Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon

RESEARCH ARTICLE

Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon

Christoph Feinauer[1], Hendrik Szurmant[2], Martin Weigt[3,4]*, Andrea Pagnani[1,5]*

1 Department of Applied Science and Technology, and Center for Computational Sciences, Politecnico di Torino, Torino, Italy, **2** Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA, United States of America, **3** Sorbonne Universités, UPMC, UMR 7238, Computational and Quantitative Biology, Paris, France, **4** CNRS, UMR 7238, Computational and Quantitative Biology, Paris, France, **5** Human Genetics Foundation, Molecular Biotechnology Center (MBC), Torino, Italy

* martin.weigt@upmc.fr (MW); andrea.pagnani@polito.it (AP)

Abstract

Interaction between proteins is a fundamental mechanism that underlies virtually all biological processes. Many important interactions are conserved across a large variety of species. The need to maintain interaction leads to a high degree of co-evolution between residues in the interface between partner proteins. The inference of protein-protein interaction networks from the rapidly growing sequence databases is one of the most formidable tasks in systems biology today. We propose here a novel approach based on the *Direct-Coupling Analysis* of the co-evolution between inter-protein residue pairs. We use ribosomal and trp operon proteins as test cases: For the small resp. large ribosomal subunit our approach predicts protein-interaction partners at a true-positive rate of 70% resp. 90% within the first 10 predictions, with areas of 0.69 resp. 0.81 under the ROC curves for all predictions. In the trp operon, it assigns the two largest interaction scores to the only two interactions experimentally known. On the level of residue interactions we show that for both the small and the large ribosomal subunit our approach predicts interacting residues in the system with a true positive rate of 60% and 85% in the first 20 predictions. We use artificial data to show that the performance of our approach depends crucially on the size of the joint multiple sequence alignments and analyze how many sequences would be necessary for a perfect prediction if the sequences were sampled from the same model that we use for prediction. Given the performance of our approach on the test data we speculate that it can be used to detect new interactions, especially in the light of the rapid growth of available sequence data.

OPEN ACCESS

Citation: Feinauer C, Szurmant H, Weigt M, Pagnani A (2016) Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. PLoS ONE 11(2): e0149166. doi:10.1371/journal.pone.0149166

Editor: Ozlem Keskin, Koç University, TURKEY

Received: March 11, 2015

Accepted: January 28, 2016

Published: February 16, 2016

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

Funding: Christoph Feinauer and Andrea Pagnani received financial support from EU Marie Curie Training Network NETADIS, (FP7 Grant 290038, http://cordis.europa.eu/fp7/home_en.html). Hendrik Szurmant received financial support from the US National Institute of General Medical Sciences, National Institutes of Health, grants GM106085 and GM019416 (http://www.nigms.nih.gov/Research/Pages/default.aspx). Martin Weigt received financial support from Agence Nationale de la Recherche project COEVSTAT (ANR-13-BS04-0012-01, http://www.agence-nationale-recherche.fr/en/anr-funded-

Introduction

Proteins are the major work horses of the cell. Being part of all essential biological processes, they have catalytic, structural, transport, regulatory and many other functions. Few proteins exert their function in isolation. Rather, most proteins take part in concerted physical

project/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-13-BS04-0012). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

interactions with other proteins, forming networks of protein-protein interactions (PPI). Unveiling the PPI organization is one of the most formidable tasks in systems biology today. High-throughput experimental technologies, applied for example in large-scale yeast two-hybrid [1] analysis and in protein affinity mass-spectrometry studies [2], allowed a first partial glance at the complexity of organism-wide PPI networks. However, the reliability of these methods remains problematic due to their high false-positive and false-negative rates [3].

Given the fast growth of biological sequence databases, it is tempting to design computational techniques for identifying protein-protein interactions [4]. Prominent techniques to date include: the genomic co-localization of genes [5, 6] (with bacterial operons as a prominent example), the Rosetta-stone method [7] (which assumes that proteins fused in one species may interact also in others), phylogenetic profiling [8] (which searches for the correlated presence and absence of homologs across species), and similarities between phylogenetic trees of orthologous proteins [9–12]. Despite the success of all these methods, their sensitivity is limited due to the analysis of coarse global proxies for protein-protein interaction. An approach that exploits more efficiently the large amount of information stored in multiple sequence alignments (MSA) seems therefore promising.

Recently, a breakthrough has been achieved using genomic sequences for the related problem of inferring residue contacts from sequence data alone. [13]. The so-called Direct-Coupling Analysis (DCA) [14, 15] allows to construct statistical models that are able to describe the sequence variability of large MSA of homologous proteins [16]. More precisely, these models reproduce the empirically measured covariations of amino acids at residue pairs. The parameters of the models unveil networks of direct residue co-evolution, which in turn accurately predict residue-residue contacts.

These models are computationally hard to infer and several approximations have therefore been developed [14, 15, 17, 18]. While models based on the mean-field approximation borrowed from statistical physics [15, 18] are fast, approximations based on pseudo-likelihood maximization [17, 19] are more accurate and used throughout this paper.

When applied to two interacting protein families, DCA and related methods are able to detect inter-protein contacts [14, 20, 21] and thereby to guide protein complex assembly [22, 23]. This is notable since contact networks in protein complexes are strongly modular: There are many more intra-protein contacts than inter-protein contacts. Moreover, DCA helps to shed light on the sequence-based mechanisms of PPI specificity [24–26].

Here we address an important question: Is the strength of inter-protein residue-residue co-evolution sufficient to *discriminate interacting from non-interacting pairs of protein families*, i.e. to infer PPI networks from sequence information? A positive answer would lever the applicability of these statistical methods from structural biology (residue contact map inference) to systems biology (PPI network inference). An obvious problem in this context is the sparsity of PPI networks, illustrated by the bacterial ribosomal subunits used in the following, cf. Fig 1: The small subunit contains 20 proteins and 21 protein-protein interfaces (11% of all 190 possible pairs). In the large subunit, 29 proteins form 29 interfaces (7% of all 406 pairs). We see that while the number of potential PPI between N proteins is $\binom{N}{2}$, the number of real PPI grows only linearly as $O(N)$. Furthermore, the number of potentially co-evolving residue-residue contacts across interfaces is much smaller than the number of intra-protein contacts. In the case of ribosomes, only 5.8% of all contacts in the small subunit are inter-protein contacts. In the large subunit this fraction drops down to 4.5%. So the larger the number of proteins, the more our problem resembles the famous search of a needle in a haystack. The noise present in the large number of non-interacting protein family pairs might exceed the co-evolutionary signal of interacting pairs.

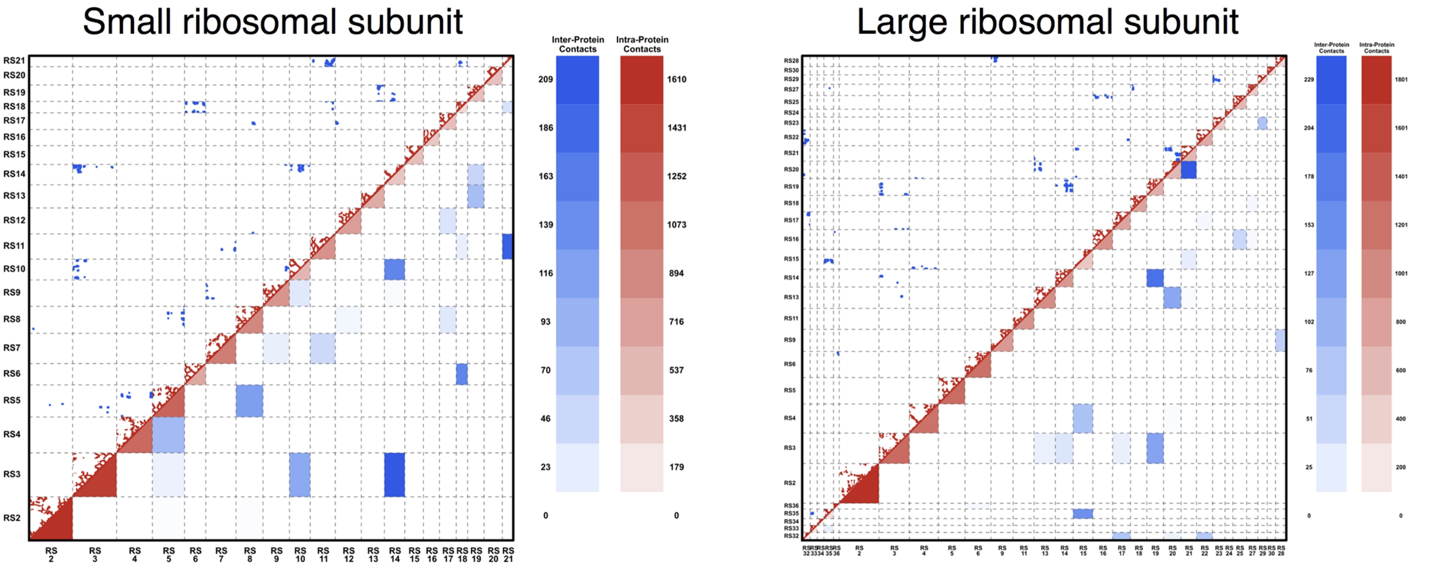

Fig 1. Contact map and protein-protein interaction network of small and large ribosomal subunits. The contact map and the protein-protein interaction network for **A** the small ribosomal subunit and **B** the large ribosomal subunit (proteins only), using a distance cutoff of 8Å between heavy atoms. The upper diagonal part shows the contact map, with red dots indicating intra-protein contacts, and blue dots inter-protein contacts. The lower triangular part shows the coarse graining into the corresponding protein-protein interaction networks, with the color levels indicating the number of intra- resp. inter-protein contacts, cf. the scales. The sparse character of both the contact network and the interaction network is clearly visible.

doi:10.1371/journal.pone.0149166.g001

It should also be mentioned that the ribosomal structure relies on the existence of ribosomal RNA, which is not included in our analysis. We therefore expect many of the small PPI interfaces to be of little importance for the ribosomal stability and that only large interfaces constrain sequence evolution and thus become detectable by co-evolutionary studies.

