Supplement. Derivation of the Functional Impact Score in the molecular field approximation (B. Reva)

A protein molecule is a subject of random mutations. Mutations can make a significant impact on a protein function and, hence, on cells and the organism. Mutations that reduce the chances of organisms to leave the offsprings are usually eliminated by natural selection. On the contrary, mutations advantageous for population survival have higher chances of being fixed in the population.

A mutation can affect a protein function either by changing the stability of a protein or by changing the stability of the specific interactions with the native protein ligands. Each residue of a protein makes a certain contribution to the stability of the protein. This contribution is simply a sum of all interactions between a residue and the rest of the molecule. Similarly, each residue makes a certain contribution to the specific interactions of the molecule with the native ligands. For those residues that do not participate in functional interactions, these contributions are equal to zero. The results of a mutation are the differences in both the energy of interactions with the rest of the molecule and in the energies of the specific interactions with ligands. The bigger the energy differences, the stronger the functional impact of the mutation.

By using the term “energy” rather than the “free energy”, we assume that all interactions with water molecules are taken into account by some molecular force field, so that one can ignore the water molecules and consider all energy functions as functions of protein coordinates. Normally, 3D coordinates of all residues in a protein are fixed, therefore the entropy contribution is negligible and one can use the term “energy” describing residue-residue interactions within a protein. Without loss of generality, we can also assume that a mutated residue has the fixed 3D coordinates, and therefore the effect of the mutation can be described as the energy difference, rather than a free energy difference. If the fixation of a mutation is energetically impossible, the globule will be destabilized and such a mutation is not likely to be fixed and observed.

The direct way to compute protein energies is to use available 3D structures of protein homologs and molecular potentials. For obvious reasons – lack of data and efficient computational modeling approaches – such an approach cannot be performed routinely and massively for many thousands of mutations. Therefore, we chose an indirect approach – we approximate the exact energy differences by the average energy differences derived from distributions of residues extracted from a protein family alignment. Multiple alignments give experimental information on distributions of residues in different sequence positions. One can treat multiple alignments as a record of multiple mutation experiments conducted by nature.

To derive the average energies from residue distributions, we assume that the observed residue distributions are equilibrium distributions and the temperature – a parameter in the distribution – is the same for all residue distributions of all protein families. In other words, we assume that all possible mutations were tried in evolution in each sequence position so that the observed distributions of residues in aligned positions of homologous sequences reflect the functional constraints on these residues. Thus, evolutionarily unfavorable (for whatever reason) residues are not observed or observed less frequently than neutral or critically important residues, while critically important
residues are conserved across many different evolutionary instances (paralogs or orthologs).

The above made assumptions are generally correct for protein sequences, which are under “negative” selection pressure. This means that protein function is optimized to stable external conditions and the role of natural selection is to eliminate unfavorable mutations that disturb the protein function. In the case of “positive” selection pressure, protein function is not optimized to new external conditions and some of the mutations that were less favorable under previous conditions become favorable under new conditions. Thus, positive selection can shift the distributions of residues from thermodynamic equilibrium so that the energy differences derived under the equilibrium assumption can be wrong. Therefore, only the validation tests can prove how reasonable the above assumptions are.

1. **Free energy and the residue coordinate.**

Following the above approach, we are determining the difference of the average energy of interactions caused by a mutation $\alpha \rightarrow \beta$ in a sequence position $i$.

To this end, we present the total energy of a residue $\alpha$ in a sequence position $i$ in the following form:

$$ u_i(\alpha, \{\beta\}_i) = \sum_{j=1, j \neq i}^{N} \varepsilon_{ij}(\alpha, \beta_j) + \varphi_i(\alpha | \{\beta\}_i) \quad (1) $$

Here the term $\varepsilon_{ij}(\beta_i, \beta_j)$ is the interaction energy between residues $\beta_i$ and $\beta_j$; $\beta_i$ is a generalized coordinate that defines a type of a residue in a sequence position $i$, and, in general, all residue-related coordinates that are necessary to define 3D coordinates of all atoms of a residue; (because, the available distributions of residues are the distributions of the residue types, 3D components of the coordinate can be ignored); the notation $\{\beta\} = (\beta_1, ..., \beta_N)$ is used to represent a protein sequence of $N$ residues and the notation $\{\beta\}_i = (\beta_{i+1}, ..., \beta_N)$ represents a protein sequence without a residue $i$.

