Computational and Structural Biotechnology Journal 19 (2021) 3556–3563

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Method article

Leri: A web-server for identifying protein functional networks from evolutionary couplings

Ngaam J. Cheung [a,c,*], Arun T. John Peter [b], Benoit Kornmann [a]

[a] Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK
[b] Institute of Biochemistry, ETH Zürich, Zürich 8092, Switzerland
[c] Leri Ltd, Oxford, UK

ARTICLE INFO

Article history:
Received 12 February 2021
Received in revised form 30 May 2021
Accepted 2 June 2021
Available online 6 June 2021

Keywords:
Residue community
Evolutionary coupling analysis
Functional network
Leri

ABSTRACT

Information on the co-evolution of amino acid pairs in a protein can be used for endeavors such as protein engineering, mutation design, and structure prediction. Here we report a method that captures significant determinants of proteins using estimated co-evolution information to identify networks of residues, termed "residue communities", relevant to protein function. On the benchmark dataset (67 proteins with both catalytic and allosteric residues), the Pearson's correlation between the identified residues in the communities at functional sites is 0.53, and it is higher than 0.8 by taking account of conserved residues derived from the method. On the endoplasmic reticulum-mitochondria encounter structure complex, the results indicate three distinguishable residue communities that are relevant to functional roles in the protein family, suggesting that the residue communities could be general evolutionary signatures in proteins. Based on the method, we provide a webserver for the scientific community to explore the signatures in protein families, which establishes a powerful tool to analyze residue-level profiling for the discovery of functional sites and biological pathway identification. This web-server is freely available for non-commercial users at https://kornmann.bioch.ox.ac.uk/leri/services/ecs.html, neither login nor e-mail required.

1. Introduction

Protein sequences, which specify their structures and function, are products of natural selection [1,2]. The sequences, therefore, harbor evolutionary information. This information in homologous proteins has enabled the accurate computation of residue contacts [3–5], and helped the prediction of tertiary structures [6–8]. Other useful clues about protein function might be gleaned from this evolutionary information. With advances in high-throughput sequencing technologies and the expansion of protein databases, we now have insight into evolutionary patterns, which have allowed researchers to detect and identify sets of amino acids performing related functions [9] and engineer proteins [1,10]. Statistical coupling analysis (SCA) is a successful approach for identifying sparse networks of interactions among co-variant amino acids (term 'protein sectors'), for example, the coupled conservation from the networks allows to make the artificial WW domain (or rsp5-domain) protein sequences that fold into a specified structure representative of the protein family and function as the natural proteins [1,9,11]. Moreover, sparse networks of evolutionarily conserved amino acids predict structural motifs for allosteric communication in proteins [12,13]. This approach is also successful in predicting the effect of mutations on protein function [14–16]. Although it is not difficult to create protein sequences, it remains challenging to design the sequences (being loosely related to natural sequences) that fold in predictable ways and function in a manner indistinguishable from the natural representative of the family. Recent findings have shown that co-evolutionary information hidden in homologous proteins can be harnessed to efficiently create sequences and predict their folding behavior *in vitro* [1,17]. Thus evolutionary information from homologous sequences provides new insights towards designed sequences that fold and function in a manner similar to natural proteins.

Biologists have been striving to uncover the mysteries of the relationship between protein primary sequence and their tertiary structures and their function. Despite the vast amount of protein sequence data made publicly available with inexpensive DNA synthesis and next-generation sequencing (NGS) [18–20], finding

* Corresponding author at: Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK.
E-mail address: yan.zhang@bioch.ox.ac.uk (N.J. Cheung).

https://doi.org/10.1016/j.csbj.2021.06.002

biologically meaningful information in the data requires elucidating the evolutionary determinants that give rise to protein functions. These determinants can be manifested as amino acid coevolutionary patterns [21]. Such coevolutionary patterns are informative in an array of applications from protein structure prediction constrained by residue contacts (derived from direct information [3,6]) to protein engineering and drug discovery. Homologous proteins preserve similar structural information (e.g., contacts between residues) and contain information (e.g., protein–protein interaction) for understanding biological activities, and advances have also developed to disentangle the information, but challenges remain what and how couplings drive protein to fold and function, especially, how the strongly coupled residues that are sparse and organized into highly ordered architectures link to protein's functional sites. Technological advances over the years have enabled certain estimations of residues ranked by their importance relevant to protein function, including the ability to infer protein sectors, groups of amino acids that are coevolving in homologous proteins [9], and assess mutation effects [15]. Assessing the network-level effects of evolutionary couplings between pairwise amino acids, however, remains a challenge. Even if experimental investigations (e.g., directed evolution [22]) could provide all components necessary for understanding protein function, discovering interactions among amino acids requires a computational approach to analyze large-scale sequence data and put pieces of information together [23,24]. The energy changes (the energy difference between the wild-type and mutated sequences, $\Delta E = E_{wt} - E_{mutant}$) computed from the positionally conserved couplings have a strong correlation with the transition temperatures for the extant and ancestral Thioredoxin (Trx) proteins, and those changes might allow engineering proteins with improved properties [25].

Here, we present an enhanced computational web-server, termed *Leri* (learning engine to recognize life), that allows researchers to detect co-evolution patterns (highly ordered networks of coupled amino acids, community structures in graph theory, termed residue communities). It can also identify functionally important residues that would be used for protein design and folding from either experimental or natural evolution [26,8]. This provides a systematic method by applying the spectrum analysis on evolutionary coupling analysis (termed SAEC) for inferring functional networks in proteins, identifying mutations as a guide for protein engineering, and facilitating researchers to understand function. Through two instructive examples, we showcase the capabilities of *Leri* in building the residue communities that are relevant to protein's functions. In the first case, we apply *Leri* to identify the functionally relevant residues in the mitochondrial morphology maintenance 1 (Mmm1) protein. In the second example, we demonstrate how *Leri*'s efficiency in capturing the residues involved in lipid transportation in the endoplasmic reticulum-mitochondria encounter structure (ERMES) complex.

