PLOS ONE

RESEARCH ARTICLE

Widespread evolutionary crosstalk among protein domains in the context of multi-domain proteins

David Jakubec[1,2], Miroslav Kratochvíl[1,3], Jiří Vymětal[1], Jiří Vondrášek[1]*

1 Department of Bioinformatics, Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, 166 10 Prague 6, Czech Republic, **2** Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University, 128 43 Prague 2, Czech Republic, **3** Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, 118 00 Prague 1, Czech Republic

* jiri.vondrasek@uochb.cas.cz

Citation: Jakubec D, Kratochvíl M, Vymětal J, Vondrášek J (2018) Widespread evolutionary crosstalk among protein domains in the context of multi-domain proteins. PLoS ONE 13(8): e0203085. https://doi.org/10.1371/journal.pone.0203085

Editor: Narayanaswamy Srinivasan, Indian Institute of Science, INDIA

Received: May 24, 2018

Accepted: August 14, 2018

Published: August 31, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by The Ministry of Education, Youth and Sports, grant number LM2015047, http://www.msmt.cz/ (J Vo) and Institute of Organic Chemistry and Biochemistry of the CAS (RVO), grant number 61388963, www.uochb.cz/ (J Vo). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Domains are distinct units within proteins that typically can fold independently into recognizable three-dimensional structures to facilitate their functions. The structural and functional independence of protein domains is reflected by their apparent modularity in the context of multi-domain proteins. In this work, we examined the coupling of evolution of domain sequences co-occurring within multi-domain proteins to see if it proceeds independently, or in a coordinated manner. We used continuous information theory measures to assess the extent of correlated mutations among domains in multi-domain proteins from organisms across the tree of life. In all multi-domain architectures we examined, domains co-occurring within protein sequences had to some degree undergone concerted evolution. This finding challenges the notion of complete modularity and independence of protein domains, providing new perspective on the evolution of protein sequence and function.

Introduction

Domains are basic functional and structural elements of proteins. In addition to sequence mutations, protein evolution is driven by combining existing domains into novel arrangements. The modular nature of domains arises from their ability to adopt well-defined three-dimensional (3D) structures (Fig 1A) that often facilitate their functions independently of their sequential surroundings. [1–3] Most eukaryotic proteins contain multiple domains [4, 5], and interactions among these domains can mediate allosteric regulation [6] or give rise to novel domain functions different from those found in isolation or other domain arrangements.

Structural changes driven by mutations in the primary sequence are one mechanism underlying the acquisition of novel domain functions. [4, 7, 8] Structural and functional constraints often require that evolution be coordinated between groups of amino acid residues in proteins (Fig 1B). [9–11] Covariation in amino acid composition between positions in multiple sequence alignments (MSAs) can be indicative of physical interactions between the residues

Competing interests: The authors have declared that no competing interests exist.

and has been used to aid prediction of protein 3D structures and conformational diversity. [12–17] Massive utilization of coevolutionary information has been made possible recently by the availability of high-quality MSAs containing data from high-throughput sequencing experiments. [14, 16, 17, 19, 20] Coevolutionary signals have been described both within and among domains that coexist within protein chains. [12, 21] In addition, coevolution has been also observed between proteins and their protein or nucleic acid interacting partners. [22–24]

