Residue Communities Reveal Evolutionary Signatures of Protein Function

N. J. Cheung
Department of Biochemistry, University of Oxford

Abstract

Naturally co-occurring amino acids, term coevolution, in a protein family play a significant role in both protein engineering and folding, and it is expanding in recent years from the studies of the effects of single-site mutations to the complete re-design of a protein and its folding, especially in structure prediction. Here we report an approach that can identify the evolutionary signatures (highly ordered networks of coupled amino acids, termed "residue communities", RCs) from protein homologous sequences. The method provides access to apply the spectrum analysis on evolutionary coupling analysis (SAEC) for inferring RCs in proteins and their important roles in protein function.

The SAEC Method

The evolutionary process of a protein family can be mathematically modeled by a sequence generator at equilibrium [1]. The generator produces a sequence τ with a probability $P(\tau)$ (Eq. 1) from a distribution over the space of all possible sequences in the family.

$$P(\tau) = \frac{1}{Z}\exp(E(\tau)) \tag{1}$$

where Z is the partition function that normalizes the distribution by summing over the Boltzmann factors of all possible sequences in the family. The total energy ($E(\tau)$) of the sequence τ is

$$E(\tau) = \sum_i \mathbf{h}_i(\tau_i) + \sum_{i<j} \mathbf{e}_{ij}(\tau_i, \tau_j) \tag{2}$$

where $\mathbf{h}_i$ and $\mathbf{e}_{ij}$ are, respectively, site-specific bias terms of an amino acid and coupling terms between pairwise amino acids.

Given the multiple sequence alignment (MSA) of the family, the model parameters $\mathbf{h}_i$ and $\mathbf{e}_{ij}$ can be optimized by maximum likelihood. Here, we leverage a site-factored pseudolikelihood approximation, instead of the full likelihood, to efficiently compute the partition function Z (Eq. 1), and the l_2-regularization is to penalize the model to avoid overfitting the sequences of the family. The objective function is defined

$$\begin{aligned} O\left\{\widehat{\mathbf{h}}, \widehat{\mathbf{e}}\right\} = \arg\max_{\mathbf{h},\mathbf{e}} \sum_{s,i} \omega_s \log P_i^s(\Omega) \\ - \lambda_{\mathbf{h}} \sum_i \mathbf{h}_i^2(\tau_i) - \frac{\lambda_{\mathbf{e}}}{2} \sum_{i,j} \mathbf{e}_{ij}^2(\tau_i, \tau_j) \end{aligned} \tag{3}$$

where Ω denotes all the sequences in the MSA, ω_s is weight of each sequence, the l_2-regularization factors $\lambda_{\mathbf{h}}$ and $\lambda_{\mathbf{e}}$ are, respectively, set to 0.01 and $0.01 \cdot q \cdot (L-1)$, q is the total number of possible states. The conditional likehood $P_i^s(\Omega)$ of sequence s at position i is defined

$$P_i^s(\Omega) = \frac{\exp\left(\mathbf{h}_i(\Omega_i^s) + \sum_{j\neq i} \sigma_j^s \mathbf{e}_{ij}(\Omega_i^s, \Omega_j^s)\right)}{\sum_a \exp\left(\mathbf{h}_i(a) + \sum_{j\neq i} \sigma_j^s \mathbf{e}_{ij}(a, \Omega_j^s)\right)} \tag{4}$$

where $\sigma_i^s = 0$ when the ith site is gapped, otherwise $\sigma_i^s = 1$.

To make the residue communities as much statistically independent as possible, the top three residue communities are defined based on two of the top five eigenvalues and their corresponding eigenvectors ($\mathbf{v}_{k|k=1,\cdots,5}$) of the matrix $\mathbf{e}_{ij}$ (Eq. 2) [2] as:

(1) community I (red) consists of residues at the ith position of $\mathbf{v}_{k=2}^i > \max(\mathbf{v}_{k=4}^i, \epsilon)$;

(2) community II (blue) includes residues at the ith position of $\mathbf{v}_{k=2}^i < -\max(\mathbf{v}_{k=4}^i, \epsilon)$; and

(3) community III (green) includes residues at the ith position of $\mathbf{v}_{k=4}^i > \max(\mathbf{v}_{k=2}^i, \epsilon)$.