Ribosomal proteins and their interactions are essential and thus conserved across all bacteria, and it appears reasonable to wonder whether this makes them a specialized example of a protein complex more amenable to co-evolutionary bias. As a second and smaller interaction network, we therefore considered the enzymes of the tryptophan biosynthesis pathway comprising a set of seven proteins in which only two pairs are known to interact (PDB-ID 1qdl for the TrpE-TrpG complex [27] and 1k7f for the TrpA-TrpB complex [28]). Also here the PPI network is very sparse; most pairs are not known to interact, but might show some degree of coordinated evolution due to the fact that in many organisms these genes show a common spatial co-localization in a single operon and also due to a number of gene fusion events, cf. the discussion below. While widespread, the tryptophan biosynthesis pathway is not essential for viability when environmental tryptophan is present.

In this paper we report the performance of DCA in the prediction of protein-protein interaction partners in the systems tested. In a first step, we analyze the performance on data from an artificial model. This allows for a systematic analysis of the performance of different approaches and of the influence of the number of sequences in the alignment. With this artificial data set we are able to establish a lower-bound on the number of sequences that would make our predictions on the PPI scale completely accurate if the generating model was the same model we use for inference. Given the growth-rate of current protein sequence databases (notably UniProt [29]), we expect that such a lower bound could be met in few years. In a second step, we apply the method to the proteins of the bacterial ribosome and to the proteins of the trp operon, and show that the results obtained for simulated data translate well to the biological sequences of this test-set.

Materials and Methods

The goal of the present work is to analyze each of the $\binom{N}{2}$ possible pairs of multiple sequence alignments from a given set of N single-protein family alignments, and to extract a pairwise score that measures the co-evolution between the proteins in the alignments. A high co-evolutionary score is then taken as a proxy for interaction. In the spirit of [30] we describe in this section consecutively the *data generation and matching*, the *model* used for analyzing data and the *inference and scoring* mechanism.

Data extraction and matching for the ribosomal and trp operon proteins

The input data is given by N multiple sequence alignments D_p consisting of M_p sequences of length L_p for every protein family p. These alignments are extracted from UniProt [29] using standard bioinformatics tools, in particular Mafft [31] and HMMer [32] (*cf.* Section A in S1 Text for details on the extraction pipeline and Tables A and B in S1 Text listing the values of N, M_p, L_p for ribosomal and trp-operon proteins). For the analysis, it is necessary to concatenate the MSAs of two putative co-evolving protein families. This means to create, for each pair of protein families (p, p'), a new alignment $D_{p,\,p'}$ of sequence length $L_p + L_{p'}$. Each line contains the concatenation of two potentially interacting proteins. More precisely, in the case where families p and p' actually interact, each line should contain a pair of interacting proteins. The general problem of producing a concatenated alignment out of single MSAs of two protein families is straightforward in two cases only: (i) we have prior knowledge which pairs of sequences represent interaction partners; (ii) no paralogs are present in the considered species (*i.e.* all species have at most a single homolog of each of the sequences to be matched). Often, as displayed schematically in Fig 2, MSAs contain multiple protein sequences within a given species and no prior knowledge can be used to know who is (potentially) interacting with whom. In prokaryotes, interacting proteins are frequently found to be coded in joint operons. This suggests to use genomic co-localization as a matching criterion. To do so, as explained in Section B in S1 Text, we approximated the *genomic distance* between sequences using UniProt accession numbers. A better distance between sequences could be defined in terms of their

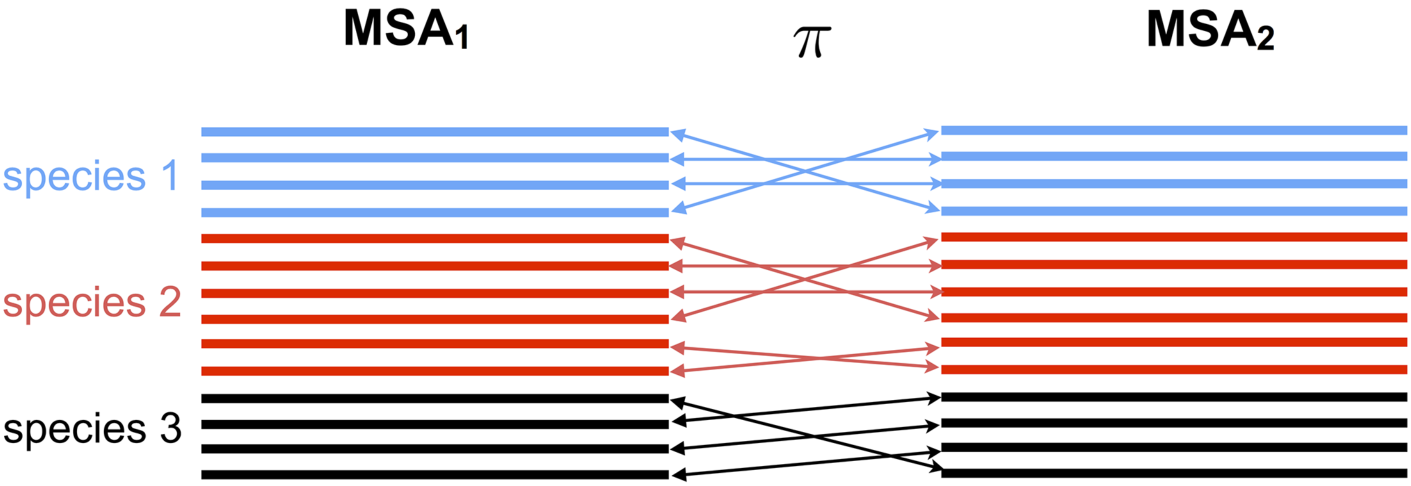

Fig 2. Concatenating two multiple sequence alignments. Sketch of the matching procedure that allows us to concatenate two different MSAs, here MSA_1,MSA_2. π represents the optimal permutation of the sequences on the second MSA computed using a standard linear programming routine.

doi:10.1371/journal.pone.0149166.g002

genomic location. Unfortunately, genomic locations are available only in the context of whole genome sequencing projects. The majority of sequences in Uniprot originate from fragments or from incomplete genome sequencing projects. These difficulties lead us to content ourselves with the proxy of accession numbers.

Having defined distances between each protein pair in the MSA, we calculate the matching which minimizes the average distance between matched sequences by linear programming. Additionally, we introduce a distance threshold used to discard matched distal protein sequence pairs. The numeric value for this threshold was determined using the small ribosomal subunit as a test case.

The average number of paralogs per species varies from system to system: For both ribosomal subunits the proteins have between 1.5 and 3 paralogous sequences per genome. The trp proteins on the other hand have considerably more paralogous sequences and the number of such sequences per genome varies between 4 and 24. This means that especially in the trp operon the matching procedure has the potential to generate much larger alignments than the competing approach of excluding species with paralogous sequences. In fact, using this last approach (which corresponds to setting our threshold parameter to 0) reduces the number of sequences in the alignments on the average by about 10% for the ribosomal proteins and by about 85% for the proteins of the trp operon (see Tables C-G in S1 Text).

Following the *ortholog conjecture*, paralogs are less likely to conserve function than orthologs and after duplication they may lose part of their interactions or gain others [33, 34]. This also means that it might be dangerous to include them in a co-evolutionary analysis, as the inclusion of non-interacting pairs would reduce the signal and increase the noise in the sequence data.