2. Materials and methods

2.1. Framework and implementation

The *leri* web-server takes a single sequence or a multiple sequence alignment (MSA) as input to infer evolutionary couplings and residue communities that could be determinants of functional specificity (Figure S1). Given an MSA (Fig. 1a, whatever provided by the user or search the query sequence against the databases), quality control, including sequence trimming and re-weighting, is required to improve the estimations of evolutionary information and reduce the background noises. Inferences (site-independent biases and couplings, Fig. 1b) are computed from the controlled MSA, and the co-evolution between pairwise residues is measured by the strength of couplings (Fig. 1c). The residues are identified as determinants of functional specificity (Fig. 1d), and those significant coupled residues are mapped to the tertiary structure of the query sequence if provided (Fig. 1e).

2.2. Generation of multiple sequence alignments

Co-variation patterns in a natural protein family highly depend on the quality of its MSA when the statistical transformation is made from the genetic record. For each analyzed protein, its multiple sequence alignment was obtained from two databases by the profiles HMM (hmmsearch version 3.3) [27] and HHblits (version 3.3.0) [28] homology search tools. Firstly, an alignment of a query protein is obtained by searching its sequence against the UniRef30 database (as of 2/2020) using HHblits [28] with the default parameters at an E-value threshold 0.001. The other alignment of the same protein is generated by the default five search iterations of *jackhmmer* in the HMM suite [27], searching the query sequence against the UniRef90 database (release 3/2019) [29]. Finally, the two alignments were concatenated and aligned according to the query sequence of each protein. Thereafter, it was trimmed based on the minimum coverage, which satisfies two basic rules [30,25]: (1) a single site with more than 90% gaps across the MSA will be removed; and (2) a sequence with the percentage of gaps more than a given threshold (80%) will be deleted from the MSA. In the alignment of each protein, the weight, $\omega(\tau)$, of each sequence was computed by the sequence identity I, which is measured by the normalized Hamming distance $D_H(\tau, \tau_j)$ between the sequence τ and all other sequences. Accordingly, the weight is defined [15] as Eq. (1).

$$\omega(\tau) = \left(\sum_j I\left[D_H(\tau, \tau_j) < \theta \right] \right)^{-1} \tag{1}$$

where θ is a threshold that controls the maximum diversity of pairwise sequences, and $\theta = 0.2$ is default, that is, the threshold of sequence identity is 80% between the two sequences.

2.3. Inference of residue communities

Here, we applied a global probabilistic model [15] with the pseudo-likelihood maximization approach to capture evolutionary information from the multiple sequence alignment (Fig. 1). A protein sequence τ is derived by a probability $P(\tau)$ from a distribution over the space of all possible sequences in its family. The probability $P(\tau)$ is defined as

$$P(\tau) = \frac{1}{Z} \exp(E(\tau)), \tag{2}$$

where $Z = \sum_i \exp\left\{ \sum_{(i<j)} \mathbf{e}_{ij}(\tau_i, \tau_j) + \sum_i \mathbf{h}_i(\tau_i) \right\}$ is the partition function that normalizes the distribution by summing over the Boltzmann factors of all possible sequences in the protein family. In the model, the measurement of the Boltzmann factors for each sequence τ is defined by the evolutionary statistical energy (a Markov Random field or a Potts model in statistical physics) as Eq. (3).

$$E(\tau) = \sum_{i<j} \mathbf{e}_{ij}(\tau_i, \tau_j) + \sum_i \mathbf{h}_i(\tau_i), \tag{3}$$

where $\mathbf{h}_i$ and $\mathbf{e}_{ij}$ are, respectively, site-specific bias terms and coupling terms between pairwise amino acids. Our method focuses on discovering evolutionarily correlated residues that underlie distinct residue communities (patterns of co-evolved amino acids, Fig. 1d) within a protein or complex. As the coupling matrix $\mathbf{e}_{ij}$ (Fig. 1b) is to measure the strength between pairwise amino acids