The energy term $\varphi_i(\alpha | \{\beta\}_i)$ in Eq.1 has a sense of the “external field” acting on a residue $\alpha_i$. It depends both on the type of the residue $\alpha_i$ and on the specific ligand interacting with the protein. The specificity of the ligand is taken into account by the sequence coordinate $\{\beta\} = (\alpha_i, \{\beta\}_i)$. For simplicity and with no loss of generality, we can assume that all interactions of a given residue with the native ligands can be summarized by the effective field $\varphi_i(\alpha | \{\beta\}_i)$. The protein evolves under restrictions imposed by interactions with its native ligands. These restrictions are taken into account by this field. For those residues that do not interact with the native ligands $\varphi_i(\alpha | \{\beta\}_i) = 0$.

Now the energy of a protein molecule interacting with a ligand can be presented in the following form

$$ U(\{\beta\}) = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \varepsilon_{ij}(\beta_i, \beta_j) + \sum_{j=1}^{N} \varphi_j(\beta | \{\beta\}_i) = u(\alpha_i, \{\beta\}_i) + U_i(\{\beta\}_i) \quad (2) $$
where the term

\[ U_i^j(\{\beta\}, \{\beta\}_j) = \sum_{j=1}^{N_i-1} E_{i,j}(\beta_{j}, \beta_{j}) + \sum_{j=1}^{N_i} \varphi_j(\beta_j | \{\beta\}_j) \]  

includes all energy terms that do not directly depend on the coordinate of a residue \(i\); the indirect dependency on a residue \(i\) is included in the coordinate \(\{\beta\}_j\) in the potential \(\varphi_j(\beta_j | \{\beta\}_j)\) when \(j \neq i\).

A protein sequence is the subject of random fluctuations. We assume that the observed distribution of residues corresponds to the minimum of a free energy that is defined as follows:

\[ F = \sum_{\{\beta\}_i} U(\{\beta\}) W(\{\beta\}) - T \sum_{\{\beta\}_i} W(\{\beta\}_i) \ln W(\{\beta\}) \]  

Here \(W(\{\beta\}) = W(\beta_1, ..., \beta_N)\) is a probability function \(\left( \sum_{\{\beta\}_i} W(\{\beta\}_i) = 1 \right)\); it gives a probability to observe a particular sequence \((\beta_1, ..., \beta_N)\) among all members of a protein family.

Using the energy terms of Eqs.1-2, the free energy of Eq.4 is presented as follows:

\[ F = \sum_{\{\beta\}_i} U(\{\beta\}) W(\{\beta\}) - T \sum_{\{\beta\}_i} W(\{\beta\}_i) \ln W(\{\beta\}) \]  

The free energy of Eqs.4-5 is normalized to one sequence, although the terms of the average energy and the entropy are applied to a protein family - an ensemble of fluctuating sequences - rather than for an individual sequence.

2. **Mean field approximation**

Our goal is to split the free energy of Eq.5 into a sum of two terms. The first term will take into account all dependencies on a residue distribution in a position \(i\); the second term will include all other dependencies. To make this possible, one has to assume that the distribution of residue types in a position \(i\) can be considered independently from the distribution of all other residues, i.e.

\[ W(\{\beta\}_i) \approx W_i(\{\alpha\}) W(\{\beta\}_i), \]  

where the probability functions \(W_i(\{\alpha\})\) and \(W(\{\beta\}_i)\) are defined as follows:

\[ W(\{\alpha\}) = \sum_{\{\beta\}_i} W(\{\beta\}_i) \quad \text{and} \quad W(\{\beta\}_i) = \sum_{\{\alpha\}} W(\{\beta\}_i) \]  