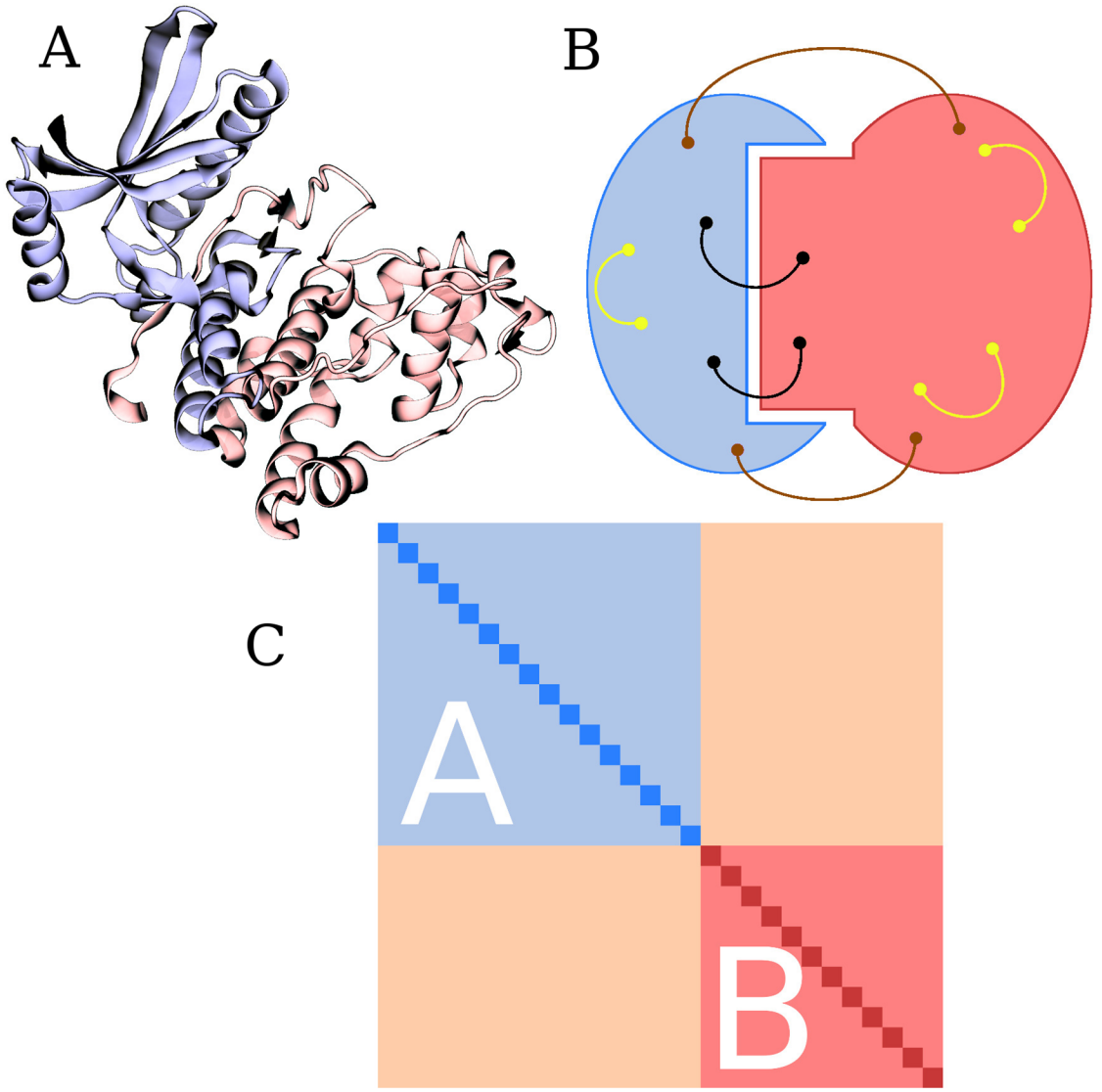

Fig 1. Representations of a two-domain architecture. A: Example structure of a two-domain protein. Chain A in Protein Data Bank [18] entry 1WAK contains two PF00069 (protein kinase) domains, shown here in blue and red. B: Correlated mutations in multi-domain proteins. Blue and red areas schematically depict individual domains involved in an interaction. Dots connected by solid lines represent pairs of coevolving positions. Coevolving positions within individual domains are shown in yellow. Coevolving positions localized in different domains are shown in black and brown. These can involve both positions forming physical contacts at the inter-domain interface (black) as well as positions separated by large distances (brown). C: General structure of a mutual information matrix for a two-domain architecture *AB*. White A and B labels denote individual domains. Diagonal dark blue and dark red elements describe the entropy of individual positions within domain *A* or *B*, respectively. Off-diagonal light blue and light red elements describe the mutual information between pairs of positions within domain *A* or *B*, respectively. Off-diagonal pink elements describe the mutual information between pairs of positions belonging to different domains.

https://doi.org/10.1371/journal.pone.0203085.g001

In this work, we present an information-theoretic analysis of coevolutionary signals among protein domains in multi-domain arrangements. Based on the functional implications these signals carry, we test the notion of evolutionary and functional independence of domains and examine their adaptability to their primary sequence context. In contrast to the aforementioned studies, we examine coevolution as a global property of a domain pair, and introduce an appropriate continuous measure to quantify its effect. Using this measure, we show that coevolution among protein domains is a much more widespread phenomenon than previously anticipated.

Materials and methods

Data set construction

We defined protein domains as sequence families recognized in release 31.0 of the Pfam database. [19] Pfam 31.0 provides a collection of 16,712 profile hidden Markov models (HMMs), each representing one protein sequence family, as well as MSAs containing alignments of sequences in UniProtKB [20] and other databases to these profile HMMs. Alignments of sequences included in release 2016_10 of the UniProt reference proteomes (URPs) to the Pfam 31.0 profile HMMs were obtained from the Pfam FTP repository. This URPs release contains a total of 26,742,727 protein sequences comprising the proteomes of 6,266 completely sequenced organisms. At least one match to a Pfam 31.0 profile HMM was recognized in 19,419,549 of these sequences. A total of 16,479 Pfam 31.0 profile HMMs matched to at least one sequence from the URPs.

Domain architecture was established for each URPs sequence. We defined the domain architecture of a protein as a vector of Pfam 31.0 sequence families identified within the protein sequence ordered according to their proximity to the N-terminus. A total of 278,458 distinct domain architectures were recognized. In order to reduce small-sample effects [25, 26], only architectures realized in at least 500 URPs sequences were considered; a total of 2,063 such architectures contained two or more domains and were thus selected for this study. A total of 4,240,857 URPs sequences were recognized as having one of these highly populated multi-domain architectures (HPAs). A total of 2,599 distinct Pfam 31.0 sequence families were identified within these sequences.

For each HPA, we compiled a list of all URPs sequences in which it was realized. Sequences of domains found in these proteins aligned to the respective Pfam 31.0 profile HMMs were retrieved from the Pfam family MSAs. Proteins that contained only standard amino acid, insert, and delete state symbols in the alignments of sequences of each of their domains to the respective Pfam 31.0 profile HMMs were identified for all HPAs. The lists of UniProt identifiers of these proteins for individual architectures are available in S1 File; the distribution of the numbers of sequences with these architectures is shown in S1 Fig.

Residue symbols found at positions corresponding to match or delete states (consensus columns) in the alignments of the respective domain sequences to the profile HMMs were extracted from each domain sequence within these proteins. This action corresponds to localizing all but the insert state positions in the respective MSAs, as residues assigned to insert states are, by definition, unaligned, and therefore irrelevant to this study. [27] Sequences of domains composed of the residues found in the consensus columns were then concatenated for each protein, creating a string composed of residues characteristic of each domain identified within the URPs sequence. For example, if protein i contained two domains A and B with respective sequences A_i and B_i, the concatenated sequence $A_i||B_i$ was created. By generating this string for each protein, we created a multi-domain MSA for each HPA.

Normalized mutual information measure

Individual columns in a MSA can be viewed as random variables, with residue symbols found in the columns acting as the values of their respective observations. The Shannon entropy H(X) of a random variable X taking on values from a finite alphabet $\mathcal{K} = \{x_1, x_2, \ldots, x_K\}$ can be estimated as

$$\mathrm{H}(X) = -\sum_{i=1}^{K} f(x_i) \log_2 f(x_i), \tag{1}$$

where $f(x_i)$ is the relative frequency of observing x_i and 0 $\log_2$ 0 is defined as zero. The mutual information MI(X, Y) of a pair of random variables X, Y can be estimated as

$$\mathrm{MI}(X, Y) = \sum_{i=1}^{K}\sum_{j=1}^{L} f(x_i, y_j) \log_2 \frac{f(x_i, y_j)}{f(x_i) f(y_j)}, \tag{2}$$

where $f(x_i, y_j)$ is the joint frequency of observing x_i and y_j simultaneously. Since 2 is chosen as the base of the logarithms in Eqs 1 and 2, values of entropy and mutual information are in bits. [28] Throughout this work, K and L are equal to 21, as all sequences in the MSAs contain only symbols for the standard amino acids and the delete state.

A mutual information matrix (MIM) showing the values of mutual information between each pair of columns within a MSA was calculated for multi-domain MSAs corresponding to the 2,063 selected multi-domain HPAs. The general structure of a MIM for a two-domain architecture is shown in Fig 1C, and an example of a MIM for an architecture consisting of two protein kinase (Pfam entry PF00069) domains is shown in Fig 2. In addition to correlated mutations within individual domains, these representations reveal positive values of mutual information between positions corresponding to different domains.