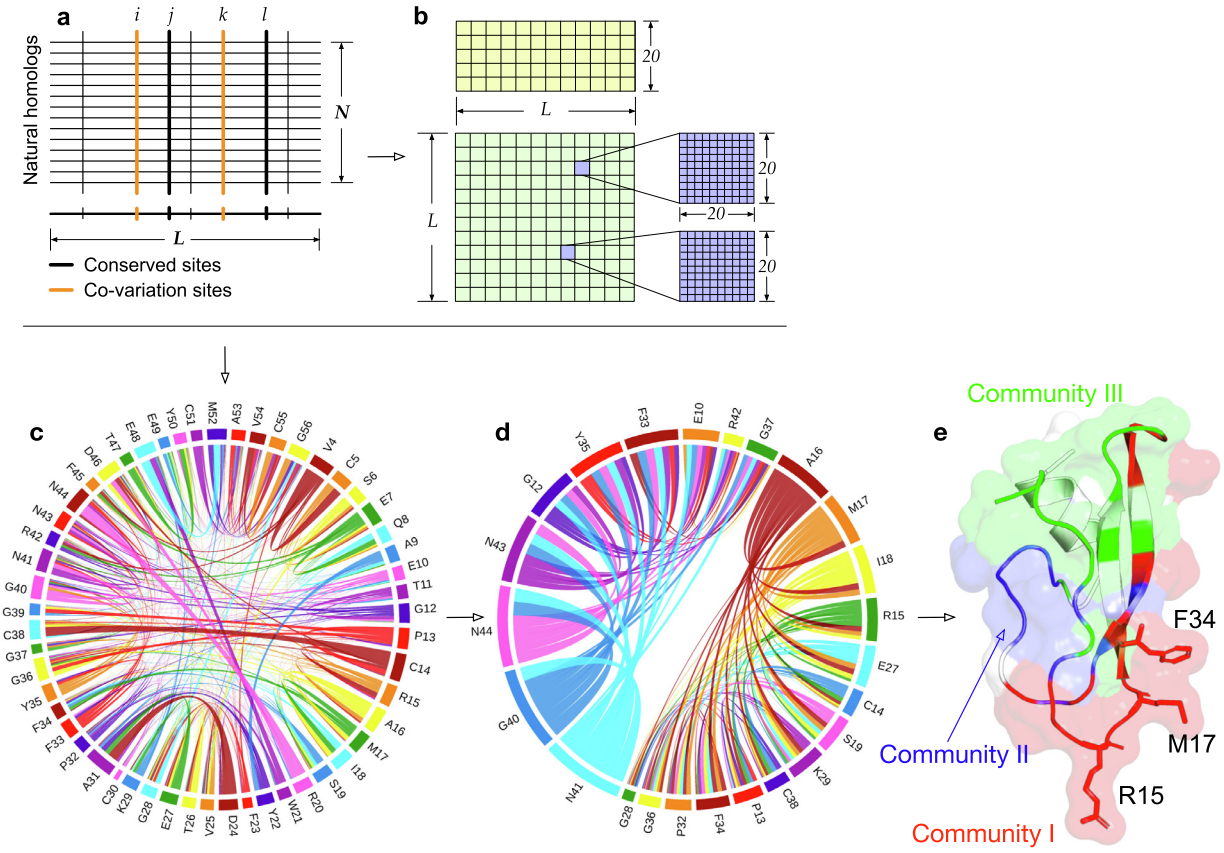

Fig. 1. Identification of function-relevant residue communities from a sequence alignment. (a) Conserved and co-varied residues in an MSA. (b) Energy-like potentials including site preferences and pairwise couplings are inferred by the SAEC method. (c) Interactions between pairwise residues. (d) The top two residue communities are computed from the pairwise couplings by spectral analysis. (e) The top three communities including a network with inter-interactions between the top two are mapped to the tertiary structure (PDB1AAP).

at positions i and j in the protein sequence alignment, it is straightforward to identify significantly coupled pairs by spectral decomposition [31], which diagonalizes the coupling matrix by linearly combining the positions of the amino acids into eigenmodes. The spectral decomposition provides a way to extract nonrandom correlated modes of the couplings, effectively reduces noises in the couplings, and partially sorts out the residues into different "residue communities". A procedure of Jacobi's iteration is carried out to determine the positive eigenvalues and the corresponding eigenvectors. Accordingly, the eigenvalues of the coupling matrix are to indicate the information in the inferred residue communities, while the corresponding eigenvectors give the weights for the positions of amino acids. In order 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. 3).

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. 2), and the l_2-regularization is to penalize the model to avoid overfitting the sequences of the family. The objective function is defined

$$O\left\{\hat{\mathbf{h}},\hat{\mathbf{e}}\right\} = \arg\max_{\mathbf{h},\mathbf{e}} \sum_{s,i} \omega_s \log P_i^s(\Omega) - \lambda_{\mathbf{h}} \sum_i \mathbf{h}_i^2(\tau_i) - \frac{\lambda_{\mathbf{e}}}{2} \sum_{i,j} \mathbf{e}_{ij}^2(\tau_i,\tau_j) \tag{4}$$

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

$$P_i^s(\Omega) = \frac{\exp\left(\mathbf{h}_i(\Omega_i^s) + \sum_{j\neq i} \sigma_j^s \mathbf{e}_{ij}\left(\Omega_i^s,\Omega_j^s\right)\right)}{\sum_a \exp\left(\mathbf{h}_i(a) + \sum_{j\neq i} \sigma_j^s \mathbf{e}_{ij}\left(a,\Omega_j^s\right)\right)} \tag{5}$$

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

We used an $\epsilon = 0.05$ as the threshold to project the amino acids reduced from the coupling matrix and extract meaningful residue communities. 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. 3) 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.

3. Results

3.1. Implementations

Leri provides visualizations of the results in HTML format. Plotly [32] is employed to visualize the data interactive web graphics, and NGL Viewer [33] is used to process and view protein structure files.

The web implementation of *Leri* allows users to choose an engine for a specific computation on a given protein sequence, as illustrated in Fig. 2. The *Leri* web-server provides four types of computations that are all driven by a query sequence, including (1) evolutionary coupling analysis and functional networks to capture evolutionary signatures (e.g. residue communities) and residues evolved for function; (2) protein single mutation for evaluating mutants for potential applications in protein engineering; (3) protein coupled mutations result in multiple mutants that have evolutionary dependence on each other; and (4) protein sequence design from the inferred residue communities can guide the engineering of functional proteins with altered (bio) chemical activities. Broadly, to create analytical results of a protein of interest, the user needs to provide its sequence or multiple sequence alignment (MSA) in FASTA format, together with an optional PDB file, on which the identified "residue communities" are mapped, and all the residue communities are computed from the MSA of each protein by the SAEC method (see Methods).

On completion of calculations, a report in HTML format is presented with a summary of detailed results by interactive graphics and visualization of the inferred residue communities in the NGL Viewer [33] (if protein structure is provided) (Fig. 3). The output data (plain text files) can be downloaded for customized use. Downloads are also available for the 3D representation of the structure with mapped residue communities. A summary is given for the basic information of the job (Fig. 3a). As detailed, the MSA can be visualized for either an uploaded sequence alignment or an alignment generated against the database, while the distribution of sequence identities between pairwise sequences are interactively shown in the histogram chart (Fig. 3a and b). To measure the information on residue conservation, the relative