Then, the first term of Eq.5 can be presented as follows:
\[
\sum_\alpha \sum_{\beta} u_i(\alpha_i, \{\beta\}_i) W(\{\beta\}_i) \approx \sum_\alpha W_i(\alpha_i) \langle u_i(\alpha_i) \rangle 
\]  \hfill (8)

where

\[
\langle u_i(\alpha_i) \rangle = \sum_{\beta_i} u_i(\alpha_i, \{\beta_i\}_i) W(\{\beta_i\}_i) \approx \\
\sum_{j=\ldots} (\sum_{\beta_j} \epsilon_{i,j}(\alpha_i, \beta_j) W_j(\beta_j)) \approx \sum_{\beta_j} \phi(\alpha_i | \{\beta_j\}_i) W(\{\beta_j\}_i) = \langle \epsilon(\alpha_i) \rangle + \langle \phi(\alpha_i) \rangle
\]  \hfill (9)

is the averaged energy of a residue type \(\alpha_i\) in a position \(i\) of a given protein family. The averaging in Eq.9 is done over the individual residue distributions \(W_j(\beta_j)\) of all residues interacting with a residue \(\alpha_i\) and over all ligands interacting with protein sequences that have a residue \(\alpha_i\) in a position \(i\).

Under the conditions of independency of Eq.6, the second term of Eq.5 gives

\[
\sum_{\beta_j} U_j(\{\beta_j\}_i) \sum_\alpha W_j(\beta_j) \approx \sum_{j=\ldots} \sum_{\beta_j} \sum_{\epsilon_{j,k}} \epsilon_{j,k}(\beta_j, \beta_k) W_j(\beta_j) W_k(\beta_k)
\]  \hfill (10)

where

\[
\langle \epsilon_{j,k} \rangle = \sum_{\beta_j} \sum_{\beta_k} \epsilon_{j,k}(\beta_j, \beta_k) W_j(\beta_j) W_k(\beta_k) \approx \sum_{\beta_j} \sum_{\beta_k} \epsilon_{j,k}(\beta_j, \beta_k) W_j(\beta_j) W_k(\beta_k)
\]  \hfill (11)

is averaged energy of pairwise interactions between residues occupying positions \(j\) and \(k\) in the protein globule; and,

\[
\langle \phi(\alpha_i) \rangle = \sum_{j=\ldots} \langle \phi_j(\beta_j | \alpha_i) \rangle
\]  \hfill (12)

\[
\langle \phi_j(\beta_j | \alpha_i) \rangle = \sum_{\beta_j} \phi_j(\beta_j | \{\beta_j\}_i) W_j(\beta_j) W(\alpha_i) \approx \sum_{\beta_j} \phi_j(\beta_j | \{\beta_j\}_i) W(\beta_j)
\]  \hfill (13)

is the averaged energy of interactions between a residue \(\beta_j\) in a position \(j\) and all ligands, when a position \(i\) is occupied by a residue \(\alpha_i\); this term takes into account the cooperativity and specificity of interactions between a ligand and a protein molecule; if a residue \(\alpha_i\) is critical for the specific interactions between a protein and a ligand, then the whole ligand-to-protein energy will be taken into account in determining distributions of residues in a position \(i\).

Under the same conditions (Eq.6), the entropy term of Eq.5 can be approximately presented as a sum of two terms:

\[
S = -\sum_{\beta_i} W(\{\beta_i\}) \ln W(\{\beta_i\}) \approx -\sum_\alpha W_i(\alpha_i) \ln W_i(\alpha_i) - \sum_{\beta_i} W(\{\beta_i\}) \ln W(\{\beta_i\})
\]  \hfill (14)