We calculated the average entropy $\overline{H}_D$ of positions corresponding to domain D for each domain within each HPA as

$$\overline{H}_D = \frac{1}{n}\sum_{i=1}^{n} \mathrm{H}(X_{i,D}), \tag{3}$$

where n is the number of positions $X_{i,D}$ corresponding to domain D. This corresponds to calculating the average values of the dark blue or dark red elements for the MIM shown in Fig 1C.

There are a total of $\binom{N}{2}$ unique domain pairs for an architecture consisting of N domains. For all such pairs of domains D, E within each architecture (regardless of whether they are sequential neighbors or not), the average value of mutual information $\overline{MI} \equiv \overline{MI}_{D,E}$ between positions corresponding to the two domains was calculated as

$$\overline{MI}_{D,E} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{MI}(X_{i,D}, Y_{j,E}), \tag{4}$$

where m, n are the numbers of positions $X_{i,D}$, $Y_{j,E}$ corresponding to domains D and E, respectively. This corresponds to calculating the average value of matrix elements in pink rectangles in the general MIM shown in Fig 1C.

We calculate the normalized average inter-domain mutual information $n\overline{MI} \equiv n\overline{MI}_{D,E}$ as a ratio of the average inter-domain mutual information $\overline{MI}_{D,E}$ and the arithmetic average of

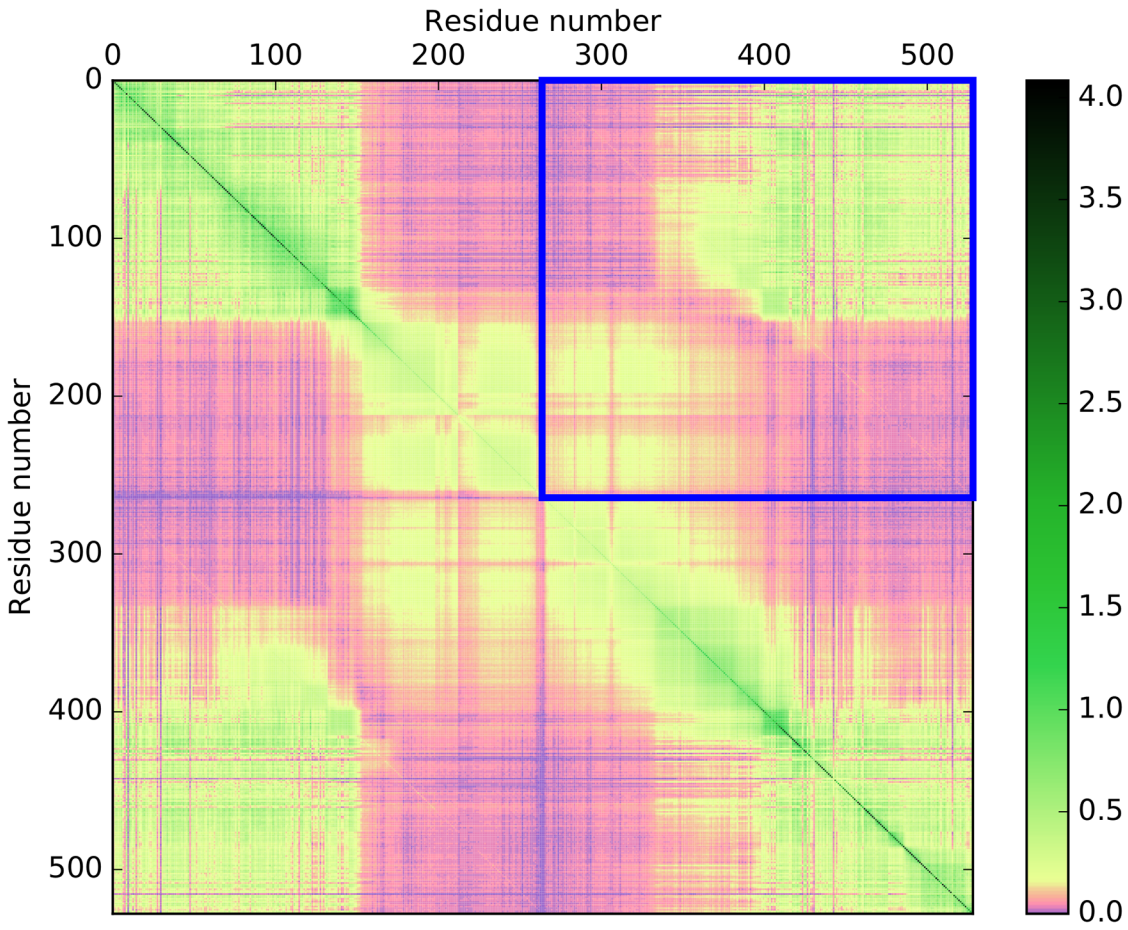

Fig 2. Mutual information matrix for the native PF00069–PF00069 (protein kinase–protein kinase) two-domain architecture. Each protein kinase domain contains 264 residues. A total of 8,753 URPs sequences have this architecture. The respective MSA consists of sequences of the two domains found within each of these URPs sequences. Note the non-linear color scale and the positive values of mutual information between positions corresponding to different domains (highlighted with the blue square). Values of entropy and mutual information are in bits.

https://doi.org/10.1371/journal.pone.0203085.g002

average entropies of positions corresponding to domains forming the respective pair, *i.e.*,

$$n\overline{MI}_{D,E} = \frac{2}{\overline{H}_D + \overline{H}_E}\overline{MI}_{D,E}. \quad (5)$$

The value of $n\overline{MI}$ obtained in this way is equivalent to the statistical measure known as symmetric uncertainty. [29] It represents, on the scale from 0 to 1, the extent of evolutionary coupling between domains *D* and *E*, independent of the internal sequence variability of each domain. There were a total of 5,205 domain pairs for which the value of $n\overline{MI}$ was calculated according to Eq 5.

