2/k\_{ij}}

\end{equation}

For the special case of ENM ``Anisotropic Network Model'' (ANM), for wich $k\_{ij}$ are either 0 or 1, we get:

\begin{equation} \label{eq:PacceptANM}

P\_{accept} = e^{-\beta \frac{1}{2} \sum\_{j\sim i} f\_{ij}^2}

\end{equation}

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

\subsection\*{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 of the dataset we selected a reference ``ancestor'' protein. To get this protein, we calculated the average structure of each multiple alignment and selected the protein with the lower ``Mean Square Deviation'' (MSD) between its structure and the average structure.

For each family of proteins, the reference protein was aligned with each of the other proteins of the family and we considered these pairs differents lineages. 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.

\subsection\*{Theoretical dataset:}

To create a theoretical dataset that is comparable with the experimental dataset, for each lineage we generated 50 mutants following a path of substitutions according to the number of mutated sites of the lineage. This path is composed of various evolutionary steps each of them comprising a single substitution. These steps was simulated 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 probability of accepting this trial mutation from Eq. \ref{eq:PacceptANM} and calculating the logical variable $\text{Accept} = P\_{accept} \ge \text{runif(1,0,1)}$. If Accept was TRUE, we accepted the mutation and the evolutionary step was finished. Else, we rejected the trial mutation and tried again.

\subsubsection\*{Regimens of selection:}

We simulated mutants with different regimens of selection:

No selection: $<Paccept> \approx 1$

Weak selection: $<Paccept> \approx 0.9$

Medium selection: $<Paccept> \approx 0.5$

Strong selection: $<Paccept> \approx 0.1$

To get these $<Paccept>$ we gave $\beta$ different values:

\begin{equation} \label{eq:beta}

\beta\_{reg} = −log(<Paccept>\_{reg})/(1/2× < f^{2}\_{ij}> × < CN >)

\end{equation}

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

\section\*{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 $\AA$ and a spring force constant of 0 to sites at a distance > 10 $\AA$. We selected 10 $\AA$ 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 $\Delta\_{lj}$ 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 ($f \leq f\_{max}$).

\begin{equation} \label{eq:drSquarei}

\left\|d\mathbf{r}\_{i}\right\|^{2}=dx^{2}\_{i}+dy^{2}\_{i}+dz^{2}\_{i}

\end{equation}

$d\mathbf{r}\_{i}=(d\mathbf{x}\_{i}d\mathbf{y}\_{i}d\mathbf{z}\_{i})^{T}$

\begin{equation}

P\_{n} \equiv \frac{(\mathbf{q}^{T}\_{n}d\mathbf{r}^{0})^2}{\sum\_{n}(\mathbf{q}^{T}\_{n}d\mathbf{r}^{0})^2}

\end{equation}