Now, using Eqs.8, 10 and 14 and combining the average energies \(\langle u_i(\alpha_i) \rangle\) and \(\langle \Delta \phi_i(\alpha_i) \rangle\) in one term \(\langle v_i(\alpha_i) \rangle\):

\[
\langle v_i(\alpha_i) \rangle = \langle u_i(\alpha_i) \rangle + \langle \Delta \phi_i(\alpha_i) \rangle
\]  \hfill (15)

the free energy of Eq.5 can be presented as a sum of two terms:

\[
F \approx F[W_i(\alpha_i)] + F[W(\{\beta_i\})]
\]  \hfill (16)
\[ F[W(\alpha_i)] = \sum_{\alpha} \langle v_i(\alpha) \rangle (W_i(\alpha) - T \sum_{\alpha} W_i(\alpha) \ln W_i(\alpha)) \] (17)

is a free energy of a residue type \( \alpha_i \) in the averaged molecular field of all other residues of the protein and the ligand, and

\[ F[W(\{\beta_i\})] = \sum_{i=1,...,N} \sum_{j \neq i} \langle \epsilon_{j,k} \rangle - T \sum_{i=1,...,N} W_i(\beta_i) \ln W_i(\beta_i) \] (18)

is a free energy of all residues of the protein without a residue \( i \).

The approximations of Eqs.6, 11,13 are known in statistical mechanics as the mean field approximation [1]. In this approximation, the energy of a given particle (residue) is determined as an average “field” of all other particles of a system (all other residues) interacting with a given residue. The essence of the approximation is in ignoring the influence of a given particle on the distributions of other particles of the system.

3. The average energy difference

At the thermodynamic equilibrium the variation of the free energy is equal to zero:

\[ \frac{\delta F}{\delta W_i(\alpha_i)} \approx \frac{\delta}{\delta W_i(\alpha_i)} (F[W_i(\alpha_i)] + F[W(\{\beta_i\})]) = \frac{\delta}{\delta W_i(\alpha_i)} (F[W_i(\alpha_i)] = 0 \] (19)

Using the free energy term of Eq.17, one obtains the general thermodynamic relation [1] between the average energy of a given residue in a given position and a probability to find this residue in this position.

\[ \langle v_i(\alpha) \rangle = F - T \ln W_i(\alpha) \] (20)

The difference of the molecular interaction energy caused by a mutation \( \alpha \) to \( \beta \) in a position \( i \) can be estimated now as follows:

\[ \Delta v_i(\alpha \to \beta) = \langle v_i(\beta) \rangle - \langle v_i(\alpha) \rangle = -T \ln \frac{W_i(\beta)}{W_i(\alpha)} \approx -T \ln \frac{n_i(\beta)}{n_i(\alpha)}, \] (21)

where \( n_i(\alpha) \) and \( n_i(\beta) \) are the observed occurrences of residues \( \alpha \) and \( \beta \) in a position \( i \).

The Eq.21 is not defined when \( n(\beta) = 0 \); actually, the energy difference determined by Eq.21 is equal to infinity. This means that a mutation to a residue that has never been seen in a given position is forbidden. This controversial result can be corrected by taking into account explicitly the discrete nature of the residue distribution and the combinatorial form of the entropy. Then, the free energy of Eq.17 is modified as follows:

\[ F[n_i(\alpha_i)] = \frac{1}{L} \ln \left[ \sum_{\alpha_i} \langle u_i(\alpha_i) \rangle + \langle \Delta \phi_i(\alpha_i) \rangle n_i(\alpha_i) - T \ln \frac{L!}{\prod_{\alpha} n_{\alpha_i}!} \right], \] (22)

In transition to Eq.22, we used \( W_i(\alpha_i) = n_i(\alpha_i)/L \) and the combinatorial definition of the entropy:
\[ S_i = \ln \prod_{\alpha} \frac{L!}{n_i(\alpha)!} \approx -L \sum_{\alpha} \left( n_i(\alpha) / L \right) \ln \left( n_i(\alpha) / L \right) = -L \sum_{\alpha} W_i(\alpha) \ln W_i(\alpha) \]  \hspace{1cm} (23)

where the factorial function \( L! \) is defined as follows: \( L! = 1 \times 2 \times \ldots \times L \), and, by definition, \( 0! = 1 \); \( L \) is the total number of residues in an alignment column; \( n_i(\alpha) \) is a number of residues of type \( \alpha \) in a column \( i \) \( (\sum_{\alpha} n_i(\alpha) = L; \ \alpha = 1,2,\ldots,21 \) indexing 20 residues types and gaps).