Information-theoretic analysis

To compare the extent of evolutionary crosstalk among protein domains in the selected multi-domain architectures with the crosstalk in sequences that share no evolutionary history, we have designed the following test. First, we split strings consisting of concatenated domain sequences at the domain boundaries. If domain architecture *AB* was found in *N* protein sequences and the

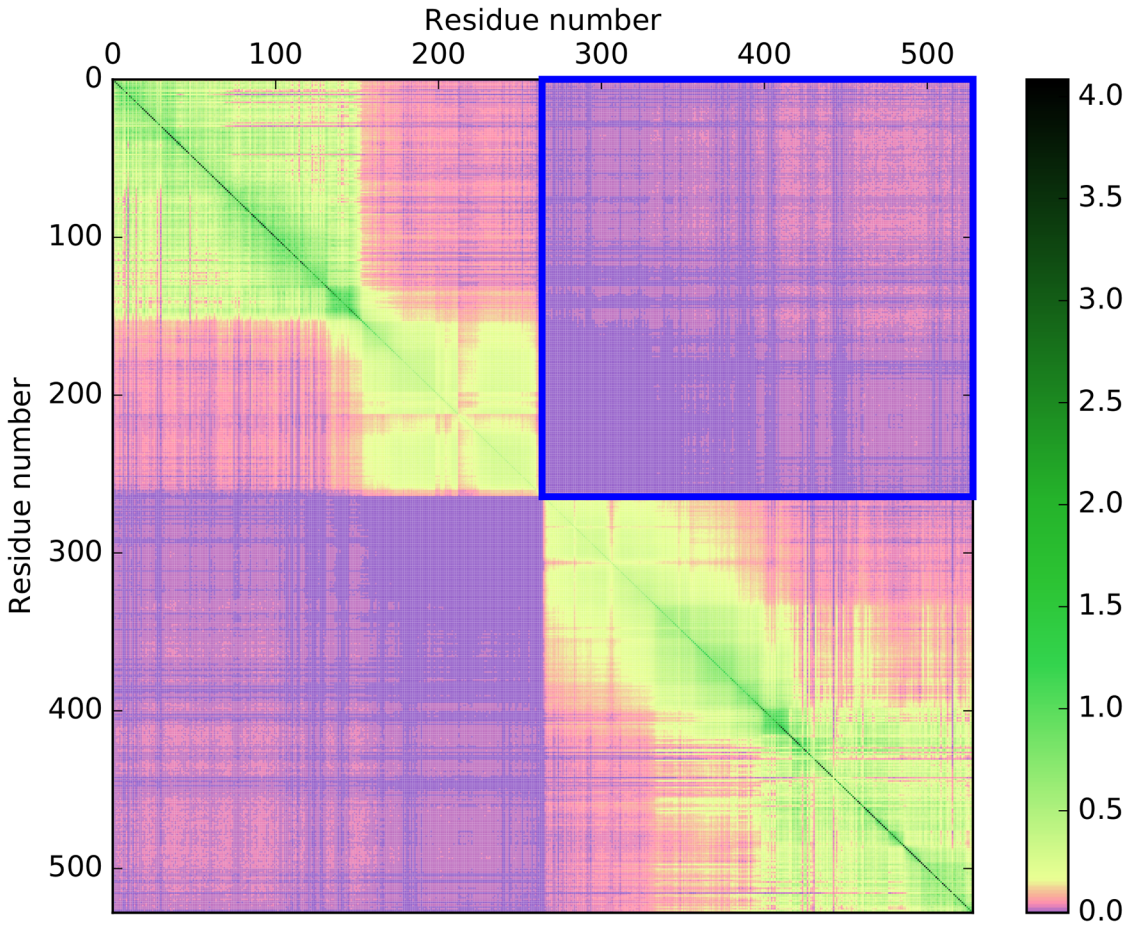

Fig 3. Mutual information matrix for the perturbed PF00069–PF00069 (protein kinase–protein kinase) domain architecture. Sequences of individual domains were randomly shuffled and rejoined. Note how the mutual information between positions corresponding to different domains (blue square) has vanished. Values of entropy and mutual information are in bits. The color scale is the same as in Fig 2.

https://doi.org/10.1371/journal.pone.0203085.g003

corresponding multi-domain MSA contained sequences $A_1||B_1$, $A_2||B_2$, . . ., $A_N||B_N$, then two sets of sequences, $\{A_1, A_2, \ldots, A_N\}$ and $\{B_1, B_2, \ldots, B_N\}$, were obtained after the original sequences had been split. After this splitting, sequences of individual domains within these sets were then randomly shuffled and rejoined so that the original architecture was reestablished. For example, if protein *i* with domain architecture *AB* originally contained a concatenated domain sequence $A_i||B_i$, the shuffling process resulted in the sequence $A_i||B_j$, with *j* being a different URPs sequence with an *AB* architecture. This way, any native evolutionary coupling among protein domains was disrupted.

For each HPA, we calculated a second MIM from the perturbed MSA of disrupted sequences in which fragments corresponding to individual domains almost certainly originated from different proteins. For each pair of domains in each architecture, the value of $n\overline{MI}$ was calculated based on the respective perturbed MSA as described above for the case of native sequences. An example of such a MIM for the architecture consisting of two protein kinase (PF00069) domains is shown in Fig 3. Here, one can clearly see that the inter-domain mutual information decreased remarkably in comparison with Fig 2.

To further examine the influence of background information noise arising among uncorrelated sequences, we studied the effects of random sequence composition fluctuations on the

resulting values of $n\overline{MI}$. A total of 100,000 amino acid sequences, each containing 100 randomly selected residues, were generated. The probabilities of choosing individual amino acids were the same at each position within the sequences and were obtained from the average primary sequence composition of proteins in release 2017_08 of the UniProtKB/Swiss-Prot database. [20] For each of the 2,599 unique Pfam 31.0 sequence families identified within the HPAs, 500 sequences were randomly chosen from the corresponding MSA of domain sequences. Each of these sequences was then concatenated with a randomly chosen sequence from the set of 100,000 sequences with random composition. In this way, a sort of a two-domain MSA was generated, in which very high-entropy positions correspond to one of the domains. The value of $n\overline{MI}$ was then calculated for each of these pseudoarchitectures.

