Identification of protein functional regions

Francesca Nerattini,[1] Matteo Figliuzzi,[2] Chiara Cardelli,[1] Luca Tubiana,[1] Valentino Bianco,[1] Christoph Dellago,[1] and Ivan Coluzza[3]

[1]*Faculty of Physics, University of Vienna, Boltzmanngasse 5, 1090 Vienna, Austria*
[2]*Sorbonne Universites, UPMC, Institut de Biologie Paris-Seine, CNRS, Laboratoire de Biologie Computationnelle et Quantitative UMR 7238, Paris, France*
[3]*CIC biomaGUNE, Paseo Miramon 182, 20014 San Sebastian, Spain, and IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain.**

Protein sequence stores the information relative to both functionality and stability, thus making it difficult to disentangle the two contributions. However, the identification of critical residues for function and stability has important implications for the mapping of the proteome interactions, as well as for many pharmaceutical applications, e.g. the identification of ligand binding regions for targeted pharmaceutical protein design. In this work, we propose a computational method to identify critical residues for protein functionality and stability and to further categorise them in strictly functional, structural and intermediate. We evaluate single site conservation and use Direct Coupling Analysis (DCA) to identify co-evolved residues both in natural and artificial evolution processes. We reproduce artificial evolution using protein design and base our approach on the hypothesis that artificial evolution in the absence of any functional constraint would exclusively lead to site conservation and co-evolution events of the structural type. Conversely, natural evolution intrinsically embeds both functional and structural information. By comparing the lists of conserved and co-evolved residues, outcomes of the analysis on natural and artificial evolution, we identify the functional residues without the need of any a priori knowledge of the biological role of the analysed protein.

Introduction

The combination of the residues along the protein sequence allow for the stability of the structure and for the biological functionality.

More precisely, it has been shown that for a protein to function, e.g. for the catalytic activity of enzymes, the sequence must contain explicit interacting spots and, simultaneously, show a balance between structural stability and flexibility. Despite the intuition might lead to concluding that structure and function evolved independently and, as a consequence, each residue can be classified as strictly functional or strictly structural, several studies showed that there is an overlap between the two categories and the interdependence is essential for the protein activity [1, 2].

A detailed characterisation of protein structural and functional residues is essential for the advancement in proteome mapping, protein engineering, as well as for the developing of new pharmaceutical applications based on targeted protein design [3, 4, 5, 6, 7].

The experimental identification of the residues directly involved in the biological process is expensive and time consuming. It requires large scale mutation assays for high-throughput screening [8, 9], while *in-silico* screening has a much lower cost.

Most computational methods [10, 11, 12, 13, 14, 15, 16, 17] analyse the large amount of protein sequence evolution data, searching for conservation or co-evolution patterns. The significance of amino acids co-evolution is based on the hypothesis that mutations of interacting residues are correlated. Hence, despite single point mutations might not conserve protein stability, multiple alterations must occur simultaneously among interacting residues [18, 19]. Co-evolution events could involve residues that are crucial for the protein activity (e.g. catalytic site residues), for the stability of the native structure (e.g. hydrophobic core residues) or, in some cases, for both. In other words, two functional residues participating in the ligand binding site of an enzyme must co-evolve to keep the efficiency of the catalytic reaction high, while two structural residues in the core of a protein or at the interface between two binding proteins cannot evolve independently without negatively affecting the protein stability or the binding affinity. Residue conservation is can be a straight forward analysis using the method of Casari et al. [20] based on a principle component analysis (PCA) of the sequence alignments. On the other hand Direct Coupling Analysis (DCA) is one of the most promising [15, 16, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37] tool to infer the direct correlations between the residues of a protein domain that arise from either functional (e.g. catalytic site residues) or structural (e.g. hydrophobic core residues) constraints optimised during natural evolution. Currently, DCA correlation maps have been used to infer the structure of a single protein or even the binding regions of protein pairs [23, 32]. In other words, the correlation information has been employed mainly to identify the structural amino acids and predict the folded structure. However, from DCA it is still not possible to a priori distinguish between structural and functional residues. This is because the two type of co-evolution events give DCA correlation signals of the same kind. Conservation and co-evolution

have separately been successfully used to identify many protein structural and functional properties but never combined int a single analysis scheme.

In this work we propose an approach combining conservation and co-evolution analysis to identify residues that are key for stability and functionality, and to further distinguish between either strictly functional (F)/structural (S) ones from the one that are involved in both (OFSR, overlapping-functional-structural-residue, according to the naming scheme adopted in Ref. [1]). By strictly structural we mean that such residues are responsible of the protein stability regardless of the set of functions. Clearly, destabilising the folded structure could have devastating effect on the protein function. On the other hand the strictly functional residues can be altered without affecting the folding capability of the protein.

The methodology that we propose is based on the hypothesis that an artificial evolution process aimed at optimising the amino acids sequence of a specific target backbone conformation, in the absence of any functional constraints, would lead exclusively to co-evolution events of the structural type.

A possible way to construct artificial evolution pathways with such characteristics is the inverse protein folding, better known as protein design [7, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]. Protein design consists in identifying sequences specific for a given backbone structure that at equilibrium would spontaneously adopt the target conformation. Generally, many different sequences can be identified for the same target in the same ways as there are thousands of sequence fragments that fold into the same domain.

Here we consider three testing case domains, we evaluate the site entropy (the single site conservation), and perform a DCA analysis to calculate pairwise couplings (a measure of the correlation due to co-evolution) on both the artificially evolved pool of sequences generated with protein design and on natural sequences found with standard protein Multiple Sequence Alignment (MSA) methods.

Strong signals in the artificial conservation and correlation analysis corresponds to residues that are key for the structural stability, while strong signals in the natural sequences analysis might have both structural and/or functional role. We classify as S the set of residues that, combining the information from both conservation and co-evolution, have a strong signal in the artificial pool and appreciably lower in the natural one, F the residues that, vice versa, have high natural and appreciably lower artificial signal and OFSR the residues which posses high and comparable signals in both natural and artificial sequences.

As testing cases we chose three domains belonging to well studied and well known protein families, namely the PDZ, FKBP and Response_reg. PDZ [49, 50, 51, 52, 53, 54] domains (to which has been previously referred to as DHR or GLGF domains) are present in signalling proteins and regulates many cellular pathways. FKBPs [55, 56, 57] have peptidyl prolyl cis-trans isomerase activity and the smallest member of the family, FKBP12, is the target of an immunosuppressant molecule, FK506, effectively used after transplant operation and to treat patients suffering from autoimmune disorders. Finally, Response_reg [58, 59, 60] are proteins that regulate the cell's functioning according to environmental changes. The choice of the families was based mainly on two fundamental parameters: all selected families have a large number of members, and reliable experimental data is available on the functional residues.

Methods

Protein Model

We employ the Caterpillar coarse-grained protein model that reduces the complexity of the amino acids to the backbone atoms: C, O, C_α, N, H. Although the full Hamiltonian of the Caterpillar model, as described in Ref. [61, 62] (see SI for details), explicitly depends on the specific orientation of the backbone atoms, in the design procedure we only consider the energy terms that are directly affected by the amino acid identities, since the protein conformation is not varied during the simulation.

Design

Protein design consists in the exploration of the vast sequence space, searching for the ensemble of sequences that would spontaneously fold into a specific target structure. Our main assumption is that artificially designed sequences will not show on average any biological function except of folding into the target structure. This assumption rests on the fact that the artificial sequences generated via our design algorithm have little overlap with natural sequences. In fact, a blast alignment did not find any scoring. Hence, if there was any functional region in the design it would have resulted in a blast scoring alignment. The second argument is that without any external selection pressure for a specific function, different "accidental" functional would average out and be indistinguishable from the noise. On the other the structural residues will leave a measurable signal since all the artificial sequences fold into the target. We artificially design three single domain proteins, chosen as representative of PDZ domain, FKBP-type peptidyl prolyl cis-trans isomerase, and Response_reg. We selected the structures corresponding to the PDB IDs: 1WI2, 2PPN [63] and 1NXW [58] respectively. It is important to notice that we opted for the crystal structure 1NXW, where the Response_reg is in complex with an acetate molecule, rather

than the native 1NXO, because of the disordered nature of some residues in the latter structure [58].