However, our matching strategy based on genomic vicinity excludes proteins coming from isolated genes; it identifies mostly protein pairs coded in gene pairs colocalized inside operons. In agreement with [4, 35–37] we assume that in such a case the maintainance of genomic colocalization is an indication for the maintainance of interaction, if the original protein pair was also interacting. While being somewhat speculative, we observe that this procedure removes most paralogs in the systems under study: Even if many genomes contain a large number of paralogs before matching (see above), in 99.8% of all genomes in the matched alignments for ribosomal protein pairs only a single sequence pair is found, while for trp protein pairs the same holds for 82% of all genomes. In other words, if there are paralogs in a species the matching algorithm tends to select one single pair, at least in the systems we studied.

We will show evidence that, in the interacting protein systems investigated here, this strategy leads to a reinforced coevolutionary signal as compared to including only genomes without paralogs. However, an independent and direct test whether protein pairs included in the alignment actually interact would constitute a big step forward, in particular since the arguments for using genomic colocalization hold chiefly for bacterial genomes.

Let us recall that the problem of finding a good matching between sequences has already been studied in the past using different strategies [24, 25]. Unfortunately, both methods are computationally too demanding to be used in a case, where hundreds or thousands of protein family pairs have to be matched.

Statistical sequence model

Within DCA, the probability distribution over amino acid sequences $x = (x_1, \ldots, x_L)$ of (aligned) length L is modeled by a so-called Potts model, or pairwise Markov Random Field,

$$\mathcal{P}(x) = \frac{1}{Z}\exp\left\{ \sum_{1\le i<j\le L} J_{ij}(x_i, x_j) + \sum_{1\le i\le L} h_i(x_i) \right\} , \qquad (1)$$

which includes statistical couplings $J_{ij}(x_i, x_j)$ between residue pairs and position-specific biases $h_i(x_i)$ of amino-acid usage [14]. The number Z is the normalization constant of $P(x)$, which is a probability distribution over all amino-acid sequences of length L. The variable x_i represents the amino acid found at position i in the sequence and can take as values any of the $q = 21$ different possible letters in an MSA (gaps are treated as a 21st amino acid). The model parameters are inferred using MSAs of homologous proteins.

In the case of two concatenated protein sequence $(x, x') = (x_1, \ldots, x_L, x'_1, \ldots, x'_{L'})$, the joint probability takes the form

$$\mathcal{P}(x, x') = \frac{1}{Z} e^{-H(x) - H'(x') - H^{int}(x, x')}. \tag{2}$$

The functions $H(x)$ and $H'(x')$ are the terms in the exponential in Eq (1) referring to each single protein. The function

$$H^{int}(x, x') = -\sum_{i \in x, j \in x'} J_{ij}(x_i, x'_j) \tag{3}$$

describes the co-evolutionary coupling between the two protein families. In the last expression, x_i is the ith amino acid in sequence x, and x'_j the jth amino acid in sequence x'. The sum runs over all inter-protein pairs of residue positions. The $q \times q$ matrices J_{ij} in this term quantify how strongly sites between the two proteins co-evolve in order to maintain their physicochemical compatibility. The matrix contains a real number for each possible amino acid combination at sites i and j and contributes to the probability in Eq 2 depending on whether an amino acid combination is favorable or not. The strongest inter-protein couplings are enriched for inter-protein contacts [14, 20]. The same kind of model can be used to predict the interaction between more than two proteins, with a corresponding number of interaction terms. However, the number of parameters in the model is proportional to $(L_1 + L_2 + .. + L_N)^2$ for N proteins while the number of samples in the concatenated MSA $D_{p_1,...,pN}$ becomes smaller because one has to find matching sequences for N proteins *simultaneously*. This leads us to consider the case $N > 2$ only for artificial proteins where the total length and sample size are controllable.

Inference and Scoring

Following [17], the parameters of the model were inferred by maximizing *pseudo-likelihood functions*. This is an alternative to directly maximizing the likelihood and considerably faster (see Section C in S1 Text Text for details). Given that the model is mathematically equivalent to the one used in [17] we can use the output of the algorithm (plmDCA) with default parameters as presented there directly for our purposes. This output consists of scores F_{ij} (the average-product corrected Frobenius norm of the matrices J_{ij}) that quantify the amount of co-evolution between sites i and j in the alignments. In order to quantify co-evolution between *proteins*, we took the F_{ij} corresponding to inter-protein site pairs (i.e. i in x and j in x') and calculated the mean of the 4 largest. These quantities, a real number for every protein pair, are used to rank protein-protein interaction partners. The number 4 was chosen because it performed well in the small ribosomal subunit, which we used as a test case when designing the algorithm. Subsequent tests on larger systems showed that any number between 1 and 6 performs almost equally well (see Section *A Global View* in *Results*).

Simulated data

As the basis for the simulated data we used a fictitious protein complex consisting of 5 proteins. Each protein has a length of 53 residues. The individual contact map of each one is given by the bovine pancreatic trypsin inhibitor (PDB ID 5pti [38]), which is a small protein performing well for the prediction of internal contacts by DCA. Each P_i has 551 internal contacts. Moreover, each protein interacts with two others in a circular way. The inter-protein contact matrices between P_i and P_{i+1} (as well as between P_1 and P_5) are random binary matrices with a density of 10% of the internal contacts. This models the sparsity of the inter-protein contacts as compared to the intra-protein contacts. A contact map for the artificial complex can be found in Fig E in S1 Text, There are no contacts between other pairs of proteins.

In order to define as realistically as possible the coupling parameters of the Potts model used for generating the artificial sequences, we used the Pfam protein family PF00014 of the pancreatic trypsin inhibitor [16]. Note that a member of this family was also used to define the structure. The couplings describing the co-evolution *within* the single proteins were directly extracted from the Pfam MSA using DCA. For the couplings corresponding to the co-evolution *between* the proteins, we used a random subset of the internal parameters and used them to couple sites that are in contact according the contact map as defined above. Non-contacting pairs of sites remain uncoupled between artificial proteins. Using this model, a joint MSA D_{12345} of sequences of length $265 = 5 \times 53$ was generated using standard MC simulations.

The process of defining the contact map, choosing the parameters and generating the sequences is described in Section E of S1 Text.

Results and Discussion

Testing the approach using simulated data

As a first test of our approach, we use *simulated data* generated by Monte Carlo (MC) sampling of a Potts model of the form of Eq (2), cf. *Materials and Methods*.

The main simplifying assumptions in this context are: (i) We assume intra- and inter-protein co-evolution strengths to be the same. (ii) We assume the distribution of inter-protein residues contacts within the possible contacts to be random. (iii) We assume the sequences to be identically and independently distributed according to our model. This model includes the assumption that non- contacting sites have zero couplings. The number of artificial sequences needed for a good performance of our method should therefore be taken at most as a lower bound for the number of biological sequences needed for a comparable performance.

In panel A of Fig 3 we show the architecture of our artificial protein complex. It is composed of five fictitious, structurally identical proteins $P_1, \ldots, P_5$, each one consisting of 53 residues. In order to simulate co-evolution between the proteins, we generate a *joint* MSA D_{12345} for all 5 proteins with a model that contains couplings between inter-protein site pairs. These couplings are modeled in a way to resemble couplings inferred from real proteins (see Materials and Methods).

To assess our capability to infer the PPI network of panel A from such data, we adopted two different strategies which we called *combined* and *paired* in panel B of Fig 3. The *combined* strategy uses plmDCA on the full-length alignments of length 265 and models the interaction between all proteins pairs *simultaneously*. Given that in this artificial setting we use the same model to generate the data as to analyze it, the approach is guaranteed to infer the model correctly for a large number of analyzed sequences and therefore to assign a higher interaction score to any interacting protein pair than to any non-interacting pair.