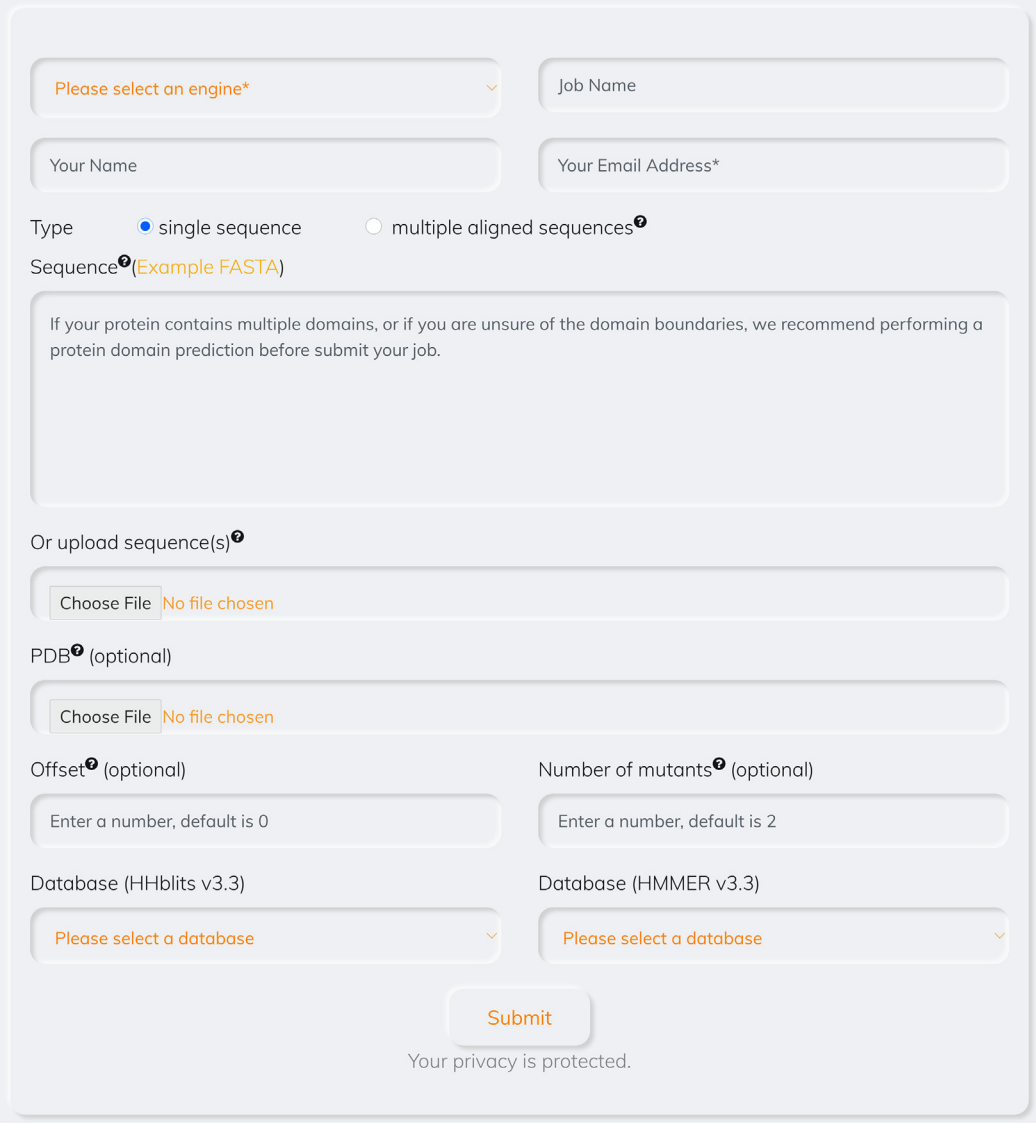

Fig. 2. Protein sequence submission page for the *Leri* web server. Users are prompted to select a computing engine and upload/paste a sequence in FASTA format to submit for calculations.