Firstly, we relax the crystallographic structure to the caterpillar protein model by discarding the side chain atoms. As fully described in Ref.s [61, 62], the Caterpillar model, in combination with Virtual Move Parallel Tempering (VMPT), that includes adaptive umbrella sampling technique [64, 65], is capable of producing a large number of sequences that should fold into realistic protein target structures (see SI for details). We extensively verified the latter hypothesis in our previous studies (e.g. Figure 4 of [61]) but also from the Random Energy model [39, 66, 67] predicting that two sequences with the same energy on the target structure are equivalent solutions to the folding problem. Hence, to guarantee foldability, we considered sequences with equivalent energy in the target structure. It is important to stress the sequences might have different folding rates, but this is true also for the natural ones.

Therefore, we employ VMPT performing Monte Carlo (MC) Parallel Tempering simulations, running in parallel 16 replicas of the system differing in temperature: T=(10.000; 5.000; 2.000; 1.000; 0.500; 0.333; 0.250; 0.200; 0.167; 0.143; 0.125; 0.111; 0.100; 0.091; 0.083; 0.077) in units of K_B and attempting temperature swaps between adjacent replicas. Additionally, for each temperature, we recover statistics from all the replicas, employing the virtual move scheme of Ref. [64].

It is important to stress that we want to generate sequences that have pair correlations induced by the target structure. Hence, the sequences generated with our methods must fold computationally but are not required to fold experimentally. The caterpillar model fulfils such a requirement [61, 62]. This property means that also that any model or force field capable of generating sequences and refold them into the natural backbone structure would be usable for the purpose of our analysis. The first evidence to support our claim comes from our previous work on heteropolymer design including the Caterpillar design Our work on design showed that provided that a heteropolymer chain is designable (we defined the rules to identify such property) then the 3D structures can be designed with high accuracy independently of the interaction matrix used to define the amino acid interactions [68, 69]. In fact, the same design strategy works for lattice and off-lattice proteins with implicit or explicit solvent, plus the above mentioned patchy polymers [68, 69, 70, 71]. This result is the first indications that the key correlations that determine the folding do not depend on the particular model used to represent the residue interactions. The only requirement, of course, is that the protein structural space is correctly represented and for that, we can bring not only the evidence produced by the Caterpillar model it-self but also made with its close cousins: the tube model of Maritan et al. [72, 73, 74, 75] and the CamTube model [76]. The caterpillar refolding resolution is between 2 and 3 ÅRoot mean square displacement. Hence, the design is not sensitive to limits in the experimental resolution below 3Åwhich for X-ray structure prediction is feasible. The second indication that prediction does not depend on the particular choice of the amino-amino acid interactions (provided that are heterogeneous for instance according to a Gaussian distribution [68, 69]) is given by the DCA it-self. In fact, the direct couplings are nothing else then specific residue-residue interactions that reproduce the correlations functions measured over the evolution process. Hence, there are several sets of direct couplings that can give the same correlation pattern and lead to the same folding. The remarkable structure prediction power of the DCA methodology indicates that optimised couplings are enough to connect sequence to structure.

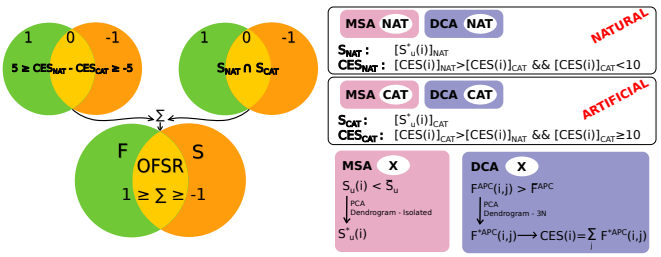


FIG. 1: Classification of functional F, structural S and overlapping functional-structural residues OFSR. Firstly, we calculate single site entropy $S_u(i)$ (red box) from MSA and co-evolution $F^{APC}(i,j)$ (blue box) from DCA analysis. We prune for highly conserved or strongly coupled residues. Therefore we select signals with entropy lower than the average one, $S_u(i) < \overline{S}_u$, and coupling strength higher than the average, $F^{APC}(i,j) > \overline{F}^{APC}$. To refine the selection and isolate the outliers, we perform a Principal Component Analysis (PCA) and construct a dendrogram on the Euclidean distance in the eigenvector space. For $S_u(i)$ we select the signals corresponding to points not belonging to the largest cluster, $S^*_u(i)$. Given the increasing complexity of the dendrogram relative to $F^{APC}(i,j)$, we select for the $3N$ signals corresponding to higher values in the dendrogram, $F^{*APC}(i,j)$. Moreover, we reduce the pair signal $F^{*APC}(i,j)$ to single site signal, by counting the Co-Evolution Signals per residue: $CES(i) = \sum_j F^{*APC}(i,j)$. We perform it on both natural and artificial families (right upper panels), and we identify $S_{NAT/CAT}$ and $CES_{NAT/CAT}$ as relevant signals (according to the selection in the relative panels). We separately analyse conservation and co-evolution, and subtract the artificial information from the natural one, according to the idea that artificial evolution does not select for function, while natural for both function and structure. We assign a score and sum up the outcome of the two analysis (on S_u and CES) and use it to categorise each residue as S (orange, if the natural character prevails on the artificial one), F (green, vice-versa) and $OFSR$ (yellow, if the natural and artificial signals are comparable).

DCA couplings and site entropy

DCAs are global statistical models for large MSA of evolutionarily related protein sequences. DCA reproduces the statistics of correlated amino acid mutations and allows to isolate the direct correlation existing between two positions in a protein sequence, ignoring the effects arising from all other positions. Single site entropy, instead, can be derived by the empirical frequencies from MSA.

In the present work, we are interested in deriving a measure of the coupling between amino acids in different positions and single site conservation. For each system, we both evaluate single site entropy and perform DCA analysis on the set of caterpillar designed sequences (containing M^{cat} sequences) and on the natural sequences (with M^{nat} entries). For DCA, we use the open source software available on the platform http://dca.rice.edu/portal/dca/home [27, 32]. The analysis on natural sequences will intrinsically capture both structural and functional constraints conserved across the families of homologous proteins, while artificial sequences reflect structural constraints relevant in folding thermodynamics only.

We compute site entropies $S_u(i)$ (see Eq. **??** in the SI) and the Average-Product Correction (APC) F_{ij}^{APC} (see Eq. **??** in the SI) from the DCA modelling to score sites according to their conservation and co-evolution. Comparing the signal from natural sequences and caterpillar sequences, we should be able to spot functional signatures. It is important to stress that the APC product does remove phylogenetic correlations only if they are present within the sets of sequences present. Phylogenetic correlations are not necessarily present also in the set of natural sequences.

Results

We consider three protein domains, and for each of them we select a pool of natural sequences and a single domain representative protein structure: PDZ domain, 38522 sequences from the Pfam [10] protein family PF00595, structure from PDB 1WI2; FKBP domain, 19610 sequences corresponding to the family PF00254, structure PDB 2PPN; and Response_reg domain, 1000 sequences taken from a BLAST [77] alignment on the sequence of the protein 1NXW from PDB (in order to reduce the enormous variability of functions and biological pathways in which Response_reg are involved) also taken as a reference for the structure. To identify residues essential for functionality and stability and further categorise them as S, F or OFSR, we need to generate an artificial protein family that, by construction, contains only structural information. To this end, we employ a protein design procedure based on the caterpillar protein model [61, 62], which has been tested on several natural protein structures, producing a large number of different sequences capable of folding into the target conformation. Each target PDB is firstly adapted to the coarse-grained caterpillar representation, removing the side chains. The protein design method consists in exploring the sequence space with a point mutation, and swap moves along the sequence with a frozen target backbone in an MC simulation. We enhance the sampling with the Virtual Move Parallel Tempering (VMPT) scheme. We denote the $\sim 10^5$ most probable sequences (best candidates for the folding) as an artificial family. Such sequences possess low energy and high variability of amino acids. We then perform the DCA analysis (blue box sketched in Fig. 1) on natural (NAT) and artificial (CAT) sequences, as well as evaluate single site entropy from MSA (red box of Fig. 1). We identify two ensembles: the most conserved residues $S_{NAT/CAT}$ and the strongest Co-Evolution Signals $CES_{NAT/CAT}$ (upper panels). For further details, see *Strong signal selection* appendix. The DCA is particularly useful in inferring the contact map of the native family structure (see Fig. 3a and in the SI Fig. **??**). When applied to the artificial families, one might expect that the predicted contact maps from DCA would be much more precise and dense of strong signals, compared to the natural ones. This expectation originates from the hypothesis that the design algorithm highly optimises the sequences for the target structure. However, the solution space of folding sequences is large and heterogeneous a number of contacts similar to the native predictions have strong DCA signals (compare Fig. 3a,b and in the SI Fig. **??**). The presence of such variability is non-trivial because it indicates that during the design optimisation the fluctuations of the residues resembles the one observed in the natural evolution. Comparing the contact maps obtained from natural and artificial families, we observe a similar distribution of data indicating that the design sequences have natural correlation patterns that are inherent in the target structure. Moreover, the differences in the predicted contacts are the first direct evidence of the different roles of the residues (i.e. S, F or OFSR).