To assess the coupling strength between two proteins, we average the four strongest residue coupling strengths between them. This leads to a score oriented toward the strongest signal while also reducing noise by averaging. In panel B of Fig 3 we show the results for MSA sizes $M = 2000, 4000, 24{,}000$ while intermediate values are reported in Fig F in S1 Text. The two lower figures—$M = 2000, 4000$—represent the lower and upper bound of what we can currently obtain from databases for the proteins analyzed by us. The largest value $M = 24{,}000$ is

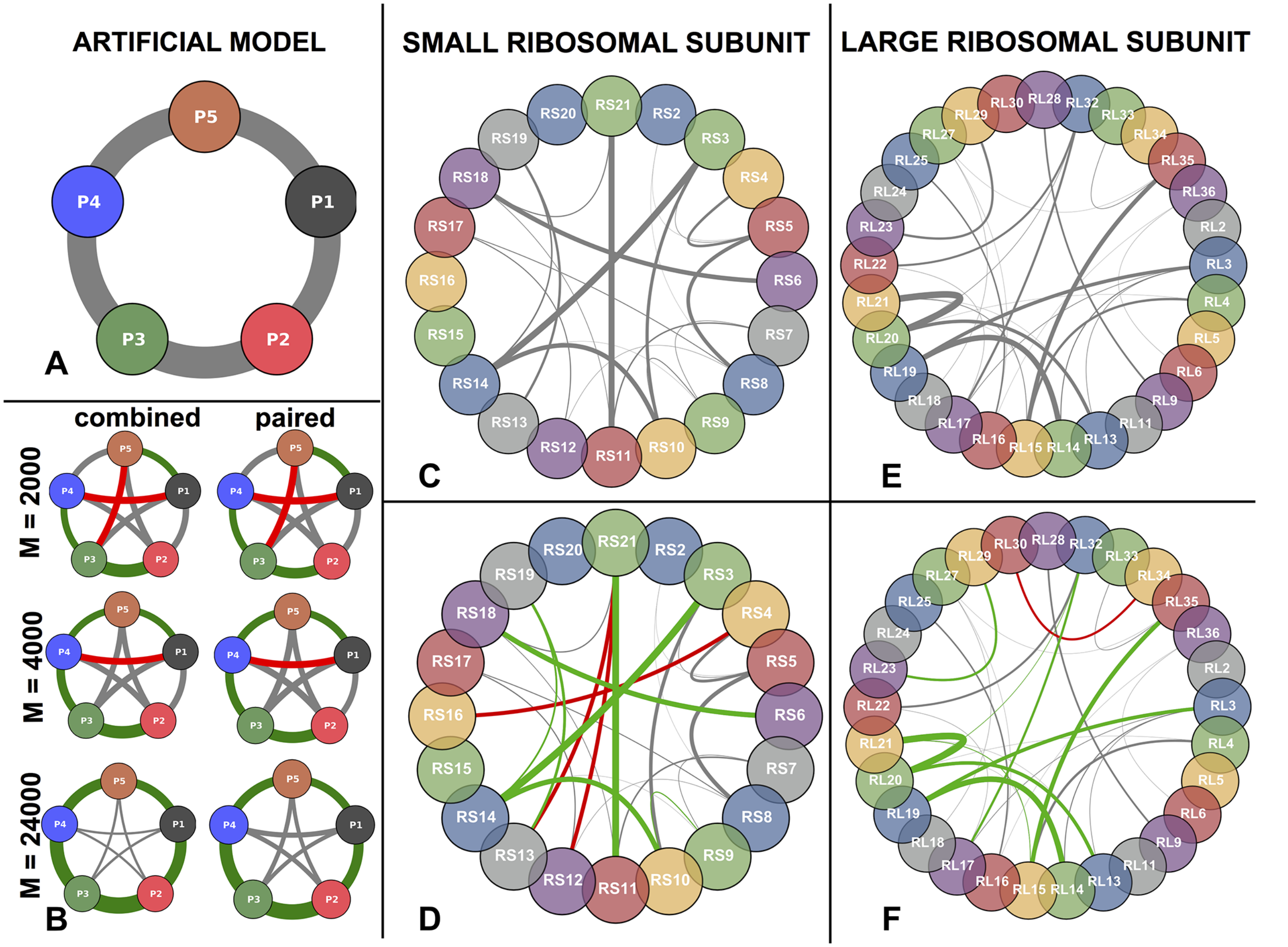

Fig 3. Residue-residue structure of both artificial and ribosomal complex. A Architecture of the *artificial* protein complex. Arcs width are proportional to the number of inter-protein residue contacts. **B** Inferred PPI network for both *paired* and *combined* strategy for different number *M* of sequences generated from the artificial model. Green arcs are true positives, red false positives, gray low-ranking predictions. Arc widths are proportional to the inter-protein interaction score. **C** SRU architecture (same color code as A). **D** Inferred PPI network (same color code as B). **E** Same as C for LRU. **F** Same as D for LRU. Arc width in panels C-F is provided by the number of inter-protein contacts, as a measure of interface size. It becomes obvious that mainly large interfaces are recognized by our approach.

doi:10.1371/journal.pone.0149166.g003

what we expect to be available in a few years from now, seen the explosive growth of sequence databases. The thickness of each link in Fig 3 is proportional to the inferred inter-protein interaction score. The five strongest links are colored in green when they correspond to actual PPI according to panel A, and in red when they correspond to non-interacting pairs. For increasing sample size the predictions become more consistent and for $M = 24{,}000$ any interacting protein-pair has a higher interaction score than any non-interacting pair.

Due to the running time of plmDCA only alignments for sequences of total length $L \lesssim 1000$ can be analyzed. This is exceeded already by the sum of the lengths of the proteins of the small ribosomal subunit. Additionally, creating a combined multiple sequence alignment for more than two proteins would lead to very low sequence numbers due to the necessary matching (see Materials and Methods). Therefore, using the combined strategy is not generally applicable. In the *paired* strategy we therefore analyze each pair of proteins separately. This means that plmDCA is applied to all $\binom{N}{2}$ protein-pair alignments D_{ab}, $1 \leq a < b \leq N$. In panel B of Fig 3 we find that the paired strategy is also able to detect the correct PPI network for large enough M. We observe, however, that the performance of the paired strategy is slightly worse. Couplings between non-interacting proteins are estimated significantly larger than using the combined strategy for large M. Even in the limit $M \to \infty$ we do not expect these links to disappear: Correlations between, e.g., P_1 and P_3 are generated via the paths 1—2—3 and 1—5—4—3, but in the paired strategy these correlations have to be modeled by direct couplings between P_1 and P_3 since the real direct coupling paths are not contained in the data.

After having answered the *'who-with-whom'* question for the artificial protein network, we address the *'how'* question of finding inter-protein contact pairs. Fig 4 panel A displays individual residue contact pairs within and between proteins in the artificial complex. Panel B shows the 10 strongest intra-protein couplings for each protein and the 10 strongest inter-protein couplings inferred by plmDCA (M = 4000, combined strategy). Green links correspond to contact pairs and red links to non-contact pairs. We see that the intra-protein prediction is perfect, whereas a few errors appear for inter-protein predictions in agreement with the results of Fig 3.

The PPI network of bacterial ribosomes

As a more realistic test we apply the method to the bacterial large and small ribosomal subunits (LRU, SRU). To define contacts and protein interaction partners we used high-resolution crystal structures with PDB-IDs 2z4k (SRU) and 2z4l (LRU) [39]. The contact network is summarized by the contact maps in Fig 1. The ribosomal RNA is ignored in our analysis.

Panels C,E of Fig 3 display the architectures of both SRU and LRU. The SRU (LRU) complex consists of 20 (29) proteins of lengths 51-218 (38-271); 21 (29) out of $\binom{20}{2} = 190$ ($\binom{29}{2} = 406$) pairs are in contact. The interfaces contain between 3-209 (1-229) residue pairs. The width of the inter-protein links in the PPI network Fig 3 in panels C,E are proportional to these numbers. The number of contacts within the individual proteins ranges from 297 to 2337 (303-2687). Globally, there are 22644 (30555) intra-protein and 1401 (1,439) inter-protein contacts, so the contacts relevant for our study comprise only 5.8% (4.5%) of all contacts.

Fig 3 panel D shows the inferred SRU PPI architecture. As expected, the biological case is harder than the artificial case where the data are independently and identically distributed according to the generating model. Even though the histograms of the inferred interaction scores for both cases are very similar (see Fig B in S1 Text), biological data are expected to show non-functional correlations due to the effect of phylogeny or sequencing efforts which are biased to model species and known pathogens. Nonetheless, among the top ten predicted interacting protein pairs the method makes only three errors (true-positive rate 70% as compared to $21/190 \simeq 11\%$ true PPI between all protein pairs, with an overall area under ROC curve (AUC) of 0.69 (see Fig 5). The method spots correctly the pairs with larger interaction surfaces whereas the small ones are lost. Two of the false-positive (FP) predictions include protein RS21, which has the smallest paired alignments with other proteins (M between 1468 and 1931). Also the third FP, corresponding to the pair RS4-RS18, is probably due to a small MSA with M = 2064. At the same time, the interaction of RS21 with RS11, which is one of the largest interfaces (199 contacts), is still detected despite the low M = 1729. The same procedure for the