We use an $\epsilon = 0.05$ as a threshold to project the amino acids reduced from the coupling matrix and extract meaningful residue communities.

References

[1] T.A. Hopf, J.B. Ingraham, F.J. Poelwijk, C.P. Schärfe, M. Springer, C. Sander, and D.S. Marks, "Mutation effects predicted from sequence co-variation," *Nature Biotechnology*, vol. 35(2), pp. 128-135, 2017.

[2] N.J. Cheung, A. T. J. Peter, and B. Kornmann, "Leri: a webserver for identifying protein functional networks from evolutionary couplings," *Computational and Structural Biotechnology Journal*, pp. 1-22, 2020.

[3] B. Kornmann, E. Currie, S.R. Collins, M. Schuldiner, J. Nunnari, J.S. Weissman, and P. Walter, "An ER-mitochondria tethering complex revealed by a synthetic biology screen," *Science*, vol. 325(5939), pp. 477-481, 2009.

Residue Communities

On the right, we show the protocol that is leveraged in our approach maps the evolutionary information into residue communities: the SAEC algorithm [2] allows us to prune the evolutionary interactions to several highly ordered networks of amino acids.

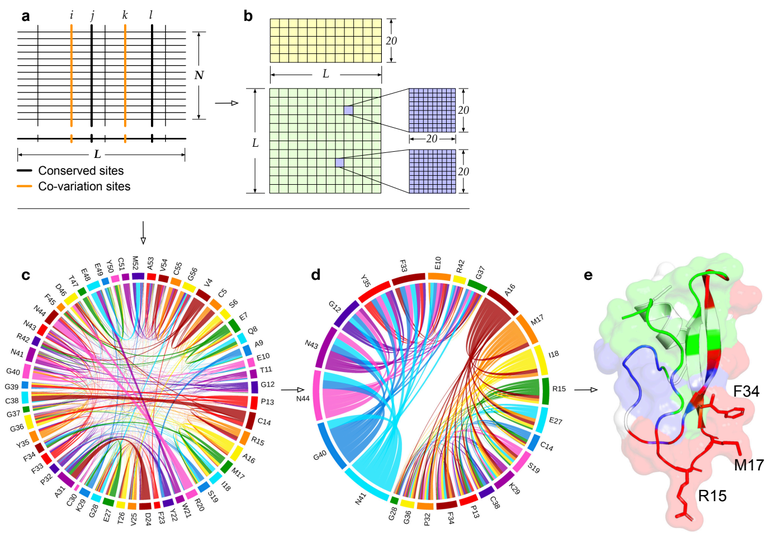
a: a multiple sequence alignment (MSA) is obtained by searching a primary sequence against a specific sequence database with appropriate parameters.

b: site-independent biases and pairwise couplings are position-specific substitution probabilities and measurements of interactions between pairwise residues.

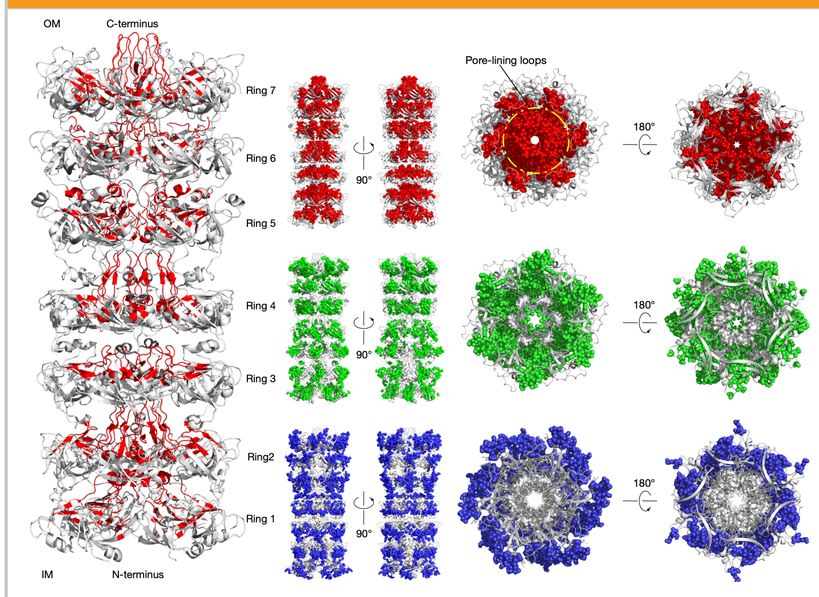
c: evolutionary couplings (ECs) are interactions with positive measurements.

d: residue communities (RCs) are pruned by the spectrum analysis on the evolutionary couplings.