To more accurately compare the natural and artificial ensembles, we consider site conservation ($S_{NAT/CAT}$) and co-evolution ($CES_{NAT/CAT}$) separately and assign a score to each of the identified residues (for $S_{NAT/CAT}$ highlighted in Figs. 7 and 8). Since natural signals bear functional and structural information, while artificial ones encode only for structure, we compare the residues conserved in the artificial evolution with the naturally conserved ones. Therefore, the complement of S_{NAT} encode for function only (green, score 1), the one of S_{CAT}, vice-versa, only for structure (orange, score -1) and the intersection $S_{NAT} \bigcap S_{CAT}$ has function-structure overlapping character (yellow, score 0). Similarly, we compare the co-evolution signals of each residue in the nat-

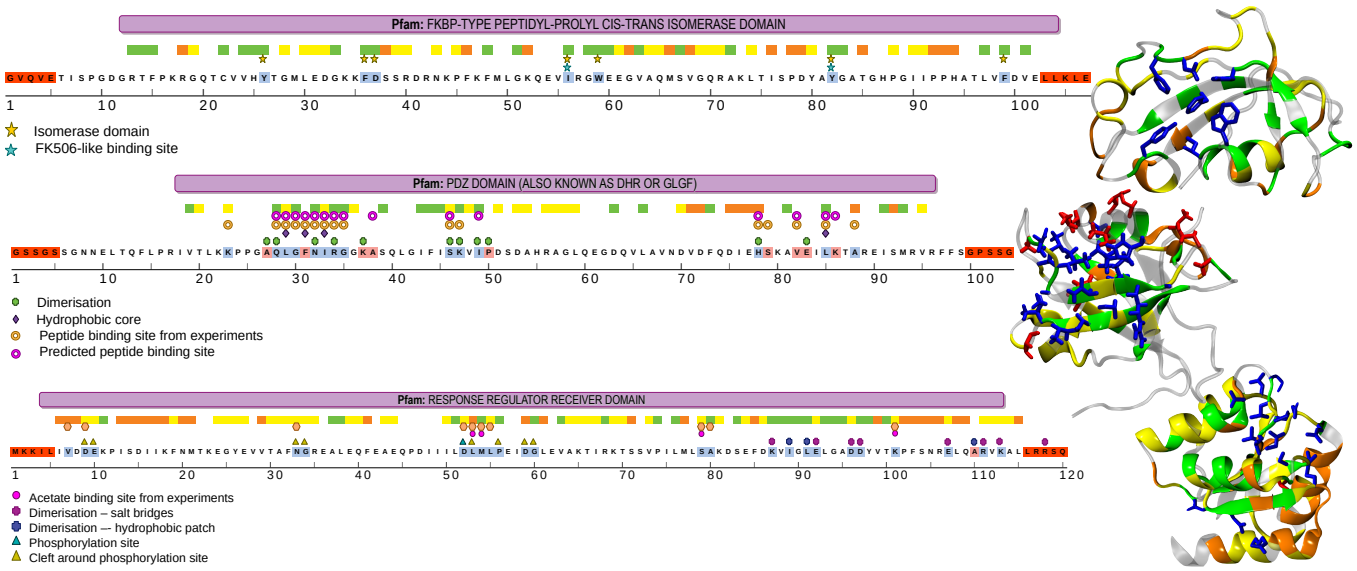


FIG. 2: Identification of critical residues (green, yellow and orange) for protein function and structure. The sequences are relative to the PDBs used as representative of each analysed family, namely 2ppn for FKBP (top), 1wi2 for PDZ (centre) and 1nxw for Response_reg (bottom), visualised on the right-hand side. The purple bar specifies the location of the domain. We have excluded the first and last five residues from our analysis, as they are often a source of the noise. We further classify the identified residues in three classes. Functional residues F, in green, have a natural signal that prevails on the artificial one. Vice-versa, structural residues S, in orange, are highly conserved and co-evolved in the artificial evolution, but poorly in the natural ones. Residues with a comparable signal between natural and artificial analysis are classified as overlapping-functional-structural-residues OFSR and visualised in yellow. At this point we want to focus only on the F and the S residues because of their strong signal compared to the one of the OFSR. The latter, can be refined by improving the definition of the cluster algorithm [78]. We use the information found in literature (both experimental and from predicting software) to mark the residues important for functional processes and/or for structural stability [51, 52, 53, 54, 58, 59, 60, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89]. Given that information, we highlight in blue the residues that show a match with our prediction, while in red the ones identified as key by literature but missing in our analysis. Over the three investigated families, our prediction has an agreement of 100%, 61% and 96% (top to bottom) in the identification of essential residues for functionality and structural stability.

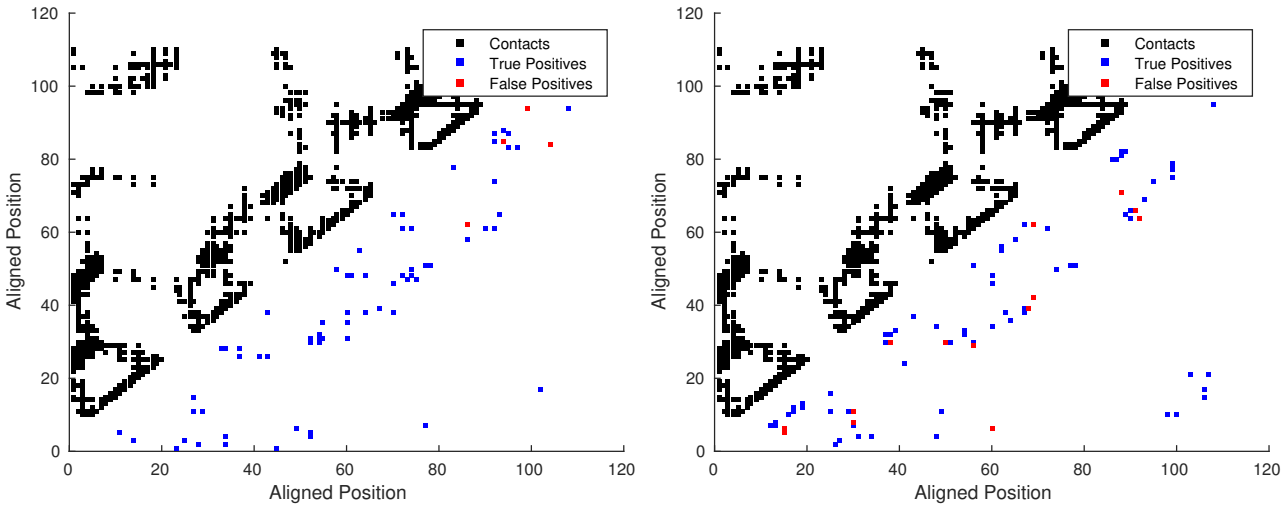


FIG. 3: Contact maps from crystallographic structure of the Response regulator and predicted from a) $F_{NAT}^{APC}(i,j)$ natural correlations and b) $F_{CAT}^{APC}(i,j)$ artificial correlations: a pair is in contact if at distance ≤ 8 Å in the crystal (black points) and is predicted to be in contact if the $F_{NAT}^{APC}(i,j)$ correlation is among the 100 strongest ones and at a distance on the chain larger than 5^{th} neighbours (coloured points). Blue squares are *true positives*, i.e. contacts both predicted and present in the crystal structure, while red ones are *false positives*, that is predicted contacts not present in the crystal.

ural and artificial evolution, and we classify as purely functional the residues with the natural signal prevail-

ing on the artificial one, function-structure overlapping if they are comparable and purely structural if the artificial signal prevails (see Fig. 1). The raw data relative to such analysis are listed in Fig. 6 of SI, and is evident from such figure that the majority of signals come from the co-evolution data, thus stating the importance of including both site conservation and co-evolution events in the evolutionary related statistical methods.