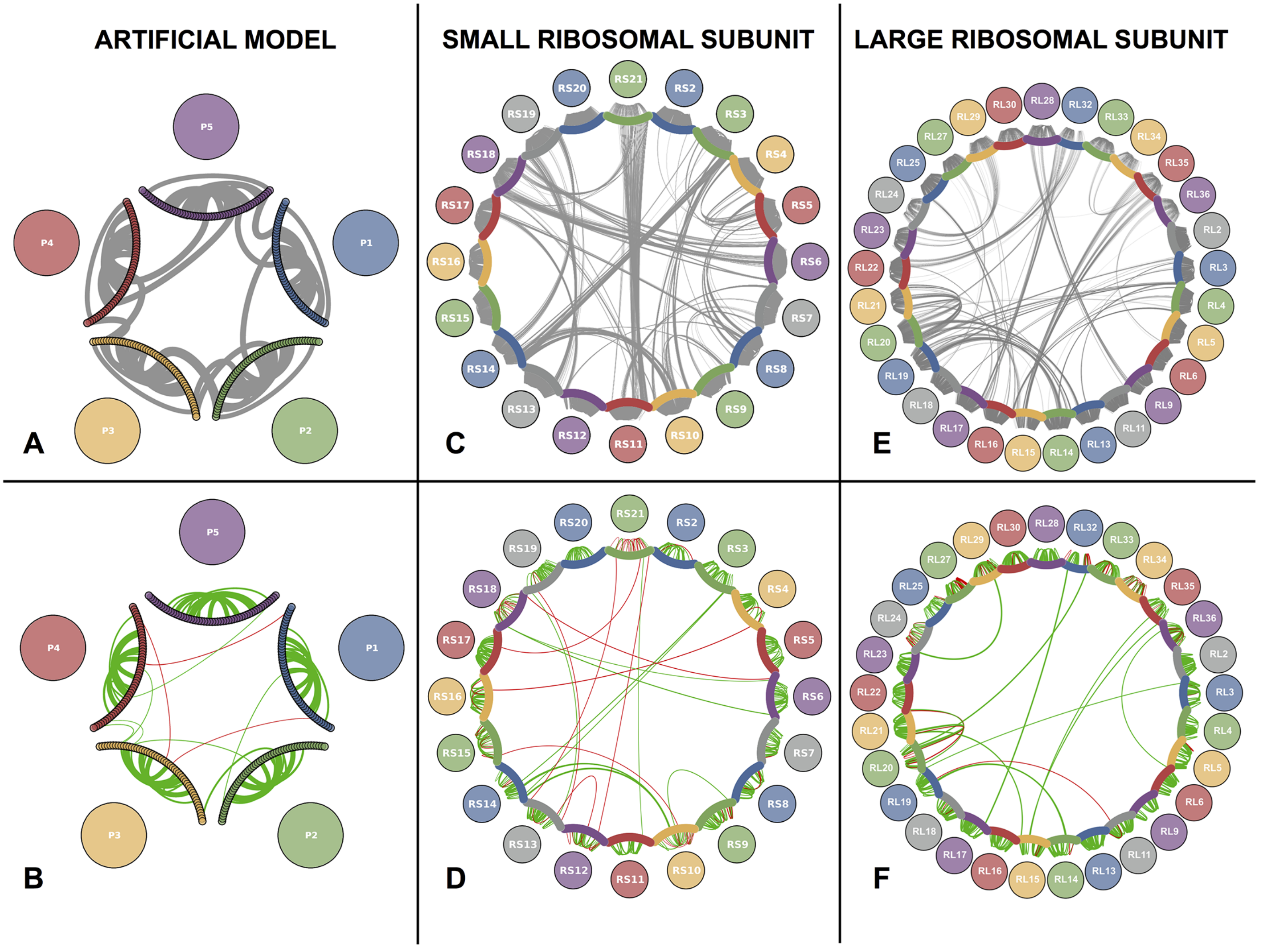

Fig 4. Architecture and inferred protein-protein interaction network of the artificial protein complex. A Residue-residue interaction structure of the generating model for the artificial data. Colored arcs represent the protein chain. Non-zero couplings in the coupling matrix of the generating model are represented as curves between the nodes. The width of the curves is proportional to the interaction score. Only the 10 strongest intra/inter-protein scores are shown. **B** Same as **A**, but based on the inferred couplings. Green arcs are true positives, red false positives. Note that not all green arcs have a corresponding arc in **A** due to our choice to display only the 10 strongest couplings, which not always correspond to the strongest score. **C** Same as **A** for SRU. All links represent a contact in the PDB structure and have equal width. **D** Same as **B** for SRU. **E** Same as **C** for LRU. All links represent a contact in the PDB structure and have equal width. **F** Same as D for LRU.

doi:10.1371/journal.pone.0149166.g004

LRU (406 protein pairs) performs even better: 9 out of the 10 first PPI predictions are correct (see Fig 3 panel F), and the AUC is 0.81.

The results on the residue scale for both SRU and LRU are depicted in panels D and F of Fig 4. Shown are the first 20 intra-protein residue contact predictions for each protein (excluding contacts with linear sequence separations below 5 to concentrate on non-trivial predictions) and the first 20 inter-protein residue contact predictions. In the SRU case of panel D for example, the results are qualitatively similar to the artificial case, albeit with a slightly reduced true-positive rate of 60% among the first 20 inter-protein residue contact predictions (compared to the ratio of 1401 actual inter-protein residue contacts and 2,403,992 possible inter-protein residue contacts, i.e., 0.058%). Again 3 out of the 8 false positives are related to RS21, which due to

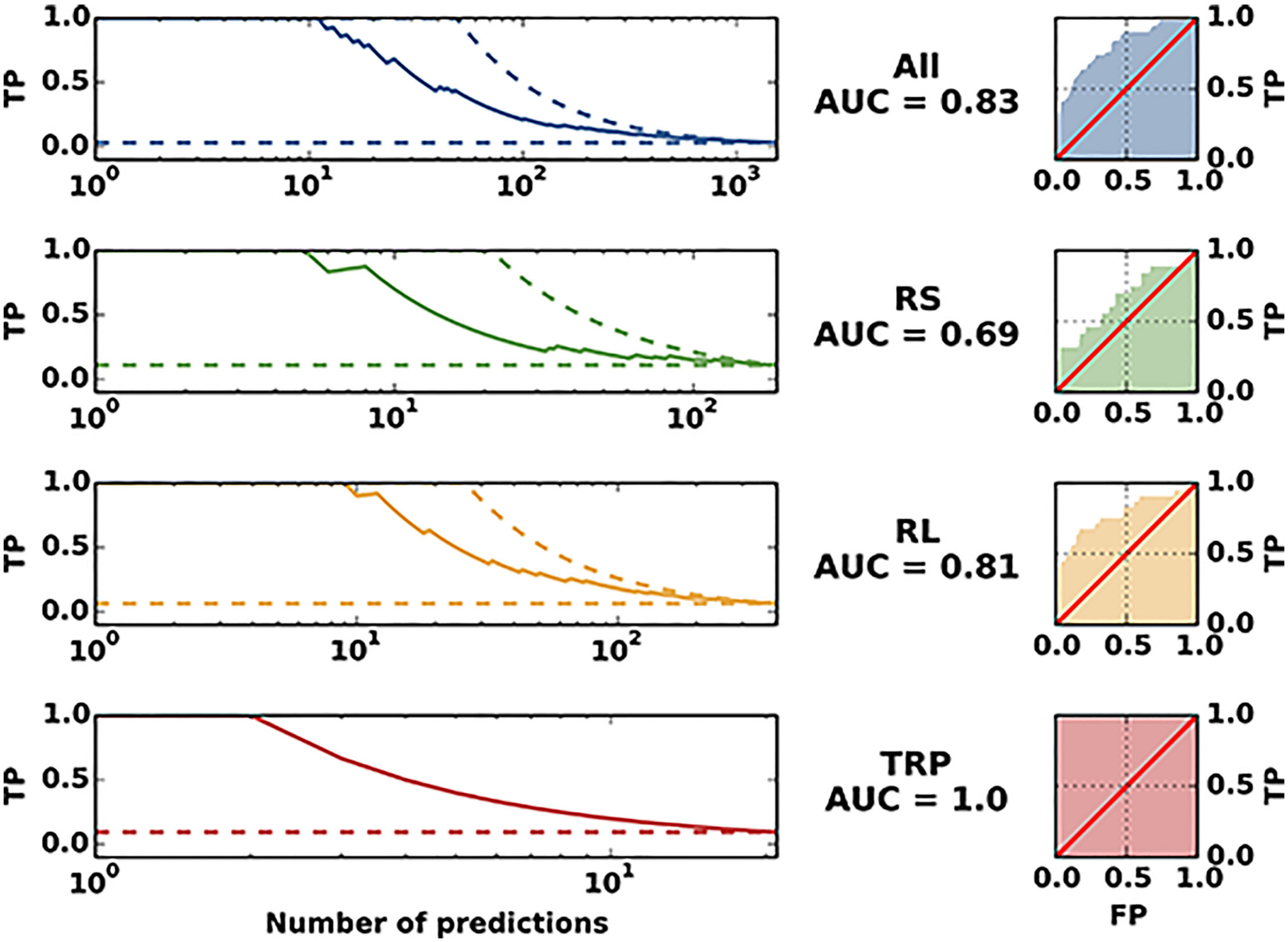

Fig 5. Performance Summary. The plots illustrate the performance in predicting protein interaction partners. The left panels show the fraction of true positives among the first *n* PPI predictions, with *n* being the number indicated on the horizontal axis (solid lines). The dashed lines show the best possible (upper dashed line) and the mean of a random prediction (lower dashed line). The right panels show ROC-curves, which indicate the dependence of the true-positive predictions (TP/P) from the false positive predictions (FP/N). The area under the curve (AUC) is a global global measure for the prediction quality; it is 1/2 for a random, and 1 for a perfect prediction. A protein pair is identified as an interacting (true positive) pair, if at least one PDB structure with at least one inter-protein contacts exists.

doi:10.1371/journal.pone.0149166.g005

the smaller MSA size is also the only one having a considerable false-positive rate in the intra-protein residue contact prediction. About 95% of the displayed 400 highest intra-protein residue contacts are actually contacts (see Fig A in S1 Text). Analogous considerations with a somewhat larger accuracy (85%) hold for LRU as displayed in Fig 4 panel F.