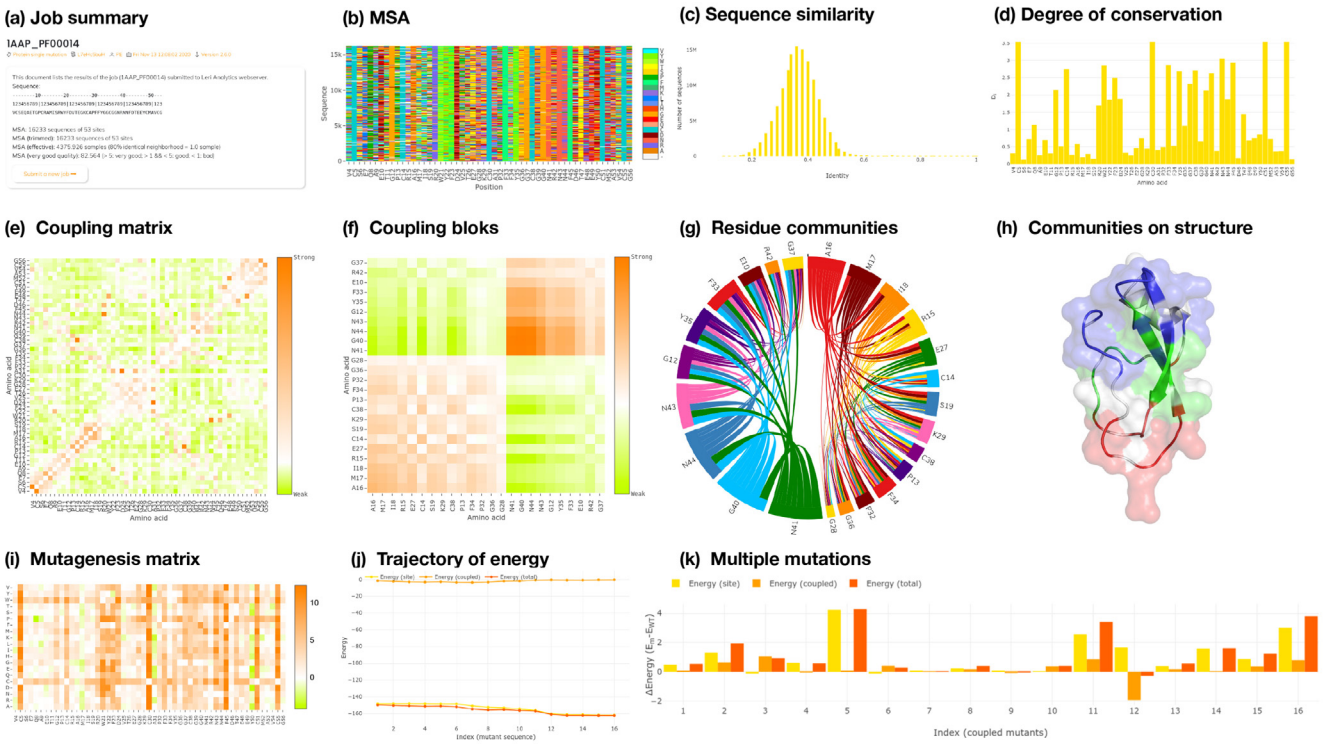

Fig. 3. Interactive results shown in HTML format for a *Leri* web-server job. (a) Basic summary of the job submitted by a user. (b) The MSA is collected from Pfam database [34] (PF00014). (c) Sequence similarity. (d) Degree of conservation at the position of each amino acid. (e) evolutionary coupling matrix inferred from the MSA by the ECA method. (f) and (g) show the top two residue communities that are computed from the coupling matrix. Interactions with both positive and negative values are illustrated in (f), while positive interactions are shown in an interactive chord graphic in (g). (h) Mapping the inferred residue communities to the tertiary structure (PDB1AAP) in red, blue, and green in interactive web interface based on NGL Viewer [33], PyMOL [35] script can be downloaded for local generation. (i) Substitution matrix for 20 common amino acids on the WT protein sequence. (j) Energy trajectory of the mutant sequence starting from the WT sequence. (k) Energy differences of coupled mutants between the WT and mutant sequences.

entropy (Kullback–Leibler divergence[36]) is computed for the most prevalent residue at the specific position (Fig. 3d). The entropy evaluates how different the observed amino acid A at the *i*th position would be if A randomly occurred with an expected probability distribution. The user can measure interactions among amino acids based on the evolutionary coupling strength (Fig. 3e) for a particular protein and infer its interaction/binding sites that are estimated in the residue communities (Fig. 3f-h). The amino acids located in those communities are functionally important and are more sensitive to substitutions, whereas the residues outside of the communities are tolerant to mutation (Fig. 3i). The differences between the wild-type and optimized mutant sequences (with the lowest sequence energy in a fixed number of iterations) are computed from any pairwise mutants (Fig. 3j), and the predicted energy differences are relevant to the thermal stability of proteins [25]. On the webserver, the number of coupled mutants

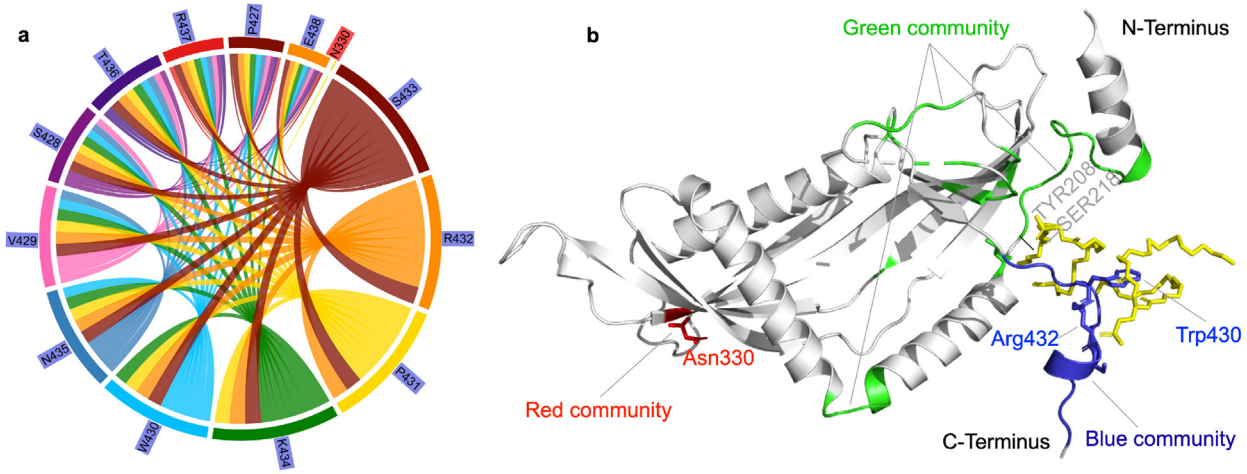

Fig. 4. Function-related amino acids of Mmm1 protein (PDB ID:5YK6) are identified by the residue communities. (a) Coupling between pairwise amino acids in the top two residue communities (the residue N330 is the only one in the red community). (b) The top three residue communities are mapped to the tertiary structure. Residues in three communities are shown in red, green, and blue, respectively. The phospholipid-bound to Mmm1 are shown in yellow stick representation.

can be customized, and the energy can guide the design of a protein of interest.

To draw interpretable conclusions from the results, we provide interactive visualizations and quantitative measurements (e.g., energies) of the data on the *Leri* web pages. In Fig. 3g, for example, we present an interactive circular plot of the top two residue communities that consist of residues in highly ordered patterns (Fig. 3f) with coupling strength between them. The communities can be also mapped on the tertiary structure (Fig. 3h) if provided, and structurally visualized for the dominant signatures via the inferences of the SAEC method.