Finally, for each residue, we sum the scores separately obtained in the site conservation and co-evolution. Therefore, we classify as purely functional (F) residues with score$\geq$ 1, purely structural (S) those with score≤ -1 and overlapping (OFSR) if score$= 0$.

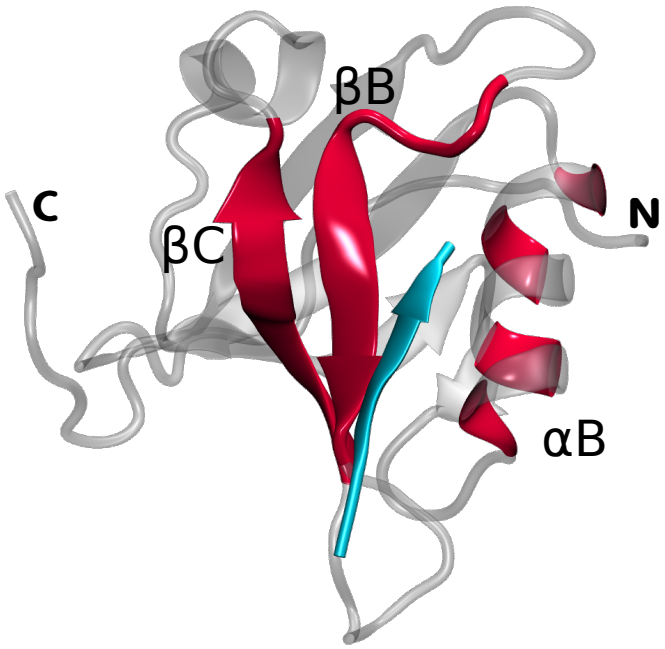


FIG. 4: PDZ domain complexed with a peptide ligand, (PDB 1tp3). In cyan is represented the ligand; in red, residues 28-35, 42-49 (both identified by the analysis as function related patches along the sequence) and 78,81,85,88 (identified as key for structure and function). We correctly identify the regions involved in the PDZ main function, namely βB, βC and we partially recover the αB stabilisation effect through our analysis.

In Fig. 2 we list the residues as functional (green), overlapping (yellow) or structural (orange) (FOS annotation) according to our analysis. We compare the outcome of our analysis with the experimental and computational literature available and encode the matching in colour assigned to each residue along the sequence: blue positions are identified as relevant for protein function and/or stability both in nature and in our prediction, while red ones are key in literature but not predicted.

Our results show that for the FKPB family we correctly identify the 100% of residues previously assigned as key for the isomerase activity and the binding of FK506 related drugs [79, 80, 81, 82]: residues=26; 36; 37; 56; 59; 82; 99. In particular, from our study, we categorise residues=37; 59; 82; 99 as strictly functional (green), matching what has been previously found in numerous computational and experimental studies. The remaining listed functional residues, based on their location in the folded structure, arrange around the binding pocket (see right-hand side of Fig. 2). Hence, the strictly functional are 60% of the 7 experimentally identified functional residues. Overall the residue annotated on the natural sequence where 71% of the total protein length giving little information above a random identification. It is the comparison with the artificial set proposed here that leads to the further FOS annotation that point the attention to key residues for further study.

PDZ domains usually bind to the C-terminus of other proteins, or too small peptides, via beta sheet augmentation, therefore establishing inter-protein hydrogen bonds and interactions that extend the PDZ beta sheet (involving βB and βC) [51, 52]. The binding partner further packs against the PDZ C-terminal alpha-helix (αB).

In the present study, we correctly identify the majority of the residues involved in such protein-protein interaction (see Fig.). Remarkably, we identify two central functional regions at positions 28 to 35 and 43 to 49, that are the regions coinciding with βB and βC and mainly involved in peptide binding. We partially recover also the further stabilisation of the peptide binding involving αB, since we identify the residues=78,81,85,88 as key.

On the overall, we correctly identify residues=23, 28, 29, 30, 32, 33, 34, 35, 46, 47, 49, 78, 85, 88, that is the 61% of the residues known from experiments and predicted by I-Tasser to be involved in either dimerisation, peptide binding or hydrophobic core [49, 51, 52, 53, 54, 86, 87, 88]. It is interesting to notice that we categorise as strictly structural only two of the residues=78,88, as mentioned earlier.

Particularly interesting are the residues=23,29,33,46 categorised as OFSR, that is a mutation that might influence one of the activity as mentioned above. Our results suggest that it would be interesting to test if such residues are necessary to stabilise the protein and the effect of a mutation is to destabilise the protein structure, which in turns reduces the protein efficiency.

It is important to stress that the missing 39% of experimentally isolated residues did not give a strong DCA signal from the natural alignment. Hence, it is not a limitation of our methodology but of the alignment data. Moreover, our lower resolution limit is comparable to the one of other methods [90]. To improve upon this point, we could try to include also the indirect couplings that Cheng et al. [13] have demonstrated to be strongly connected to functionality. In particular, we could include higher order correlation than just pairs like in the DCA. An other approach would be to try to divide the natural sequences into phylogenic groups and test for the appearance of novel correlations.

Additionally, if dimer structures were available or speculated, one could further refine the classification, identifying among the functional residues the subgroup of sites involved in protein-protein interactions (e.g. dimerisation or small-peptide binding). One could produce artificial families through our design algorithm using as a target the structure of a dimer or the protein bound to the peptide instead of the single protein one. The sites that would change their classification from functional to structural would be the one that is essential to stabilise the bound configuration. In the case of the PDZ domain, we expect that the missing functional residues should become now classified as structural. We have not done this analysis yet and is planned for future work.

Response_reg domains are usually involved in several biological pathways in the human body, thus, for the sake of simplicity, we used the natural sequences collected with a multiple sequence alignment with the protein 1nxw via BLAST, mainly analysing DNA-binding response regulators. Previous studies show that the protein is activated through a phosphorylation event on the residue 52. However, the DNA-binding activity occurs also in the absence of phosphorylation, mediated by the dimerisation of the domain [58, 59, 60]. Therefore, dimerisation is here inherently connected to protein function.

We correctly identify the 96% of residues previously indicated as involved in phosphorylation, dimerisation and binding as key for protein function and structure. We miss out only the position 110, that is a hydrophobic patch mediating the dimer interface. The analysis on the Response_reg domain differs from the previous ones since it shows large areas of strictly structural relevance (patches at 6-8, 13-21 and 101-110).

Summarising, we were able to correctly identify the vast majority of the residues experimentally proven to be involved in protein functions and protein-protein interaction. Moreover, as a further indication of the validity of the method, in all cases, we find that the residues identified with our analysis fall within the domain of the family (purple bar in Fig. 2). As a general remark, thanks to the low computational cost of the design algorithms, the same procedure could be performed by designing many protein pairs bound to each other in different conformations. Again, from comparing the new analysis with the single protein one, the function residues that would turn into structural would be involved in unknown protein-protein interactions.

It is important to stress that if we would have picked at random the known functional residues with N trials (in our case N is the number of proposed residues), according to the binomial distribution, we would have identified:1 ± 1 for the FKPB family 3 ± 2 for the PDZ family and 4 ± 2 for the Response_reg. All the numbers are well below the actual identified residues. Finally, it is important to stress that the artificial sequences cannot be replaced by random ones because they would not have shown any annotation signal for the functional regions (see Fig. 8).