The PPI network of the tryptophan biosynthetic pathway

As a distinct test case for our methodology we analyzed the 7 enzymes (TrpA, B,C,D,E,F,G) that comprise the well characterized tryptophan biosynthesis pathway. In contrast to the ribosomal proteins, these enzymes are only conditionally essential in the absence of environmental tryptophan and their genes are only expressed under deplete tryptophan conditions. In this particular system, only two protein-protein interactions are known and resolved structurally:

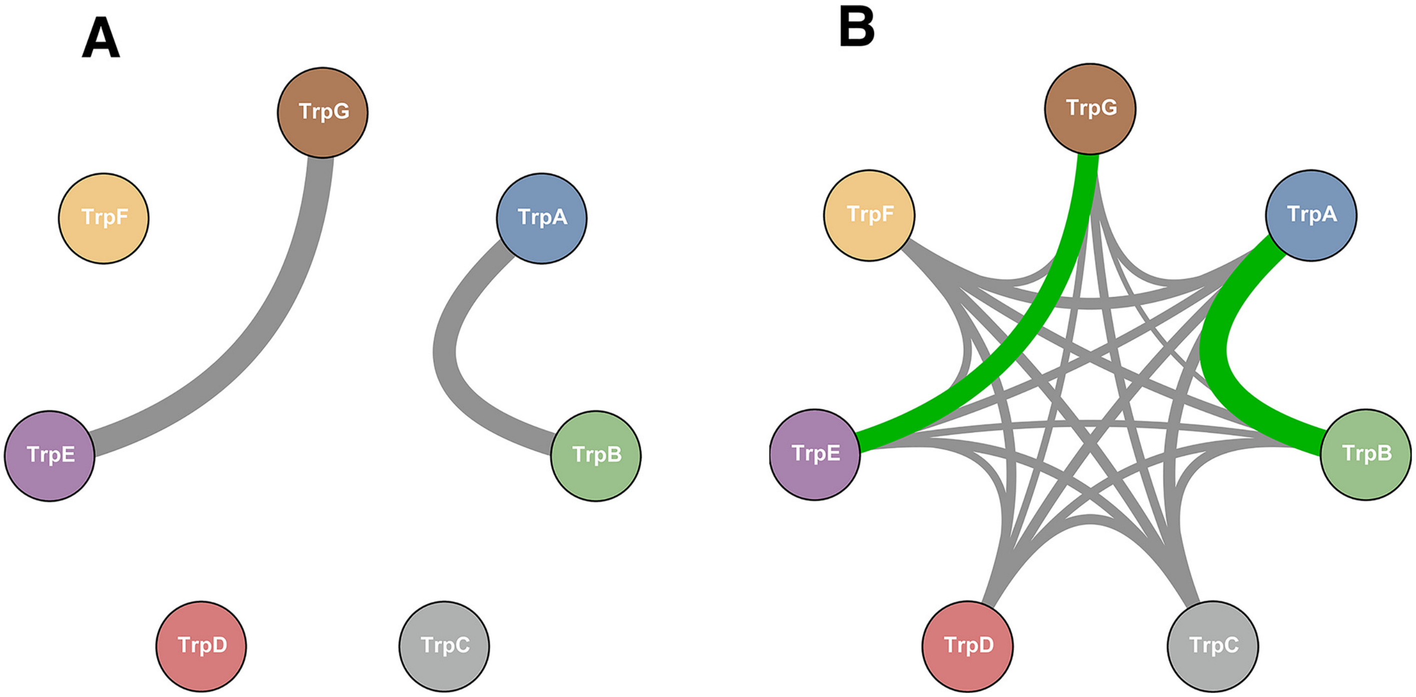

Fig 6. Tryptophan biosynthesis pathway. A Architecture of the known protein-protein interaction among the 7 enzymes which are coded in the Trp operon. The widths of the arcs are proportional to the number of inter-protein residues (which in this case is almost equal for the two interacting pairs). **B** Inferred PPI network, here the width of the arcs is proportional to the interaction score. Green arc correspond to the protein pairs for which a known structure exist.

doi:10.1371/journal.pone.0149166.g006

TrpA-TrpB (PDB-ID 1k7f [28]) and TrpG-TrpE (PDB-ID 1qdl [27]). Whereas the TrpG-TrpE pair catalyzes a single step in the pathway and their interaction is thus essential for correct functioning, the TrpA-TrpB pair catalyzes the last two steps in tryptophan biosynthesis. Both enzymes function in isolation but their interactions are known to increase substrate affinity and reaction velocity by up to two orders of magnitude. All other proteins catalyze individual reactions, but one might speculate that the efficiency of the pathway could benefit from co-localization of enzymes involved in subsequent reactions. Interestingly, the Pfam database [16] reports that in many species pairs of genes in the operon appear to be fused, suggesting that some of the fused pairs are actually PPI candidates. An example is the TrpCF protein, which is fused in *Escherichia coli* and related species (but not in the majority of species).

After applying our method to all 21 protein pairs we find elevated interaction scores only for TrpA-TrpB and TrpE-TrpG, which are the only known interacting pairs (see Fig 6 and Table J in S1 Text for the interaction scores of all pairs). Those two pairs have interaction scores of 0.375 and 0.295, while the other pairs are distributed between 0.071 and 0.167. Even though we do not define a significance threshold for prediction (see Section *A global view*), these two pairs would be discernible as interesting candidates even if we did not have the 3D structures.

We speculate therefore that the fusions in many species do not imply strong inter-protein co-evolution. To further investigate this aspect, we took a closer look at the protein pair TrpC-TrpF. For this protein pair, a high resolution structure of a fused version exists (PDB-ID 1pii [40]). We ran our algorithm on the complete multiple sequence alignment, the multiple sequence alignment with fused sequence pairs removed and only on the fused sequences. In none of these cases did we observe a statistically significant interaction score or a statistically significant prediction of inter-protein contacts present in the structure of the fused protein.

Our results are corroborated by the finding that all scores measuring the co-evolution between a ribosomal protein and an enzyme from the tryptophan synthesis pathway are small (see the following subsection). No indication for an interaction between the two systems is found, as to be expected from the disjoint functions of the two systems.

A global view

It is interesting to assemble a larger-scale system out of the three systems (SRU, LRU, Trp). To this end, we created all possible pairings between the proteins used in the present study (SRU vs. RU, SRU vs. Trp, LRU vs Trp, SRU vs SRU, LRU vs. LRU, and Trp vs. Trp). This leads to a total of 1540 pairs, out of which only 49 pairs are known to interact (which we defined as true positives). We present the findings in Fig 5 and in Figs G-I in S1 Text. Fig G in S1 Text shows the true-negative rate, which is the fraction of true negatives in the indicated number of predictions with the *lowest* interaction scores. As it can be seen our scoring produces a false negative just after 420 true negatives. Fig 5 and Fig H in S1 Text show true positive rates for the complete system and the individual systems. We also show true positive rates for alternative ways to calculate the interaction score between protein pairs, i.e. a different number of inter-protein residue-residue interaction scores to average. We notice that in the complete system, the performance is similar to the performance in individual systems. All of the 10 highest-scoring protein pairs are known to interact, and 75% of the first 20 protein-pairs. After these first 20 pairs, the true positive rate drops to around 45% in the first 40 predictions. This is analogous to the case of protein contact prediction, where methods based on the same model are able to extract a number of high confidence contacts but see a large drop in performance afterwards [15]. The area under the AUC for the whole system is 0.83 (see Fig 5). This is stable when averaging different numbers of residue-contact scores to arrive at a protein-protein interaction score, but the performance seems to worsen when using more than 6. This is probably because only a few inter-protein residue contacts have a large score and averaging over too many only adds noise. It can also be seen that averaging over 4 performs very well in the small ribosomal subunit, which is why we have chosen this value for the large part of the analysis. On the larger-scale system, though, any number between 1 and 6 performs almost identically.