At thermodynamic equilibrium, the variation of the free energy as a result of a mutation of \( \alpha \rightarrow \beta \) in a position \( s \) is close to zero.

\[
\delta F = \frac{1}{L} \left[ \langle v(\beta) \rangle \langle n(\beta) \rangle + 1 + \langle v(\alpha) \rangle \langle n(\alpha) \rangle - 1 - \langle v(\beta) \rangle - \langle v(\alpha) \rangle \langle n(\beta) \rangle - \langle n(\alpha) \rangle \langle n(\beta) \rangle + 1 \right] =

- T \left( \ln \frac{n(\gamma) \ldots n(\alpha) - 1 \ldots n(\beta) + 1 \ldots n(\gamma)}{n(\gamma) \ldots n(\alpha) - 1 \ldots n(\beta) + 1 \ldots n(\gamma)} \right)

= \frac{1}{L} \left[ \langle v(\beta) \rangle - \langle v(\alpha) \rangle - T \ln \frac{n(\alpha)}{n(\beta) + 1} \right] \approx 0
\]

Finally, using the definitions of Eqs.9 and 15, this gives

\[
\Delta v_i(\alpha \rightarrow \beta) = \langle v_i(\beta) \rangle - \langle v_i(\alpha) \rangle =

\langle \Delta v(\beta) \rangle + \langle \Delta \phi(\beta) \rangle - \langle \Delta v(\alpha) \rangle - \langle \Delta \phi(\alpha) \rangle = -T \ln \frac{n_i(\beta) + 1}{n_i(\alpha)} \]  \hspace{1cm} (24)

Now, the energy cost of the mutation \( \beta \) that was never observed in a given position \( i \) is estimated as \( \Delta v_i(\alpha \rightarrow \beta) = T \ln n_i(\alpha) \).

The value of the average energy difference of Eq.25 is big when \( n_i(\beta) \ll n_i(\alpha) \) that is when a mutated residue is conserved across many sequences of a protein family. Therefore we call the mutation impact term of Eq.25 “the conservation score” (superscript “c”) to underline the family conservation.

4. Concluding remarks

1. The protein stability impact term \( \langle \Delta v(\beta) \rangle - \langle \Delta v(\alpha) \rangle \) and the ligand interaction impact terms \( \langle \Delta \phi(\beta) \rangle - \langle \Delta \phi(\alpha) \rangle \) and \( \langle \Delta \phi(\beta) \rangle - \langle \Delta \phi(\alpha) \rangle \) are combined together in Eq.25. These impact factors are indistinguishable within the approach based on the residue distribution functions.

2. The impact term of Eq.25 is the difference between the average energies, rather than the difference between the individual energies that can differ from sequence to sequence. This means that the mutation impact is the same for all sequences of a protein family that have the residue \( \alpha \) in a position \( i \) mutated to the same residue \( \beta \).

3. The mutation impact term of Eq. 25 depends both on the types of the original and the mutated residues, and on the position of the mutated residue in the protein family.
alignment. The position of a residue in a multiple alignment of cognate sequences is well correlated with the position of a residue in a protein globule. This means that a residue occupying a position on a protein surface is aligned with residues occupying similar surface positions and a residue occupying a buried position inside a protein globule is aligned with the residues that occupy similar buried positions. Thus, the functional form of Eq.25 takes into account the major factor – the difference in physico-chemical properties of amino-acid residues in the context of the 3D structure.

References

1. R. Kubo, H. Ichimura. Statistical mechanics. An advance Course with Problems and Solutions. Elsevier First edition 1965; fifth impression 2004 (ISBN: 0-444-87103-9)