3.2. Phospholipid-binding site inferred by coupling networks

The endoplasmic reticulum-mitochondria encounter structure (ERMES) complex, consisting of at least four proteins [37]: Mdm10, which is integral outer mitochondrial membrane (OMM) protein, Mmm1, which is integral to the ER membrane, and Mdm12 and Mdm34, which are cytosolic proteins, have been functionally connected to phospholipid biosynthesis [37]. Mmm1 binds phospholipid molecules [38] (Fig. 4b) through a conserved domain called SMP (Synaptotagmin-like, Mitochondrial, and lipid-binding Proteins) that also presents in Mdm12 and Mdm34. This first case illustrates how the inferred communities of residues can be used to identify functional sites in Mmm1 protein. The couplings between pairwise amino acids were inferred from the alignment of naturally occurring sequences. The highly coupled residues fall within three networks (Fig. S1), and the top two networks consist of residues that have distinct intra-interactions (Fig. 4a) and are highlighted in red and blue in Fig. 4b, while the other network (green in Fig. 4b) includes residues that have inter-interactions with those in other two networks.

The residues in the well-ordered loop (residues Y208-S218) are important to mediate the self-association of the Mmm1 dimer and phospholipid-binding at the highly conserved C-terminus [39]. As illustrated in Fig. 4b, those residues are captured from the MSA of the Mmm1 using the proposed SAEC method. That is, the inferences are consistent with experimental results. Interestingly, we find that the residues positioned at the blue communities engage in the binding specificity of phospholipid through the conserved Trp430 and Arg432 residues (Fig. 3b) that are demonstrated to be biologically essential [39].

3.3. Communities of coupled residues in the Mmm1-Mdm12 complex

As a further way to explore the inferences, we assessed whether residue communities play a role in protein–protein interaction and lipid transport. In practice, we performed calculations on the Mmm1-Mdm12 complex to identify the determinants in those communities. To achieve accurate inferences, the MSA of each domain in the complex are searched against the sequence profile database (see Methods). The estimated residue communities of each domain are mapped to the structure of the Mmm1-Mdm12 complex (Fig. 5).

From the overall calculations on sequence profiles, we observed distinct residue communities and determinants in the complex of Mmm1-Mdm12 (Fig. 5a). As investigated [39–41], Mmm1 interacts with Mdm34 (Fig. S2) through Mdm12 via relatively weak or

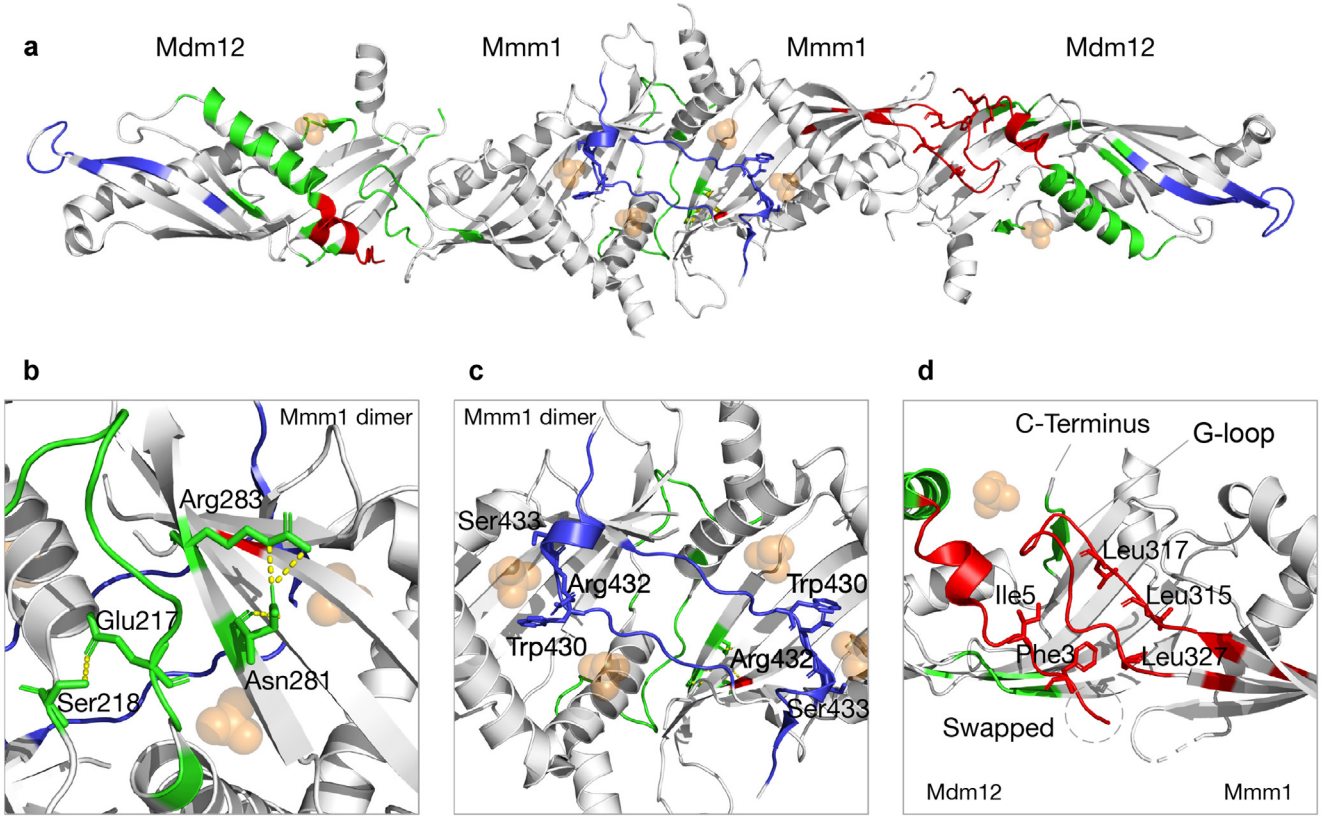

Fig. 5. Communities of coupled residues mapping to the tertiary structure of Mmm1-Mdm12 complex (PDB ID:5YK7). (a) The top three communities of coupled residues are mapped in red, green, and blue. (b) A pattern of co-evolution and the co-evolved residues inferred by the SAEC method are shown in green as sticks 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 sticks in blue. (d) Critical determinants of the interaction between Mmm1 and Mdm12.