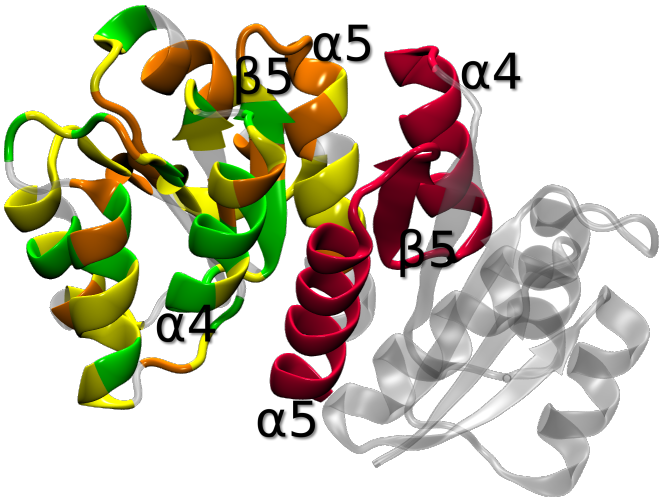


FIG. 5: Dimer form of a response regulator receiver domain, (PDB 1pkx). The dimerisation occur through the α4-β5-α5 interface. α4-β5-α5 of chain A is highlighted in red, while chain B shows the classification obtained with our analysis: functional in green, intermediate in yellow and structural in orange.

Conclusion

We have analysed three major protein families: FKBP, PDZ and Response_reg. From the natural sequences, we extracted the information relative to highly conserved sites and highly co-evolved residue pairs. The analysis resulted in a long list of residues that could potentially have a functional and/or a structural role for the proteins.

For each protein family, we produced an artificial set of sequences that are equivalent solutions to the folding problem.

The researchers that are interested in using our methodology at the moment need three tools all of which are available as open software. The first information necessary is the alignment of the target protein. Then the target protein structure is stripped of the sequence and redesigned using the protein package ViPS available online[94]. At this point, the user will have two lists of aligned proteins one natural and the other artificial. In order to generate the Entropy and FAPC weight, the user can use the Matlab DCA script, free for download [95], and run it on both lists. Finally, the last step is to use the Matlab script in the supplementary materials to compare the weights of the two lists.

We applied the same conservation and co-evolution analyses performed on the natural alignments, obtaining a correlation pattern with many features in common with

the natural ones. However, it is in the differences that we find the main result of this work. The residues that have a strong co-evolution signature only in the natural alignments are residues that do not have a structural role in the protein because otherwise, they would show a similar signature in the analysis of the artificial set. Hence, it is reasonable to assume that if such residues where conserved, they must have a role related to the function of the protein or structural elements that were not included in the design process. For instance, residues that are essential for dimerisation would appear as functional when the protein is designed alone and instead categorised as structural when the dimer is designed as a whole.

The results demonstrate the validity of our automated approach to identify functional residues in protein families. Large scale analysis of the whole proteome using an automated algorithm based on our methodology could give an essential contribution to the identification of functional protein regions. By designing protein complexes, our method could also be used to classify functional residues for their involvement in protein-protein interactions. Retrospectively, the successful annotation of the functional residues proves the validity of our initial hypothesis that has important implications on the structure of the protein sequence space. It further further supports the fundamental prediction that the existing proteins are indeed just a small fraction of all possible proteins and the natural evolution has explored only specific sections with biologically relevant functions.

Outlook

As possible development of the here presented method, one could combine the single site and pair information in one signal, that is the conditioned entropy. Conditioned entropy is an expression of the conditional probability of finding the amino acid a at site i, given the amino acids in all the other positions, therefore including not only the single site statistics but also the pair conservation. This approach would reduce the number of analysis needed for the identification of protein's essential regions. It would be interesting also to test the DCA approach based on the Hopfield-Potts method from Cocco et al. [21].

Recently, Possenti et al. [91] have presented a new method to compute the entropic contribution of each residue of several testing proteins, divided in two contributions structural and all the potential functions further classified in proteolytic cleavage, solubility, and functionality. The latter is extracted from experimental data. The publication demonstrates that it is possible to separate the information contained in the protein evolution process and it would be advantageous to combine such methodology with the one presented here, where we can extract the information about the functional term needed for their analysis. It is also interesting to mention the recent methodology Co-Factor by Zhang et al [90] that is based on structural protein homology. Their work raises the interesting question of whether protein function is encoded in the structure or the sequence or both. The fact that both co-evolution signals and structural alignment are capable of identifying structural residues suggests the latter hypothesis. in fact, from our work and mutation experiments, it is evident that it is possible to design proteins to fold into a structure without its original function. Again, this further confirms that proteins evolution resamples the same subset of solutions compared to the overall protein space. This observation is fascinating and opens the door to exciting speculation that deserves a dedicated study beyond the scope of the present paper.

Finally, we think that our methodology could be used as a starting point for further improvements in particular by reintroducing the information on the indirect couplings measured in the natural alignments similarly to what done by Cheng et al. [13]. Moreover, by redesigning the proteins in known bound conformations (e.g. dimers or larger complexes), it should be possible to reclassify functional or missed residues as structurally crucial for the stability of the dimer.

Acknowledgements

All simulations presented in this paper were carried out on the Vienna Scientific Cluster (VSC).

Funding

We acknowledge support from the VSC School, as well as from the Austrian Science Fund (FWF) project 26253-N27. V. B. acknowledges the support from FWF Grant No. M 2150-N36. IC gratefully acknowledges support from the Ministerio de Economià y Competitividad (MINECO) (FIS2017-89471-R).

Strong signal selection

Firstly, we evaluate single site and pairwise signals (S_u from MSA, Eq.s **??**, and F^{APC} from DCA, Eq. **??**) on both the I) natural sequences and the II) artificial families. The results of this two analysis is then combined in the last step, consisting in III) selecting residues with overall high signals, both natural and artificial, as key for structure and function, and further isolate S and F residues from the overlapping OFSR ones by comparing the intensity of their signals in the natural and artificial analysis.

From the MSA we extract the empirical frequencies for the conservation of each site, and evaluate the single site entropy [92], $S_u(i)$, defined in Eq. **??**. The DCA

analysis is performed over the aligned sequences ignoring the the first 5 and the last 5 residues of each sequence, since they are source of noise. Moreover, to remove trivial correlations, we neglect the couplings between first and second neighbours along the chain. The main result of the DCA analysis are the estimated couplings between the residues pairs.

The DCA couplings are used to the Frobenius Average Product Correction [23, 93], $F^{APC}(i,j)$, defined in Eq. **??**. The F^{APC} is a method to correct for the phylogenic relations between residues and correlates with pair co-evolution events. F^{APC} is based on the idea that correlation between pairs of amino acids is the sum of a true statistical dependency and a background dependency due to the phylogenic relationships. In the F^{APC} it is assumed that the background dependency is a product of independent factors associated with the two positions.

The $F^{APC}(i,j)$ and the entropy $S_u(i)$ provide signals for each residue pair and single residue respectively.

It is important to notice that, for the entropy distribution, outliers that are larger than the average $\overline{S}_c$ correspond to sites of high variability, but since we are interested in the highly conserved ones we exclude such points. Similarly we focus only on the outliers corresponding to strong co-evolution events, hence we exclude the $F^{APC}(i,j)$ outliers below the distribution average $\overline{F}^{APC}$.

We perform a standard Principal Component Analysis (PCA), using residue index (indices) and conservation (co-evolution) signals, implemented in the open-source software Octave (see *PCA source code* section in SI for the source). The residue index are used to map back the strong signals into the protein sequence. Once the $S_u(i)$ and the $F^{APC}(i,j)$ are projected over their respective principal components, we prune for the outliers. We plot a dendrogram of the values where the distances are simple euclidean distances between the points in the eigenvector space. The outliers of $S_u(i)$ are identified by isolating the sites that do not belong to the largest group at low distances (see highlighted region in Fig. 7 and 8). We denote the signals of the corresponding residues as $S_c^*(i)$, and we label as S_{NAT} the list outcome of the analysis on natural signals and as S_{CAT} the artificial one.

The $F^{APC}(i,j)$ outliers are isolated following a similar procedure, but, after constructing the dendrogram ordered according to the distance from the other points in the eigenvector space, we take the $3N$ residue pairs at higher values. The reason is that the F^{APC} dendrogram is much richer than the entropy one and it is difficult to isolate the clusters.

Currently, the procedure is fully automatic and we also tested it against a simple quartile separation that produced similar selections. It would be interesting to apply machine learning algorithm or different cluster algorithms [78], to further test the robustness and/or improve the selection efficiency.

For each of the selected residues, we count the number of pairs in which it is involved among the $F^{APC}(i,j)$ signals, and we name them as Co-Evolution Signals $CES(i) = \sum_j F^{*APC}(i,j)$, both for natural and artificial analyses.