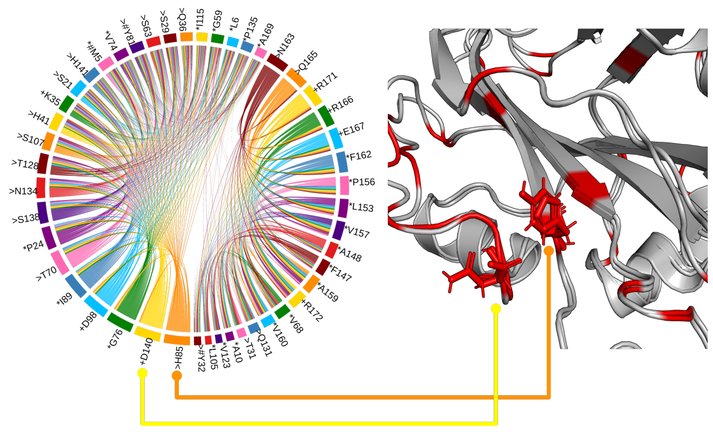
e: RCs mapped to the tertiary structure



Results



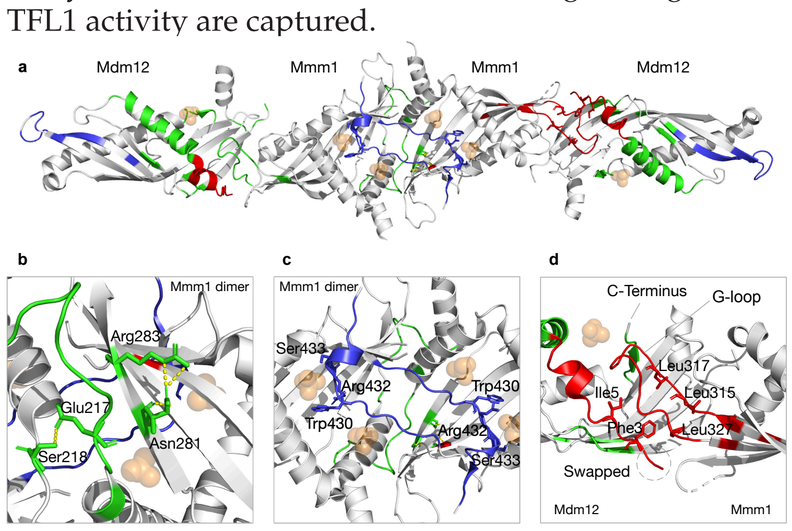
We tested our method on four different proteins and show the relationship between the residues that are relevant to protein function and the estimated evolutionary signatures (RCs).



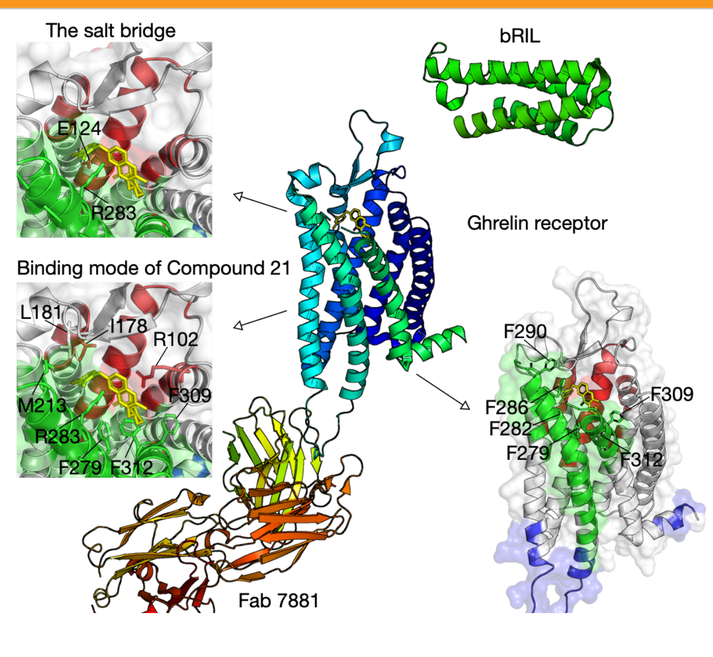
Top-left: the RCs are identified for the LetB (PDB: 6V0C), and the pore-facing residues at the RC (red) contribute to the formation of the tunnel that mediates lipid transport in different states, while the RCs (in blue and green) involve in the dynamic of the LetB between the open and closed states.