transient interactions. In addition, The studies [39–41] suggested that Mdm34 and Mdm12 interact with each other through their N termini, especially via β-strand swapping [41] (residues in red at the N-terminus of the Mdm12 as shown Fig. 5d). The identified residues between the Mmm1 and Mdm12 suggest the three proteins, Mmm1, Mdm12, and Mdm34, can bind to each other at the interaction sites within the predicted residue communities (Fig. 5a). The binding interface between Mmm1 monomers is characterized by the blue and green communities, and the green community consists of a well-ordered loop that contacts the head region of the other molecule of the Mmm1 dimer. Moreover, the G-loop, as shown in the red community, (the extended hairpin loop at the C-terminus [41]) of the Mmm1 monomer plugs into Mdm12, and it introduces conformational change when the complex assembles [39] (Fig. 6). The co-evolution is distinctly revealed by two residues, Asn281 and Arg283 (Fig. 5b), among Mmm1 orthologs.

The G-loop (consisting of residues 425–432, Fig. 5c) of the Mmm1 covers the concave surface at the center of the dimeric SMP domain, and in the loop, there are two absolutely conserved Trp430 and Arg432 residues that are essential for the recognition of phospholipids. An extensive hydrogen-bonding network (consisting of the conserved Arg253, Arg415, Trp411, Trp430, Arg432, and Ser433) coordinates the phosphate group and carboxyl oxygen of the distal phospholipid. Among those residues, three residues (Trp430, Arg432, and Ser433) are statistically significant in the residue communities that are inferred by the SAEC method, which is consistent with their contribution to homodimerization for the lipid coordination in Mmm1[39]. The G-loop of Mmm1 can plug into the head region of Mdm12 and cover the solvent-exposed concave surface of Mdm12 in an extended conformation (Fig. 5d and Fig. 6). In particular, the hydrophobic amino acids Leu315, Leu317, and Leu327 in Mmm1 form extensive and coordinate nonpolar contacts with the side chains of Phe3, and Ile5 of Mdm12 (Fig. 5d). Those hydrophobic contacts play a critical role in the Mdm12-Mmm1 interaction.

A conformational change of the G-loop makes an extended structure to plug into the head region of the scMdm12Δ and cover the solvent-exposed concave surface of scMdm12Δ [39]. The spectrum of the couplings among amino acids is consistent with the conformational changes of the G-loop as illustrated in Fig. 6. To demonstrate that residue community (in red) is relevant to the conformational change, the two structures of Mmm1 as a single domain and in the complex are aligned to each other for comparison (Fig. 6). The statistical qualities of the residue communities (in green) do not capture the important information from the G-loop when Mmm1 adopts a shape as a monomer, while the Mmm1 changes its shape via the G-loop (residue community in red) in response to the formation of a complex between the Mmm1 and Mdm12 proteins.

4. Discussion

Leri allows systematic and optimized estimations of the partitioned interactions among amino acids (residue communities) that may not be predictable from physically contacted residues (e.g., the distance between C_β-C_β of pairwise residues), such as co-evolving amino acids at function sites/interfaces, and formation of function interface from a network of residue interactions. These inferences can, in turn, provide mechanistic explanations for experimental observations by identifying the key determinants involved in protein function. Accordingly, *Leri* provides a computational framework in which predictions and experimental evidence can be integrated to infer the key principles at the network level from homologous protein sequences.

In this study, we demonstrated that coupling between amino acid pairs can be grouped into networks, termed residue communities, in which the amino acids have stronger interactions in intra-networks but weaker interactions in inter-networks. We also demonstrate that identified residue communities are functionally relevant in protein specificity, e.g protein–protein interactions, phospholipid-binding activity, and conformational changes.

Although the inferences of the SAEC method are able to predict key residues related to protein function in the highly ordered communities, it might not be successful in identifying the communities if the multiple aligned homologous sequences suffer from limitations such as lack of diversity in the sequences and/or biases arising from evolutionary constraints across a protein family. Moreover, there is still a challenge that remains in computational resources, and this may make it difficult in exploring the whole space of a large protein and interpreting the signatures that are identified in the communities. In spite of those challenges, the proposed SAEC approach, supported by the *Leri* webserver, is readily applicable for analysis and identification of the residues communities that are relevant to protein functions. We anticipate that the analyses from the SAEC method will benefit the scientific community of biochemists and biophysicists for a better understanding of proteins and their subsequent engineering.

Data and code availability

The web-server is freely available for non-commercial use at https://kornmann.bioch.ox.ac.uk/leri/services/ecs.html. The stan-

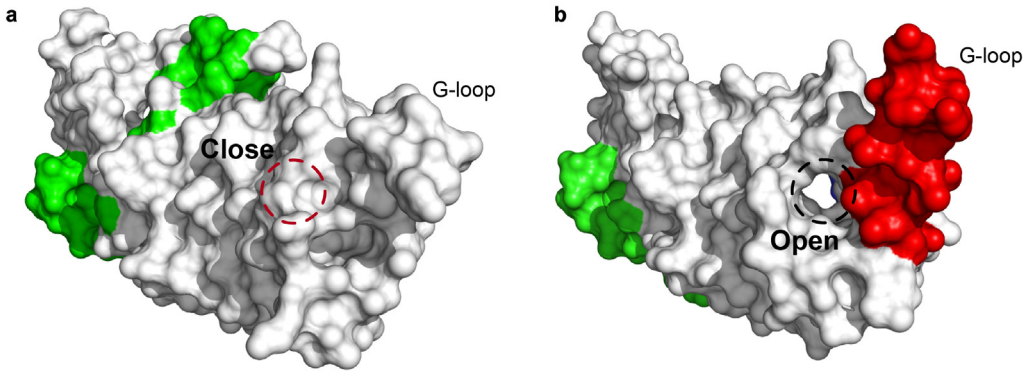

Fig. 6. The G-loop of Mmm1. The inferred residue communities are highlighted in green and red, and the dashed circles in red and black are to indicate the position of the channel (open or close). (a) The G-loop in a single domain (PDB5YK6) is not captured by the SAEC method. (b) The G-loop causing a conformational change in the Mmm1-Mdm12 complex (PDB5YK7) is identified from the homologous sequences.

dalone package of *Leri* working on the Linux system is available upon request for non-commercial use.