We then use the $CES(i)$ value of each residue to construct two lists: a natural and an artificial one. We consider a residue natural if $CES(i)_{NAT} > CES(i)_{CAT}$ and $CES(i)_{CAT} < 10$, while artificial if $CES(i)_{NAT} < CES(i)_{CAT}$ and $CES(i)_{CAT} \geq 10$, where the threshold of 10 is used to filter the false positives of $CES(i)_{CAT}$.

The CES_{NAT} selection list of natural signals identified with the above mentioned procedure, according to our hypothesis, intrinsically contains structural and functional information, embedded among the predicted couplings. Hence, potentially, some of the correlated residues might be involved in a protein function.

In Fig. **??** (see section *Contact maps* of SI for further information) we show the contacts predicted from the analysis on natural sequences, together with total native contacts. From the comparison we observe that we predict the majority of the native contacts (blue dots), while we still have some false positive predictions that do not correspond to natural structural contacts. The latter are the most interesting for the present study and we speculate that should indeed hide specific conserved residues necessary for the protein functionality but not for the structural stability.

On the other hand, the list CES_{CAT} outcome of the artificial analysis contains only structural information, therefore being effectively a list of the predicted strongest bonds in the native structure of the domain.

The entirety of residues in the two lists are considered as key for the protein structure and/or function.

We then operate a pre-classification step: we process conservation and co-evolution independently, and the assign a score (-1 for structure predominant character, 1 for function predominant and 0 in between) to each residue. As for the conservation signal, we assign 1 if the residue exclusively belong to the S_{NAT} list, -1 if is exclusively in S_{CAT} and 0 if it is shared among the two. Similarly, we subtract the natural $CES(i)$ value to the artificial one for each residue of the $CES_{CAT/NAT}$ lists, and we assign a score basing on the discrepancy gap. We set a threshold and we consider functional (score= 1) a residue whose natural/artificial gap is $CES_{NAT} - CES_{CAT} \geq 5$, structural (score= -1) one with $CES_{NAT} - CES_{CAT} \leq -5$, and overlapping (score= 0) a residue whose signals are comparable $5 > CES_{NAT} - CES_{CAT} > 5$.

Finally, if in common, we sum the scores assigned to each residue in the conservation and co-evolution lists and we categorise as F the positions with a total sum $\Sigma \geq 1$, S the ones with $\Sigma \leq -1$ and OFSR residues with $\Sigma = 0$.

The residues selected with the above mentioned procedure, along with their classification, are listed in Fig. 6. For each residue in the $CES_{NAT/CAT}$ lists, the relative CES_{NAT} and CES_{CAT} values are written in the adjacent columns, so that we can compute the difference between them and assign a red colour (strong signal) if the absolute value of such difference is ≥ 5. If the signal is strong for residues selected in the natural analysis, we assign a green colour (score= 1, functional) to the relative CES_{NAT} column. Viceversa, if the signal is strong in the caterpillar analysis, we assign the orange colour (score= -1, structural). When the signal is weak, i.e. the difference is smaller than 5, we assign the yellow colour (score= 0, intermediate). For the entropy, we assign green when the signal is present only in the S_{NAT} list, orange if present only in the S_{CAT} list, yellow if in common. For the final assignment, we combine the two analysis, on entropy and CES, and we assign green if score≥ 1, orange if score≤ -1 and yellow if score= 0. In principle the intermediates should be classified as structural as well since are in common between the two ensemble of sequences. However, they have small CES and the value could be sensitive to the cluster algorithm used to identify the strongest $F^{APC}(i,j)$ signals. Hence, at this stage we would like to focus on the functional and structural ones for the analysis of the prediction power of our methodology. Preliminary tests performed with an independent cluster algorithm [78] identified the same functional and structural residues as done here (data not show).

Please note that the functional residues that we ultimately identify do not depend on the particular choice of the PDB taken as reference structure, since the domain structure is conserved in the family and the alignment guarantees the mapping of the functional residues on the selected domain type of every member of the family.

* Electronic address: icoluzza@cicbiomagune.es

[1] Csaba Magyar, va Tüdős, and Istvn Simon. Functionally and structurally relevant residues of enzymes: are they segregated or overlapping? *FEBS Letters*, 567(2-3):239–242, 6 2004. ISSN 00145793. doi: 10.1016/j.febslet.2004.04.070. URL https://www.sciencedirect.com/science/article/pii/S0014579304005460#BIB7http://doi.wiley.com/10.1016/j.febslet.2004.04.070.

[2] Peteris Zikmanis and Inara Kampenusa. Relationship between Metabolic Fluxes and Sequence-Derived Properties of Enzymes. *International Scholarly Research Notices*, 2014: 1–9, 2014. ISSN 2356-7872. doi: 10.1155/2014/817102. URL http://www.ncbi.nlm.nih.gov/pubmed/27437461http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4897147https://www.hindawi.com/archive/2014/817102/.

[3] James A. Wells and Christopher L. McClendon. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450(7172):1001–1009, 2007. ISSN 14764687. doi: 10.1038/nature06526.

[4] Peter Vanhee, Almer M. van der Sloot, Erik Verschueren, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. Computational design of peptide ligands. *Trends in Biotechnology*, 29(5):231–239, 2011. ISSN 01677799. doi: 10.1016/j.tibtech.2011.01.004. URL http://dx.doi.org/10.1016/j.tibtech.2011.01.004.

[5] Chun Meng Song, Shen Jean Lim, and Joo Chuan Tong. Recent advances in computer-aided drug design. *Briefings in Bioinformatics*, 10(5):579–591, 2009. ISSN 14675463. doi: 10.1093/bib/bbp023.

[6] A. Lavecchia and C. Giovanni. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry*, 20(23):2839–2860, 6 2013. ISSN 09298673. doi: 10.2174/09298673113209990001. URL http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed11&NEWS=N&AN=2013445557http://www.eurekaselect.com/openurl/content.php?genre=article&issn=0929-8673&volume=20&issue=23&spage=2839.

[7] Ivan Coluzza. Computational protein design: a review. *Journal of Physics: Condensed Matter*, 29(14):143001, 4 2017. ISSN 0953-8984. doi: 10.1088/1361-648X/aa5c76. URL http://iopscience.iop.org/10.1088/1361-648X/aa5c76http://iopscience.iop.org/article/10.1088/1361-648X/aa5c76http://stacks.iop.org/0953-8984/29/i=14/a=143001?key=crossref.6b75a4256c5bfe8a8e20e6ef3834f61e.

[8] Michael E. Cusick, Niels Klitgord, Marc Vidal, and David E. Hill. Interactome: Gateway into systems biology. *Human Molecular Genetics*, 14(SUPPL. 2):171–181, 2005. ISSN 09646906. doi: 10.1093/hmg/ddi335.

[9] Alia Qureshi Emili and Gerard Cagney. Large-scale functional analysis using peptide or protein arrays. *Nature Biotechnology*, 18(4):393–397, 4 2000. ISSN 1087-0156. doi: 10.1038/74442. URL http://www.nature.com/articles/nbt0400_393.

[10] Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 1 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1344. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1344http://nar.oxfordjournals.org/content/early/2015/12/15/nar.gkv1344.full.

[11] Scott McGinnis and Thomas L. Madden. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(WEB SERVER ISS.):20–25, 2004. ISSN 03051048. doi: 10.1093/nar/gkh435.

[12] Elliott Lever and Denise Sheer. The role of nuclear organization in cancer. *The Journal of pathology*, 220(September):114–125, 2010. ISSN 1096-9896. doi: 10.1002/path.

[13] Ryan R Cheng, Faruck Morcos, Herbert Levine, and Jos N Onuchic. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of Sciences*, 111(5):E563–E571, 2014. ISSN 0027-8424.

[14] David De Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14 (4):249, 2013. ISSN 1471-0064.

[15] Bosco K Ho, David Perahia, and Ashley M Buckle. Hybrid approaches to molecular simulation. *Current opinion in structural biology*, 22(3):386–393, 2012. ISSN 0959-440X.

[16] Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072, 2012. ISSN 1546-1696.

[17] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 1 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102472&tool=pmcentrez&rendertype=abstract.