Results

There were a total of 5,205 domain pairs for which the values of $n\overline{MI}$ were calculated before and after intra-architectural domain sequence shuffling. The distributions of resulting $n\overline{MI}$ values are shown in Fig 4A and 4B, respectively; the raw obtained values are available in S1 File. The difference between these values was calculated for each domain pair. The distribution

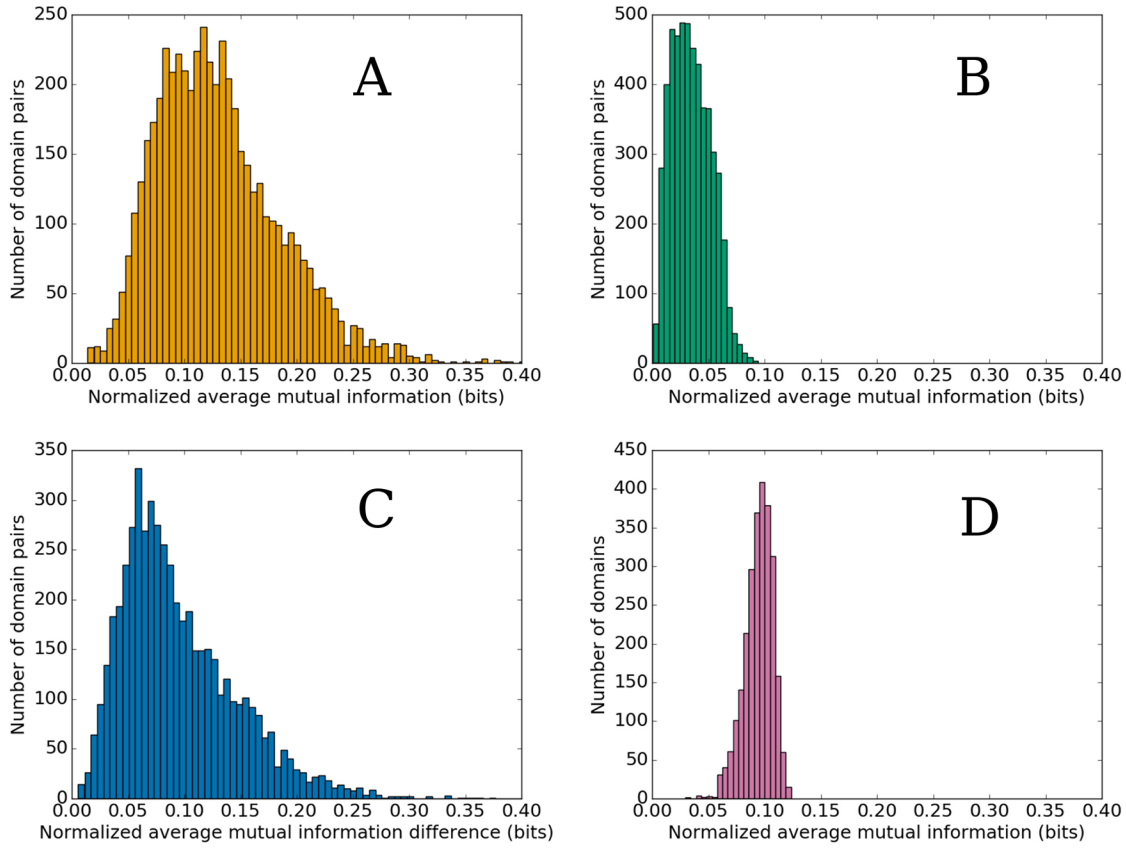

Fig 4. Distributions of the values of $n\overline{MI}$ for various domain pairings. A: Distribution of the values ($N = 5,205$) of $n\overline{MI}$ for domain pairs in native (non-disrupted) multi-domain protein sequences. B: Distribution of the values ($N = 5,205$) of $n\overline{MI}$ for domain pairs after intra-architectural domain sequence shuffling. C: Distribution of differences ($N = 5,205$) between the values of $n\overline{MI}$ for individual domain pairs before and after intra-architectural domain sequence shuffling. D: Distribution of the values ($N = 2,599$) of $n\overline{MI}$ for domain–random sequence pairs.

https://doi.org/10.1371/journal.pone.0203085.g004

of these differences is shown in Fig 4C. For illustration, the value of this difference is ≈ 0.068 bits for the exemplary PF00069–PF00069 two-domain architecture.

It is clear from Fig 4C that all differences between the values of $n\overline{MI}$ before and after intra-architectural domain sequence shuffling are greater than zero, *i.e.*, the $n\overline{MI}$ is always greater before the domain sequence shuffling. As a hypothesis, this is confirmed by performing the Wilcoxon signed-rank test [30] on the distribution of these differences, which yields both the value of the test statistic and the p-value of 0.0. The difference between the distributions shown in Fig 4A and 4B can also be confirmed by performing the two-sided two-sample Kolmogorov–Smirnov (KS) test [31], which yields the value of the KS statistic of ≈0.869 and the corresponding p-value effectively zero.

It should be noted that both the distribution of the values of $n\overline{MI}$ in native sequences (Fig 4A) and the distribution of the $n\overline{MI}$ differences (Fig 4C) have means of around 0.1 bits, which can be considered significant, given that maximum entropy of individual positions in the MSAs is on the order of 10^0 bits.

Fig 4D shows the distribution of the values of $n\overline{MI}$ for sequences consisting of natural domains and random sequences. Here, one can see that the distribution differs significantly from that observed for native multi-domain sequences (Fig 4A) and appears to be more similar to the distribution obtained after domain sequence shuffling (Fig 4B). However, a deeper statistical inspection performed using the two-sided two-sample KS test and the two-sided Mann–Whitney U test [32] shows that both non-random distributions are significantly different (all p-values are effectively zero). This result shows that, even though each individual position in the artificial random sequences has nearly maximum possible entropy, and thus has a large potential to generate considerable values of mutual information with positions in genuine protein sequences due to statistical noise, the observed values of $n\overline{MI}$ differ significantly from those observed in natural multi-domain architectures. Therefore, it seems unlikely that the evolutionary coupling observed among domains in genuine multi-domain proteins would be a result of random fluctuations in amino acid residue frequencies.

In addition, the apparent similarity of the distributions shown in Fig 4B and 4D implies that the domain sequence shuffling procedure has reduced the inter-domain sequence covariation score almost to the level expected to result from statistical noise. However, it should be noted that there is a weak–intermediate positive linear correlation between the values of $n\overline{MI}$ before and after domain sequence shuffling for individual domain pairs (Fig 5). Therefore, it seems that not all contained information could be eliminated using this approach. This could be related to a similar positive linear correlation observed between the non-normalized values of $\overline{MI}$ (Eq 4) and the respective average domain pair entropies $\frac{\overline{H}_D+\overline{H}_E}{2}$ (Fig 6). We provide an explanation for the observation of these correlations in Discussion.

It is worth noting that the values of $n\overline{MI}$ and the respective average domain pair entropies are virtually uncorrelated (Pearson correlation coefficient $r \approx -0.001$; Spearman's rank correlation coefficient $\rho \approx -0.005$).