A further question is whether it is possible to define a threshold allowing to reliably discriminate between interacting and non interacting pairs in terms of the interaction score. Fig I in S1 Text shows two normalized histograms of the interaction scores. The rightmost tail of the interacting pairs distribution is well separated from the rightmost tail of non-interacting one, but the highest scores of non-interacting pairs are strongly overlapping with the lowest scores of the interacting ones. The situation is therefore analogous to what is observed in the case of the inference of contacts within single protein families [14, 15, 17, 18], where the same technique is known to produce relatively few high confidence contacts in the topmost scoring residue pairs. To conclude, while high scores seem to reliably predict interacting pairs, and low scores non-interacting pairs, there is a large zone where clear discrimination between interacting and non-interacting pairs is not possible.

Limitations of the method

The main limitation of the method we present in this paper is arguably the concatenation of the two MSAs. In the test cases we analyzed this step was rendered difficult due to the presence of paralogs in the majority of species. The solution based on the minimization of the genomic distances within the concatenated MSA as outlined in subsection “Data extraction and matching for the ribosomal and trp operon proteins” can be problematic when used naively.

An example is the homo-dimerization of OmpR-class Response Regulators [14]. In this class of response regulators, DCA discovers a very clear homo-dimerization signal. However, the corresponding Pfam family PF00072 also contains a large number of response regulators not belonging to the OmpR class, functioning either in the monomeric form or in different dimeric structures. Using the full Pfam MSA, the OmpR-class homodimerization signal fades out due to the mixing of these different classes. The analysis as presented in this paper would therefore not capture this protein-protein interaction.

This suggests that more robust methods than genomic proximity should be developed. This is of course especially true if one is interested in eukaryiotic systems, were we have no evidence that genomic colocalization is an indication for interaction.

A second limitation of our method is the apparent inability to actually discern between interacting and non interacting pairs. Everything that the method can do at this point is to state that some interactions are more likely than others and show that this ordering is considerably better than a random one. This is a problem related to all DCA methods as reported for instance in [14, 15, 17, 18] in the context of inference of contact maps in single proteins. It is probable, though, that a cutoff can be chosen if more data becomes available and the method is applied to a data-set much larger in scale than the one presented here.

Conclusions

To conclude, we have shown that DCA performs well in the systems tested when used to predict protein-protein interaction partners. In the small and large ribosomal subunit our tests resulted in a true positive rate of 70% and 90% in the first 10 predictions (AUC of 0.69 and 0.81) while in the trp operon the two largest interaction scores corresponded to the only two interactions experimentally known (AUC 1). The performance is summarized in Fig 5. The Figure shows both the quality of the first predictions, but also a drop in performance after a fraction of all interacting pairs (about 40% in our test case). This is analogous to the case of protein contact prediction by DCA and related methods, where the performance drops after a limited number of high-confidence predictions [15]. In the same context and with the same caveat, a good performance in predicting inter-protein contacts on the residue level has been shown. The artificial data have shown that the performance of our approach depends crucially on the size of the alignments. Only for very large MSA ($M = 24{,}000$ sequences in our data) a perfect inference of the artificial PPI network was achieved. MSA for real proteins pairs are typically much smaller. Even for pairs of ribosomal proteins, which exist in all bacterial genomes, only about 1500-3200 sequence pairs could be recovered. This places these data towards the lower detection threshold of PPI. We therefore expect the performance of the presented approach to improve in the near future thanks to the ongoing sequencing efforts (the number of sequence entries in Uniprot [29] has been growing from about 10 millions in 2010 to 90 millions in early 2015) and improved inference schemes. The performance of the same algorithm on different and dissimilar systems suggests that the approach could be used to detect interactions experimentally unknown so far. In fact, if we trust our results on the trp operon we can already draw some speculative biological inferences. While there are many high-resolution structures of the ribosome available, one might have expected that in the trp operon there could be more transient previously unreported interactions in the tryptophan biosynthesis pathway beyond the two interactions that have been structurally characterized. As mentioned, various enzyme fusions can be observed in the databases, suggesting that there is an evolutionary benefit to co-localizing the enzymes of the pathway in the cell. An obvious benefit of such co-localization would be that the pathway intermediates do not have to diffuse throughout the cell from one enzyme component to the next. In the tryptophan biosynthesis pathway in particular, there are numerous phosphorylated intermediates that need to be protected from unspecific cellular phosphatase activity. Organizing the enzymes in the pathway in a multi-protein complex would seem like an efficient way to protect the intermediates from decay. However, our data indicate that the only statistically relevant co-evolutionary signals that can be observed are restricted to the known strong interactions between TrpA with TrpB and TrpE with TrpG. This could be interpreted in a number of ways: *(i)* The most obvious explanation is that there are no additional protein-protein interactions beyond those that are known and that no multi-enzyme complex exists for the tryptophan biosynthesis pathway. Alternatively *(ii)* it seems plausible that there are numerous structural solutions to form a tryptophan biosynthesis complex and that there is no dominant structure from which a co-evolutionary pattern can be observed in the sequence databases. Lastly *(iii)* it is not out of question that the enzymes of the pathway do not directly form a complex but that they are jointly interacting with an unidentified scaffold component. Of course we cannot exclude that our method is not able to capture other potentially present interactions.

From a methodological point of view, one possible algorithmic improvement is creating better MSAs for protein pairs. The vast majority of protein families show genomic amplification within species. This raises the issue of which sequence in one MSA should be matched with which sequence in the other MSA when concatenating the two MSAs, as shown in Fig 2. In the absence of prior knowledge and as long as only prokaryotes are concerned, we showed that it is possible to use the simple criterion of *matching by genomic proximity*. This criterion is based on the observation that two sequences are more likely to interact if they are genomically co-localized. Our results have shown that in the case of the ribosomal network better inference results can be obtained by using this matching criterion than by using a random matching or using a conservative matching taking only species with a single sequence in both MSAs into account, cf. Fig 7. However, we found it beneficial for the predictive performance to introduce

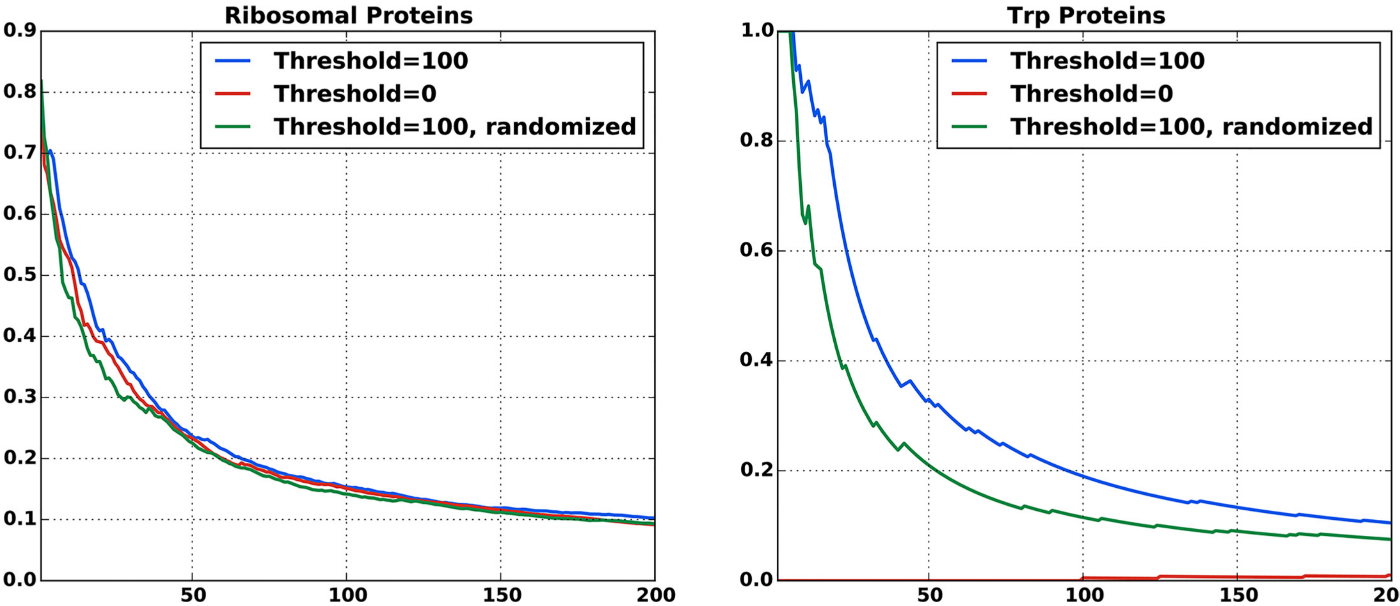

Fig 7. Efficacy of the different matching procedures. True-positive rates for inter-protein residue contact prediction for different matching procedures. Shown are means for all protein pairs that have at least 100 residue pairs in contact. The ribosomal and the trp proteins were tested independently. The red curves correspond to a matching including only protein sequences without paralogs inside the same species ("matching by uniqueness in genome"). The low performance of this approach on Trp proteins is due to a very low number of species without homologs, which leads to very small matched alignments. The blue curves show the results for our matching procedure as described in the text. The green curves correspond to alignments that have been obtained by first applying our matching procedure and then randomizing the matching within individual species. The definition of "contact" was the same as used above (a distance of less than 8.0Åbetween two heavy atom in the residues).

doi:10.1371/journal.pone.0149166.g007

a threshold distance above which we simply discarded candidate sequences. This is not based on biological principles.