Declaration of Competing Interest

Potential conflicts of interest. N.J.C. (Y. Z.) is a founder of Leri Ltd based in Oxford, UK. All other authors report no conflicts of interest relevant to this article.

Acknowledgements

We acknowledge Dr. Sabine van Schie and Dr. Christian Covill-Cooke for discussion and proofreading the manuscript. We also thank all members of the Kornmann's Group for helpful discussions. ATJP is supported by the Spark grant of the Swiss National Science Foundation (SNSF).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csbj.2021.06.002.

References

[1] Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature 2005;437(7058):512–8.
[2] Shah NH, Kuriyan J. Understanding molecular mechanisms in cell signaling through natural and artificial sequence variation. Nature Struct Mol Biol 2019;26(1):25–34.
[3] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Proc National Acad Sci 2009;106(1):67–72.
[4] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc National Acad Sci 2011;108(49):E1293–301.
[5] Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E 2013;87(1):012707.
[6] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PloS One 2011;6(12).
[7] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 2012;149(7):1607–21.
[8] Stiffler MA, Poelwijk FJ, Brock KP, Stein RR, Riesselman A, Teyra J, Sidhu SS, Marks DS, Gauthier NP, Sander C. Protein structure from experimental evolution. Cell Syst 2020;10(1):15–24.
[9] Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell 2009;138(4):774–86.
[10] Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, Hilvert D, Monasson R, Cocco S, Weigt M, et al. An evolution-based model for designing chorismate mutase enzymes. Science 2020;369(6502):440–5.
[11] McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature 2012;491(7422):138–42.
[12] Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Struct Biol 2003;10(1):59–69.
[13] Tee W-V, Tan ZW, Lee K, Guarnera E, Berezovsky IN. Exploring the allosteric territory of protein function. J Phys Chemi B 2021.
[14] Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. Mol Biol Evol 2016;33(1):268–80.
[15] Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. Nature Biotechnol 2017;35(2):128.
[16] Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. Elife 2018;7:e34300.
[17] Tian P, Louis JM, Baber JL, Aniana A, Best RB. Co-evolutionary fitness landscapes for sequence design. Angewandte Chemie International Edition 2018;57(20):5674–8.
[18] Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnol 2008;26(10):1135.
[19] Blow N. DNA sequencing: generation next-next. Nature Methods 2008;5(3):267–74.
[20] Ansorge WJ. Next-generation DNA sequencing techniques. New biotechnology 2009;25(4):195–203.
[21] Baldessari F, Capelli R, Carloni P, Giorgetti A. Coevolutionary data-based interaction networks approach highlighting key residues across protein families: The case of the G-protein coupled receptors. Comput Struct Biotechnol J 2020;18:1153–9.
[22] Lutz S, Bornscheuer UT. Protein engineering handbook. John Wiley & Sons; 2012.
[23] Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. Nature 2016;537(7620):320–7.
[24] Quijano-Rubio A, Yeh H-W, Park J, Lee H, Langan RA, Boyken SE, et al. De novo design of modular and tunable protein biosensors. Nature 2021;591(7850):482–7.
[25] Cheung NJ, Yu W. Sibe: a computation tool to apply protein sequence statistics to predict folding and design in silico. BMC Bioinformatics 2019;20(1):1–11.
[26] Lutz S. Beyond directed evolution-semi-rational protein engineering and design. Current Opinion Biotechnol 2010;21(6):734–43.
[27] Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol 2011;7(10).
[28] Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 2017;45(D1):D170–6.
[29] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31(6):926–32.
[30] Cheung NJ, Yu W. De novo protein structure prediction using ultra-fast molecular dynamics simulation. PloS one 2018;13(11).
[31] Abdi H. The eigen-decomposition: Eigenvalues and eigenvectors. Encyclopedia Measur Stat 2007:304–8.
[32] Plotly Technologies Inc., Collaborative data science (2015). https://plotly.com..
[33] Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics 2018;34(21):3755–8.
[34] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. Pfam: the protein families database. Nucleic Acids Res 2014;42(D1):D222–30.
[35] Schrödinger, LLC, The PyMOL molecular graphics system, version 1.8, schrödinger, llc. (November 2015)..
[36] Kullback S, Leibler RA. On information and sufficiency. Annals Math Stat 1951;22(1):79–86.
[37] Kornmann B, Walter P. ERMES-mediated ER-mitochondria contacts: molecular hubs for the regulation of mitochondrial biology. J Cell Sci 2010;123(9):1389–93.
[38] Kornmann B, Currie E, Collins SR, Schuldiner M, Nunnari J, Weissman JS, Walter P. An ER-mitochondria tethering complex revealed by a synthetic biology screen. Science 2009;325(5939):477–81.
[39] Jeong H, Park J, Jun Y, Lee C. Crystal structures of Mmm1 and Mdm12–Mmm1 reveal mechanistic insight into phospholipid trafficking at ER-mitochondria contact sites. Proc National Acad Sci 2017;114(45):E9502–11.
[40] AhYoung AP, Jiang J, Zhang J, Dang XK, Loo JA, Zhou ZH, Egea PF. Conserved SMP domains of the ERMES complex bind phospholipids and mediate tether assembly. Proc Nat Acad Sci 2015;112(25):E3179–88.
[41] Jeong H, Park J, Lee C. Crystal structure of Mdm12 reveals the architecture and dynamic organization of the ERMES complex. EMBO Rep 2016;17(12):1857–71.