Discussion

While most studies of coevolution to date (for example, [11–17, 21–24] and many others) have treated it as a discrete property, in this paper we study the degree of coevolution as a continuous property. Unlike the mentioned studies, our aim has not been the identification of precise "coevolving" pairs of residues in close spatial proximity, but rather the mapping of the overall tendency of residues in a pair of domains to respond to mutations in their partner. The value of $n\overline{MI}$ introduced here is an intrinsically global measure which quantifies this effect for a pair

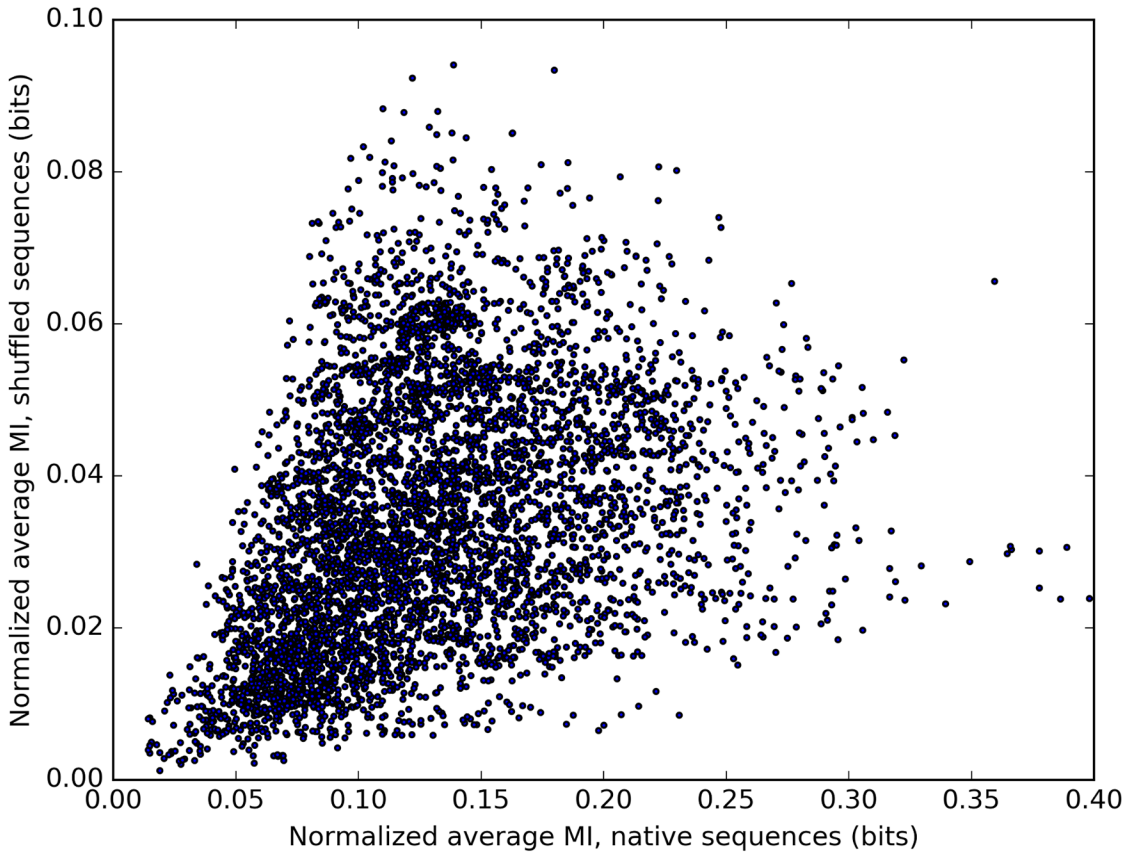

Fig 5. Correlation between the values of $n\overline{MI}$ before and after intra-architectural domain sequence shuffling. Number of domain pairs (data points) $N = 5,205$. The value of the Pearson correlation coefficient $r \approx 0.345$; the value of the Spearman's rank correlation coefficient $\rho \approx 0.432$.

https://doi.org/10.1371/journal.pone.0203085.g005

of domains without specifying which precise residue pairs contribute to this measure significantly. This makes it serve a different and unique role compared to recent coevolutionary analyses for 3D contact prediction, such as DCA [12] and PSICOV. [33]

In this analysis, we deliberately ignored the issue of transitivity, where two domains may appear to be coevolving if they share a common partner which is coevolving with both of them. It may not be possible to reliably quantify and filter out such effects from purely numerical data when coevolution is treated as a continuous property and some signal is observed for all pairs of domains (a possible linear optimization-based solution is highly numerically unstable). In addition, a pair of domains could, in principle, be coevolving with some other conserved non-domain region, such as a linker, and these effects could not be filtered out even if an appropriate method existed. It is worth noting that most (1,390 out of 2,063) architectures studied in this work consist of only two protein domains, which rules out the possibility of domain-related transitivity effects.

We explain the correlation between the values of $n\overline{MI}$ before and after domain sequence shuffling, and the related correlation between the non-normalized values of $\overline{MI}$ and the respective average domain pair entropies as follows. When two variables (positions in a MSA) each have a large entropy, there is a greater chance that mutual information will appear between the variables due to random noise, as mutual information can only increase with increasing entropy. Therefore, if one calculates the value of $\overline{MI}$ for a pair of domains with