We believe that our *naive* matching strategy can be improved substantially. Even if closeness of sequence pairs on the genome is a good proxy for interaction in some cases, for example if they belong to the same operon, excluding all distal pairs is a very crude criterion. This criterion is known to be erroneous in many cases, for example in the bacterial two component signal transduction system [24–26]. It would therefore be interesting to include the matching into the inference procedure itself, *e.g.* to find a matching that maximizes the inter-protein sequence covariation, cf. [24] for a related idea. However, for highly amplified protein families this leads to a computationally hard optimization task. Simple implementations get stuck in local minima and do not lead to improvements over the simple and straight-forward scheme proposed here.

Supporting Information

S1 Text. Supplementary Information Text. Multiple Sequence Alignments, Matching procedure, Inference Technique, Ribosomal Protein Interaction Partner Prediction, Artificial Data. (PDF)

Acknowledgments

CF and AP are supported from the EU Marie Curie Training Network NETADIS, (FP7 Grant 290038). HS was supported by grants GM106085 and GM019416 from the US National Institute of General Medical Sciences, National Institutes of Health. MW was supported by the Agence Nationale de la Recherche project COEVSTAT (ANR-13-BS04-0012-01). CF and AP acknowledge Riccardo Zecchina and Carlo Baldassi for many interesting discussions.

Author Contributions

Conceived and designed the experiments: CF HS MW AP. Performed the experiments: CF MW AP. Analyzed the data: CF MW AP. Contributed reagents/materials/analysis tools: CF MW AP. Wrote the paper: CF HS MW AP.

References

1. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Poc Natl Acad Sci. 2001; 98(8):4569–4574. doi: 10.1073/pnas.061034498
2. Ho Y, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 2002; 415(6868):180–183. doi: 10.1038/415180a PMID: 11805837
3. Braun P, et al. An experimentally derived confidence score for binary protein-protein interactions. Nature methods. 2008; 6(1):91–97. doi: 10.1038/nmeth.1281 PMID: 19060903
4. Harrington ED, Jensen LJ, Bork P. Predicting biological networks from genomic data. FEBS letters. 2008; 582(8):1251–1258. doi: 10.1016/j.febslet.2008.02.033 PMID: 18294967
5. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends in biochemical sciences. 1998; 23(9):324–328. doi: 10.1016/S0968-0004(98)01274-2 PMID: 9787636
6. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nature biotechnology. 2000; 18(6):609–613. doi: 10.1038/76443 PMID: 10835597
7. Marcotte CJV, Marcotte EM. Predicting functional linkages from gene fusions with confidence. Applied bioinformatics. 2002; 1(2):93–100. PMID: 15130848
8. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Poc Natl Acad Sci. 1999; 96(8):4285–4288. doi: 10.1073/pnas.96.8.4285

9. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Current Opinion in Structural Biology. 2002; 12(3):368–373. Available from: http://www.sciencedirect.com/science/article/pii/S0959440X02003330. doi: 10.1016/S0959-440X(02)00333-0 PMID: 12127457

10. Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins: Structure, Function, and Bioinformatics. 2002; 47(2):219–227. doi: 10.1002/prot.10074

11. Yeang CH, Haussler D. Detecting Coevolution in and among Protein Domains. PLoS Comput Biol. 2007 11; 3(11):e211. doi: 10.1371/journal.pcbi.0030211 PMID: 17983264

12. Juan D, Pazos F, Valencia A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Poc Natl Acad Sci. 2008; 105(3):934–939. doi: 10.1073/pnas.0709671105

13. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nature Reviews Genetics. 2013;. doi: 10.1038/nrg3414 PMID: 23458856

14. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Poc Natl Acad Sci. 2009; 106(1):67–72. doi: 10.1073/pnas.0805923106

15. 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. Poc Natl Acad Sci. 2011; 108 (49):E1293–E1301. doi: 10.1073/pnas.1111471108

16. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Research. 2014; 42(D1):D222–D230. doi: 10.1093/nar/gkt1223 PMID: 24288371

17. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707. doi: 10.1103/PhysRevE.87.012707

18. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, et al. Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. PLoS ONE. 2014; 9(3):e92721. doi: 10.1371/journal.pone.0092721 PMID: 24663061

19. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. Proteins: Struct, Funct, Bioinf. 2011; 79:1061. doi: 10.1002/prot.22934

20. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife. 2014; 3. doi: 10.7554/eLife.02030 PMID: 24842992

21. 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. doi: 10.7554/eLife.03430 PMID: 25255213

22. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-resolution protein complexes from integrating genomic information with molecular simulation. Poc Natl Acad Sci. 2009; 106(52):22124–22129. doi: 10.1073/pnas.0912100106

23. Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Poc Natl Acad Sci. 2012; 109(26):E1733–E1742. doi: 10.1073/pnas.1201301109

24. Burger L, Van Nimwegen E. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. Molecular Systems Biology. 2008; 4(165):165. doi: 10.1038/msb4100203 PMID: 18277381

25. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. PloS one. 2011; 6(5):e19729. doi: 10.1371/journal.pone.0019729 PMID: 21573011

26. Cheng RR, Morcos F, Levine H, Onuchic JN. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Poc Natl Acad Sci. 2014; 111(5):E563–E571. doi: 10.1073/pnas.1323734111

27. Knöchel T, Ivens A, Hester G, Gonzalez A, Bauerle R, Wilmanns M, et al. The crystal structure of anthranilate synthase from Sulfolobus solfataricus: Functional implications. Proceedings of the National Academy of Sciences. 1999; 96(17):9479–9484. Available from: http://www.pnas.org/content/96/17/9479.abstract. doi: 10.1073/pnas.96.17.9479

28. Weyand M, Schlichting I, Marabotti A, Mozzarelli A. Crystal structures of a new class of allosteric effectors complexed to tryptophan synthase. Journal of Biological Chemistry. 2002; 277(12):10647–10652. doi: 10.1074/jbc.M111285200 PMID: 11756456

29. Consortium TU. UniProt: a hub for protein information. Nucleic Acids Research. 2015; 43(D1):D204–D212. Available from: http://nar.oxfordjournals.org/content/43/D1/D204.abstract. doi: 10.1093/nar/gku989 PMID: 25348405

30. Feinauer C, Skwark MJ, Pagnani A, Aurell E. Improving Contact Prediction along Three Dimensions. PLoS Comput Biol. 2014 10; 10:e1003847. doi: 10.1371/journal.pcbi.1003847 PMID: 25299132

31. Katoh K, Misawa K, Kuma Ki, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002; 30(14):3059–3066. doi: 10.1093/nar/gkf436 PMID: 12136088

32. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011;p. gkr367. doi: 10.1093/nar/gkr367

33. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS computational biology. 2011 June; 7(6):e1002073. Available from: http://europepmc.org/articles/PMC3111532. doi: 10.1371/journal.pcbi.1002073 PMID: 21695233

34. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLoS Comput Biol. 2012 05; 8(5):e1002514. doi: 10.1371/journal.pcbi.1002514 PMID: 22615551

35. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences. 1999; 96(6):2896–2901. Available from: http://www.pnas.org/content/96/6/2896.abstract. doi: 10.1073/pnas.96.6.2896

36. Lathe WC, Snel B, Bork P. Gene context conservation of a higher order than operons. Trends in biochemical sciences. 2000; 25(10):474–479. doi: 10.1016/S0968-0004(00)01663-7 PMID: 11050428

37. Rogozin IB, Makarova KS, Wolf YI, Koonin EV. Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. Briefings in Bioinformatics. 2004; 5(2):131–149. Available from: http://bib.oxfordjournals.org/content/5/2/131.abstract. doi: 10.1093/bib/5.2.131 PMID: 15260894

38. Wlodawer A, Walter J, Huber R, Sjölin L. Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and X-ray refinement of crystal form II. Journal of Molecular Biology. 1984; 180(2):301–329. doi: 10.1016/S0022-2836(84)80006-6 PMID: 6210373

39. Borovinskaya MA, et al. Structural basis for aminoglycoside inhibition of bacterial ribosome recycling. Nature Struct Mol Biol. 2007; 14(8):727–732. doi: 10.1038/nsmb1271

40. Wilmanns M, Priestle JP, Niermann T, Jansonius JN. Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: Indoleglycerolphosphate synthase from Escherichia coli refined at 2.0 Å resolution. Journal of Molecular Biology. 1992; 223(2):477–507. Available from: http://www.sciencedirect.com/science/article/pii/0022283692906657. doi: 10.1016/0022-2836(92)90665-7 PMID: 1738159