[18] Tanja Kortemme, Lukasz A Joachimiak, Alex N Bullock, Aaron D Schuler, Barry L Stoddard, and David Baker. Computational redesign of protein-protein interaction specificity. *Nature Structural & Molecular Biology*, 11(4):371–379, 4 2004. ISSN 1545-9993. doi: 10.1038/nsmb749. URL http://www.nature.com/doifinder/10.1038/nsmb749.

[19] Tanja Kortemme and David Baker. Computational design of protein-protein interactions. *Current Opinion in Chemical Biology*, 8(1):91–97, 2004. ISSN 13675931. doi: 10.1016/j.cbpa.

2003.12.008.

[20] Georg Casari, Chris Sander, and Alfonso Valencia. A method to predict functional residues in proteins. *Nat. Struct. Mol. Biol.*, 2 (2):171–178, feb 1995. ISSN 1545-9993. doi: 10.1038/nsb0295-171. URL http://www.nature.com/articles/nsb0295-171.

[21] Simona Cocco, Remi Monasson, and Martin Weigt. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS computational biology*, 9(8):e1003176, 2013. ISSN 1553-7358.

[22] Angel E Dago, Alexander Schug, Andrea Procaccini, James A Hoch, Martin Weigt, and Hendrik Szurmant. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences*, 109(26):E1733–E1742, 2012. ISSN 0027-8424.

[23] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87 (1):012707, 1 2013. ISSN 1539-3755. doi: 10.1103/PhysRevE.87.012707. URL https://link.aps.org/doi/10.1103/PhysRevE.87.012707.

[24] Bryan Lunt, Hendrik Szurmant, Andrea Procaccini, James A Hoch, Terence Hwa, and Martin Weigt. Inference of direct residue contacts in two-component signaling. In *Methods in enzymology*, volume 471, pages 17–41. Elsevier, 2010. ISBN 0076-6879.

[25] Faruck Morcos, Terence Hwa, Jos N Onuchic, and Martin Weigt. Direct coupling analysis for protein contact prediction. In *Protein Structure Prediction*, pages 55–70. Springer, 2014.

[26] Faruck Morcos, Biman Jana, Terence Hwa, and Jos N Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, 110(51):20533–20538, 2013. ISSN 0027-8424.

[27] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, Jos N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108 (49):E1293–E1301, 2011. ISSN 0027-8424.

[28] Andrea Procaccini, Bryan Lunt, Hendrik Szurmant, Terence Hwa, and Martin Weigt. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. *PloS one*, 6(5):e19729, 2011. ISSN 1932-6203.

[29] Alexander Schug, Martin Weigt, Jos N Onuchic, Terence Hwa, and Hendrik Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106(52):22124–22129, 2009. ISSN 0027-8424.

[30] Joanna I Sułkowska, Faruck Morcos, Martin Weigt, Terence Hwa, and Jos N Onuchic. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, 109(26):10340–10345, 2012. ISSN 0027-8424.

[31] Hendrik Szurmant and James A Hoch. Statistical analyses of protein sequence alignments identify structures and mechanisms in signal activation of sensor histidine kinases. *Molecular microbiology*, 87(4):707–712, 2013. ISSN 1365-2958.

[32] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0805923106. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0805923106.

[33] Carol A Rohl and David Baker. De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J. Am. Chem. Soc.*, 124(11):2723–2729, mar 2002. ISSN 0002-7863. doi: 10.1021/ja016880e. URL https://doi.org/10.1021/ja016880e.

[34] Nikolaos G Sgourakis, Oliver F Lange, Frank DiMaio, Ingemar André, Nicholas C Fitzkee, Paolo Rossi, Gaetano T Montelione, Ad Bax, and David Baker. Determination of the Structures of Symmetric Protein Oligomers from NMR Chemical Shifts and Residual Dipolar Couplings. *J. Am. Chem. Soc.*, 133(16):6288–6298, apr 2011. ISSN 0002-7863. doi: 10.1021/ja111318m. URL https://doi.org/10.1021/ja111318m.

[35] Steffen Lindert and J Andrew McCammon. Improved cryoEM-Guided Iterative Molecular DynamicsRosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction. *J. Chem. Theory Comput.*, 11(3):1337–1346, mar 2015. ISSN 1549-9618. doi: 10.1021/ct500995d. URL https://doi.org/10.1021/ct500995d.

[36] Justin Chen, Jiming Chen, Giovanni Pinamonti, and Cecilia Clementi. Learning Effective Molecular Models from Experimental Observables. *J. Chem. Theory Comput.*, 14(7):3849–3858, jul 2018. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b00187. URL https://doi.org/10.1021/acs.jctc.8b00187.

[37] Jeffrey R Wagner, Christopher T Lee, Jacob D Durrant, Robert D Malmstrom, Victoria A Feher, and Rommie E Amaro. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem. Rev.*, 116(11):6370–6390, jun 2016. ISSN 0009-2665. doi: 10.1021/acs.chemrev.5b00631. URL https://doi.org/10.1021/acs.chemrev.5b00631.

[38] E. I. Shakhnovich. Proteins with selected sequences fold into unique native conformation. *Physical Review Letters*, 72(24):3907–3910, 6 1994. ISSN 00319007. doi: 10.1103/PhysRevLett.72.3907. URL http://link.aps.org/doi/10.1103/PhysRevLett.72.3907http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.72.3907.

[39] A. M. Gutin and E. I. Shakhnovich. Ground state of random copolymers and the discrete random energy model. *The Journal of Chemical Physics*, 98(10):8174–8177, 5 1993. ISSN 0021-9606. doi: 10.1063/1.464522. URL http://aip.scitation.org/doi/10.1063/1.464522papers2://publication/uuid/AEF0E5B1-DB7F-4FEB-9367-BFD43B38854D.

[40] Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a Novel Globular Protein Fold with Atomic Level Accuracy. *Science*, 302 (5649):1364–1368, 11 2003. ISSN 0036-8075. doi: 10.1126/science.1089427. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14631033.

[41] Ho Ki Fung, William J. Welsh, and Christodoulos A. Floudas. Computational de novo peptide and protein design: Rigid templates versus flexible templates. *Industrial and Engineering Chemistry Research*, 47(4):993–1001, 2 2008. ISSN 08885885. doi: 10.1021/ie071286k. URL http://pubs.acs.org/doi/abs/10.1021/ie071286k.

[42] I Samish, C Macdermaid, J Perez-Aguilar, and J Saven. Theoretical and computational protein design. *Annual Review of Physical Chemistry*, 62:129–149, 2011. ISSN 1545-1593. doi: 10.1146/annurev-physchem-032210-103509. URL papers://ae875177-834e-4ba8-8523-120292c79891/Paper/p4643.

[43] Andrew R. Thomson, Christopher W. Wood, Antony J. Burton, Gail J. Bartlett, Richard B. Sessions, R. Leo Brady, and Derek N. Woolfson. Computational design of water-soluble α-helical barrels. *Science*, 346(6208):485–488, 10 2014. ISSN 0036-8075. doi: 10.1126/science.1257452. URL http://www.sciencemag.org/lookup/doi/10.1126/science.1257452http://www.sciencemag.org/cgi/doi/10.1126/science.1257452.

[44] Po-Ssu Ssu Huang, Gustav Oberdorfer, Chunfu Xu, Xue Y. Pei, Brent L. Nannenga, Joseph M. Rogers, Frank DiMaio, Tamir Gonen, Ben Luisi, and David Baker. High thermodynamic stability of parametrically designed helical bundles. *Science*, 346(6208):481–485, 10 2014. ISSN 1095-9203. doi: 10.1126/science.1257481. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1257481http://www.sciencemag.org/content/346/6208/481.abstract.

[45] Marco Chino, Ornella Maglio, Flavia Nastri, Vincenzo Pavone, William F. Degrado, and Angela Lombardi. Artificial Diiron Enzymes with a De Novo Designed Four-Helix Bundle Structure. *European Journal of Inorganic Chemistry*, 2015(21):3371–3390, 7 2015. ISSN 10990682. doi: 10.1002/ejic.201500470. URL http://doi.wiley.com/10.1002/ejic.201500470.

[46] Thomas Gaillard and Thomas Simonson. Full Protein Sequence Redesign with an MMGBSA Energy Function. *J. Chem. Theory Comput.*, 13(10):4932–4943, oct 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00202. URL https://doi.org/10.1021/acs.jctc.7b00202.