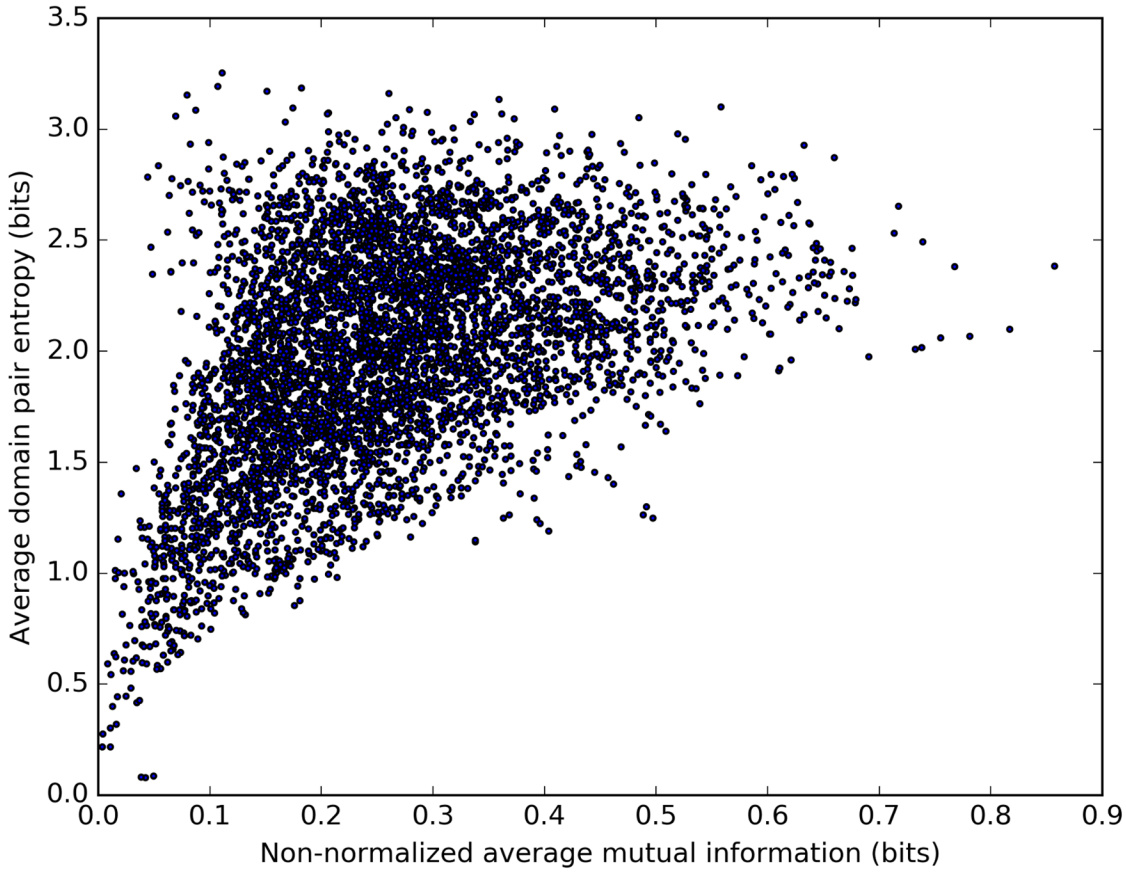

Fig 6. Correlation between the non-normalized values of $\overline{MI}$ and average domain pair entropies. Number of domain pairs (data points) $N = 5,205$. The value of the Pearson correlation coefficient $r \approx 0.464$; the value of the Spearman's rank correlation coefficient $\rho \approx 0.460$.

https://doi.org/10.1371/journal.pone.0203085.g006

large average entropies, one can expect the result to be greater as a consequence of an increased statistical noise (Fig 6). As the intra-architectural domain sequence shuffling has no effect on entropies of individual positions, the increased chance to generate mutual information from random noise remains unchanged after domain sequences are shuffled. This mutual information can compensate some of the loss introduced by the domain sequence shuffling. Therefore, domain pairs with large average entropies can produce larger values of $\overline{MI}$ both in native and in shuffled multi-domain sequences, leading to the observed correlation (Fig 5).

In addition to the random noise factor described above, there is another contribution to the correlation observed in Fig 6, caused by the natural bounds on the value of mutual information between two variables. This value can never exceed the intrinsic entropy of either variable. Therefore, two positions with small entropies can never produce a large value of mutual information, whereas positions with large entropies may yield both small and large mutual information. This asymmetry can contribute to the observed correlation.

The value of $n\overline{MI}$ calculated for a pair of protein domains after intra-architectural domain sequence shuffling serves as a proxy for the value expected if domains from different proteins were paired randomly and thus shared no evolutionary history with each other. We found that the corresponding value of $n\overline{MI}$ calculated from the alignment of native (non-disrupted) multi-domain sequences is always greater (Fig 4C). This result implies that, in the context of

multi-domain proteins, a portion of domain sequence variation can always be attributed to coordinated evolution among different domains.

Coordinated evolution can be intuitively understood by the need to preserve essential protein function in cases in which multiple domains form a ligand-binding or catalytic site (Fig 1B). [12, 21] We propose the following possible explanation for observation of this phenomenon even among domains lacking such apparent functional constraints: Co-localizing multiple domains into the same polypeptide chain may influence the folding pathway or open new paths to optimize protein function *via* inter-domain interactions. These interactions may enable direct or allosteric modulation of the function of the complete protein or its individual domains. If a vital function of a multi-domain protein depends on the cooperative action of its domains, evolution may opt to distribute mutations needed to preserve this function across the domains in a coordinated fashion. Residues preferentially mutated in this way may constitute nodes of energetic connectivity in the protein structure analogous to those observed at the single domain level. [34] It should be noted that $n\overline{MI}$, as defined, is a measure representing overall evolutionary coupling of two domains, and does not provide detailed insight into which specific amino acid residue pairs contribute significantly to this coupling.

In either case, additional domains act as buffers or reservoirs of evolutionary capacity that can be utilized to either mitigate the impact of mutations required to maintain proper protein function or, alternatively, to optimize the respective functions of individual domains. The precise mechanism through which this functional modulation is realized and its full impact on protein evolution remain to be established.

Conclusion

We showed that, in the context of multi-domain proteins, evolution of domain sequences proceeds in a coordinated fashion. We proved this by comparing a mutual information-based measure between native multi-domain sequences and artificial sequence constructs which share no common evolutionary history and further showed that the observed evolutionary coupling is distinct from statistical noise.

Supporting information

S1 File. Primary data archive for the studied architectures. This archive contains a file presenting the list of UniProt identifiers of the URPs sequences included in the multi-domain MSAs for each of the 2,063 studied multi-domain architectures, and files containing the values of $\overline{MI}$ and $n\overline{MI}$ for individual domain pairs before and after domain sequence shuffling.
(GZ)

S1 Fig. Distribution of the numbers of sequences in the MSAs. Total number of architectures (MSAs) $N = 2,063$.
(TIF)

Acknowledgments

We thank to M. Oliveberg, M. M. Babu, and T. Majerova for reading the manuscript and for their valuable comments and suggestions. We are also grateful to M. Oliveberg for the concept of Fig 1B. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated. Computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.

Author Contributions

Conceptualization: David Jakubec, Miroslav Kratochvíl, Jiří Vymětal, Jiří Vondrášek.

Data curation: David Jakubec, Miroslav Kratochvíl.

Formal analysis: David Jakubec, Miroslav Kratochvíl, Jiří Vymětal.

Funding acquisition: Jiří Vondrášek.

Investigation: David Jakubec, Miroslav Kratochvíl, Jiří Vymětal.