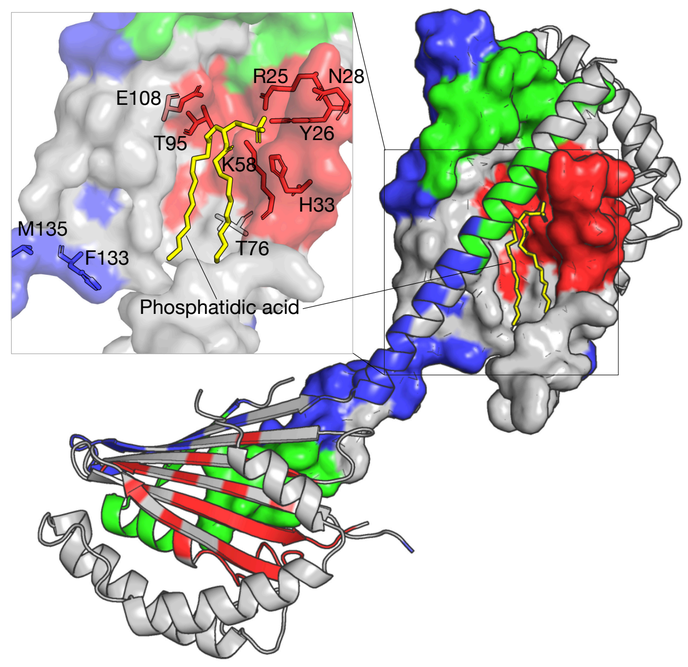
Top-right: the RCs that cover the ligand-binding pocket of the ghrelin receptor (PDB: 6KO5) are identified, e.g., bifurcated by a salt bridge between E124 and R283.

Bottom-left: the residues His88 and Asp144 that are likely the most critical residues for distinguishing FT and TFL1 activity are captured.



Summary: We present an approach (SAEC) that can identify residue communities as evolutionary signatures of protein function from couplings between amino acids using the spectrum analysis. We evaluate the inference method on the different proteins with important biological functions, and the results demonstrate that the inferred residue communities suffice to specify the evolutionary signatures of protein evolution and provide access to design functional sequences. Such residue communities may guide the engineering of functional proteins with altered (bio)chemical activities.





Bottom-right: the lipid transfer protein Ups1 shuttles phosphatidic acid between mitochondrial membranes. The inferred RCs are mapped to the tertiary structure (PDB: 4YTX), and the residues in the RC (red) are located at the tunnel-like binding cavity within the hydrophobic core and experimentally demonstrated to contribute the gain-of-function of Ups1.

The RCs are predicted for the endoplasmic reticulum-mitochondria encounter (ERMES) complex [3]. The communities of coupled residues mapping to the tertiary structure of Mmm1-Mdm12 complex (PDB ID: 5YK7). (**a**) Top three communities of coupled residues are mapped in red, green and blue. (**b**) Pattern of co-evolution, and the co-evolved residues that inferred by the SAEC method is shown in green as stick with polar interaction in yellow dot lines. (**c**) The inferred residue community highlighted in blue consists of function-related residues. The two highly conserved residues Trp430 and Arg432 are shown as stick in blue. (**d**) Critical determinants of the interaction between Mmm1 and Mdm12.

As illustrated (below), the G-loop (red) of Mmm1 that results in conformation changing in Mmm1-Mdm12 complex (PDB: 5YK7) is distinguishable from that in the Mmm1 monomer (PDB: 5YK6).