Methodology: David Jakubec, Miroslav Kratochvíl, Jiří Vymětal.

Project administration: David Jakubec, Jiří Vondrášek.

Software: David Jakubec, Miroslav Kratochvíl, Jiří Vymětal.

Supervision: Jiří Vondrášek.

Validation: David Jakubec, Miroslav Kratochvíl.

Visualization: David Jakubec, Miroslav Kratochvíl.

Writing – original draft: David Jakubec, Jiří Vondrášek.

Writing – review & editing: David Jakubec, Miroslav Kratochvíl, Jiří Vymětal, Jiří Vondrášek.

References

1. Björklund ÅK, Ekman D, Light S, Frey-Skött J, Elofsson A. Domain Rearrangements in Protein Evolution. J Mol Biol. 2005; 353: 911–923. https://doi.org/10.1016/j.jmb.2005.08.067 PMID: 16198373
2. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in the modular evolution of proteins. Trends Biochem Sci. 2008; 33: 444–451. https://doi.org/10.1016/j.tibs.2008.05.008 PMID: 18656364
3. Bornberg-Bauer E, Albà MM. Dynamics and adaptive benefits of modular protein evolution. Curr Opin Struct Biol. 2013; 23: 459–466. https://doi.org/10.1016/j.sbi.2013.02.012 PMID: 23562500
4. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol. 2004; 14: 208–216. https://doi.org/10.1016/j.sbi.2004.03.011 PMID: 15093836
5. Han J-H, Batey S, Nickson AA, Teichmann SA, Clarke J. The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol. 2007; 8: 319–330. https://doi.org/10.1038/nrm2144 PMID: 17356578
6. Schueler-Furman O, Wodak SJ. Computational approaches to investigating allostery. Curr Opin Struct Biol. 2016; 41: 159–171. https://doi.org/10.1016/j.sbi.2016.06.017 PMID: 27607077
7. Bashton M, Chothia C. The Generation of New Protein Functions by the Combination of Domains. Structure. 2007; 15: 85–99. https://doi.org/10.1016/j.str.2006.11.009 PMID: 17223535
8. Lees JG, Dawson NL, Sillitoe I, Orengo CA. Functional innovation from changes in protein domains and their combinations. Curr Opin Struct Biol. 2016; 38: 44–52. https://doi.org/10.1016/j.sbi.2016.05.016 PMID: 27309309
9. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. Biochemistry. 2005; 44: 7156–7165. https://doi.org/10.1021/bi050293e PMID: 15882054
10. Liu Y, Bahar I. Sequence Evolution Correlates with Structural Dynamics. Mol Biol Evol. 2012; 29: 2253–2263. https://doi.org/10.1093/molbev/mss097 PMID: 22427707
11. Neuwald AF. Gleaning structural and functional information from correlations in protein multiple sequence alignments. Curr Opin Struct Biol. 2016; 38: 1–8. https://doi.org/10.1016/j.sbi.2016.04.006 PMID: 27179293
12. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 2011; 108: E1293–E1301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

13. Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci U S A. 2013; 110: 20533–20538. https://doi.org/10.1073/pnas.1315625110 PMID: 24297889

14. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013; 110: 15674–15679. https://doi.org/10.1073/pnas.1314045110 PMID: 24009338

15. Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. Proc Natl Acad Sci U S A. 2015; 112: 13567–13572. https://doi.org/10.1073/pnas.1508584112 PMID: 26487681

16. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. Elife. 2015; 4: e09248. https://doi.org/10.7554/eLife.09248 PMID: 26335199

17. Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, et al. Protein structure determination using metagenome sequence data. Science. 2017; 355: 294–298. https://doi.org/10.1126/science.aah4043 PMID: 28104891

18. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, et al. The Protein Data Bank. A Computer-Based Archival File for Macromolecular Structures. Eur J Biochem. 1977; 80: 319–324. https://doi.org/10.1111/j.1432-1033.1977.tb11885.x PMID: 923582

19. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44: D279–D285. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716

20. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017; 45: D158–D169. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622

21. Yeang C-H, Haussler D. Detecting Coevolution in and among Protein Domains. PLOS Comput Biol. 2007; 3: e211. https://doi.org/10.1371/journal.pcbi.0030211 PMID: 17983264

22. Yang S, Yalamanchili HK, Li X, Yao K-M, Sham PC, Zhang MQ, et al. Correlated evolution of transcription factors and their binding sites. Bioinformatics. 2011; 27: 2972–2978. https://doi.org/10.1093/bioinformatics/btr503 PMID: 21896508

23. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. 2014; 3: e03430. https://doi.org/10.7554/eLife.03430

24. Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H, Weigt M. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. Proc Natl Acad Sci U S A. 2017; 114: E2662–E2671. https://doi.org/10.1073/pnas.1615068114 PMID: 28289198

25. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics. 2009; 25: 1125–1131. https://doi.org/10.1093/bioinformatics/btp135 PMID: 19276150

26. Mao W, Kaya C, Dutta A, Horovitz A, Bahar I. Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. Bioinformatics. 2015; 31: 1929–1937. https://doi.org/10.1093/bioinformatics/btv103 PMID: 25697822

27. Eddy SR, the HMMER development team. HMMER User's Guide Version 3.2.1; June 2018. http://eddylab.org/software/hmmer/Userguide.pdf.

28. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics. 2005; 21: 4116–4124. https://doi.org/10.1093/bioinformatics/bti671 PMID: 16159918

29. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C: The Art of Scientific Computing. Cambridge: Cambridge University Press; 1992.

30. Wilcoxon F. INDIVIDUAL COMPARISONS BY RANKING METHODS. Biometrics Bull. 1945; 1: 80–83. https://doi.org/10.2307/3001968

31. Massey FJ, Jr. THE KOLMOGOROV–SMIRNOV TEST FOR GOODNESS OF FIT. J Am Stat Assoc. 1951; 46: 68–78. https://doi.org/10.1080/01621459.1951.10500769

32. Mann HB, Whitney DR. ON A TEST OF WHETHER ONE OF TWO RANDOM VARIABLES IS STOCHASTICALLY LARGER THAN THE OTHER. Ann Math Stat. 1947; 18: 50–60. https://doi.org/10.1214/aoms/1177730491

33. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28: 184–190. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153

34. Lockless SW, Ranganathan R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science. 1999; 286: 295–299. https://doi.org/10.1126/science.286.5438.295 PMID: 10514373