[47] David Mignon, Nicolas Panel, Xingyu Chen, Ernesto J Fuentes, and Thomas Simonson. Computational Design of the Tiam1 PDZ Domain and Its Ligand Binding. *J. Chem. Theory Comput.*, 13

(5):2271–2289, may 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.6b01255. URL https://doi.org/10.1021/acs.jctc.6b01255.

[48] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, Jason W Labonte, Michael S Pacella, Richard Bonneau, Philip Bradley, Roland L Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J Gray. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.*, 13(6):3031–3048, jun 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00125. URL https://doi.org/10.1021/acs.jctc.7b00125.

[49] Baruch Z Harris and Wendell A Lim. Mechanism and role of PDZ domains in signaling complex assembly. *Journal of Cell Science*, 114:3219–3231, 2001. URL http://jcs.biologists.org/content/joces/114/18/3219.full.pdf.

[50] Jing-Song Fan and Mingjie Zhang. Signaling Complex Organization by PDZ Domain Proteins. *Neurosignals*, 11(6):315–321, 2002. ISSN 1424-862X. doi: 10.1159/000068256. URL https://www.karger.com/Article/Pdf/68256https://www.karger.com/Article/FullText/68256.

[51] David Cowburn. Peptide recognition by PTB and PDZ domains. *Current Opinion in Structural Biology*, 7(6):835–838, 1997. ISSN 0959440X. doi: 10.1016/S0959-440X(97)80155-8.

[52] Z. Songyang, A. S. Fanning, C. Fu, J. Xu, S. M. Marfatia, A. H. Chishti, A. Crompton, A. C. Chan, J. M. Andersen, and L. C. Cantley. Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science*, 275(5296):73–77, 1997. ISSN 00368075. doi: 10.1126/science.275.5296.73.

[53] Declan A. Doyle, Alice Lee, John Lewis, Eunjoon Kim, Morgan Sheng, and Roderick MacKinnon. Crystal structures of a complexed and peptide-free membrane protein- binding domain: Molecular basis of peptide recognition by PDZ. *Cell*, 85(7):1067–1076, 1996. ISSN 00928674. doi: 10.1016/S0092-8674(00)81307-0.

[54] Joo H. Morais Cabral, Carlo Petosa, Michael J Sutcliffe, Sami Raza, Olwyn Byron, Florence Poy, Shirin M Marfatia, Athar H Chishti, and Robert C Liddington. Crystal structure of a PDZ domain. *Nature*, 382(6592):649–652, 8 1996. ISSN 0028-0836. doi: 10.1038/382649a0. URL http://www.ncbi.nlm.nih.gov/pubmed/8757139%5Cnhttp://www.nature.com/nature/journal/v382/n6592/pdf/382649a0.pdfhttp://www.nature.com/doifinder/10.1038/382649a0.

[55] K. P. Wilson, M. M. Yamashita, M. D. Sintchak, S. H. Rotstein, M. A. Murcko, J. Boger, J. A. Thomson, M. J. Fitzgibbon, J. R. Black, and M. A. Navia. Comparative X-ray structures of the major binding protein for the immunosuppressant FK506 (tacrolimus) in unliganded form and in complex with FK506 and rapamycin. *Acta Crystallographica Section D Biological Crystallography*, 51(4):511–521, 7 1995. ISSN 09074449. doi: 10.1107/S0907444994014514. URL http://scripts.iucr.org/cgi-bin/paper?S0907444994014514.

[56] Michael T.G. Ivery. Immunophilins: Switched on protein binding domains? *Medicinal Research Reviews*, 20(6):452–484, 11 2000. ISSN 0198-6325. doi: 10.1002/1098-1128(200011)20:6⟨452::AID-MED2⟩3.0.CO;2-6. URL http://doi.wiley.com/10.1002/1098-1128%28200011%2920%3A6%3C452%3A%3AAID-MED2%3E3.0.CO%3B2-6.

[57] Jacqueline Dornan, Paul Taylor, and Malcolm Walkinshaw. Structures of Immunophilins and their Ligand Complexes. *Current Topics in Medicinal Chemistry*, 3(12):1392–1409, 8 2003. ISSN 15680266. doi: 10.2174/1568026033451899. URL http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1568-0266&volume=3&issue=12&spage=1392.

[58] Colin J Bent, Neil W Isaacs, Timothy J Mitchell, and Alan Riboldi-Tunnicliffe. Crystal structure of the response regulator 02 receiver domain, the essential YycF two-component system of Streptococcus pneumoniae in both complexed and native states. *Journal of bacteriology*, 186(9):2872–9, 5 2004. ISSN 0021-9193. doi: 10.2210/PDB1NXW/PDB. URL https://www.rcsb.org/structure/1nxwhttp://www.ncbi.nlm.nih.gov/pubmed/15090529http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC387779.

[59] Alejandro Toro-Roman, Timothy R. Mack, and Ann M. Stock. Structural Analysis and Solution Studies of the Activated Regulatory Domain of the Response Regulator ArcA: A Symmetric Dimer Mediated by the α4-β5-α5 Face. *Journal of Molecular Biology*, 349(1):11–26, 5 2005. ISSN 00222836. doi: 10.1016/j.jmb.2005.03.059. URL https://www-sciencedirect-com.uaccess.univie.ac.at/science/article/pii/S0022283605003505#fig5http://linkinghub.elsevier.com/retrieve/pii/S0022283605003505.

[60] Minh-Phuong Nguyen, Joo-Mi Yoon, Man-Ho Cho, and Sang-Won Lee. Prokaryotic 2-component systems and the OmpR/PhoB superfamily. *Canadian Journal of Microbiology*, 61(11):799–810, 11 2015. ISSN 0008-4166. doi: 10.1139/cjm-2015-0345. URL http://www.nrcresearchpress.com/doi/10.1139/cjm-2015-0345.

[61] Ivan Coluzza. A Coarse-Grained approach to protein design: Learning from design to understand folding. *PLoS ONE*, 6(7):e20853, 1 2011. ISSN 19326203. doi: 10.1371/journal.pone.0020853. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3128589&tool=pmcentrez&rendertype=abstract.

[62] Ivan Coluzza. Transferable Coarse-Grained Potential for De Novo Protein Folding and Design. *PLoS ONE*, 9(12):e112852, 12 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0112852. URL http://dx.plos.org/10.1371/journal.pone.0112852http://www.ncbi.nlm.nih.gov/pubmed/25436908.

[63] Szilvia Szep, Sheldon Park, Eric T. Boder, Gregory D. Van Duyne, and Jeffery G. Saven. Structural coupling between FKBP12 and buried water. *Proteins: Structure, Function, and Bioinformatics*, 74(3):603–611, 2 2009. ISSN 08873585. doi: 10.1002/prot.22176. URL http://doi.wiley.com/10.1002/prot.22176.

[64] Ivan Coluzza and Daan Frenkel. Virtual-move parallel tempering. *ChemPhysChem*, 6(9):1779–1783, 9 2005. ISSN 14394235. doi: 10.1002/cphc.200400629. URL http://www3.interscience.wiley.com/journal/111081506/abstract?CRETRY=1&SRETRY=0http://www.ncbi.nlm.nih.gov/pubmed/16110517.

[65] Daan Frenkel. Speed-up of Monte Carlo simulations by sampling of rejected states. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17571–17575, 12 2004. ISSN 0027-8424. doi: 10.1073/pnas.0407950101. URL http://www.pnas.org/content/101/51/17571%5Cnhttp://www.pnas.org/cgi/doi/10.1073/pnas.0407950101.

[66] E. I. Shakhnovich and A. M. Gutin. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophysical Chemistry*, 34(3):187–199, 1989. ISSN 03014622. doi: 10.1016/0301-4622(89)80058-4. URL papers2://publication/uuid/9C8CCDEB-1915-446E-A779-02585F0500EE.

[67] Vijay S Pande, Alexander Yu. Grosberg, and Toyoichi Tanaka. Heteropolymer freezing and design: Towards physical models of protein folding. *Reviews of Modern Physics*, 72(1):259–314, 2000. ISSN 0034-6861 (print), 1538-4527 (electronic), 1539-0756. doi: http://dx.doi.org/10.1103/RevModPhys.72.259. URL papers2://publication/uuid/C216088C-3A62-4BB4-8DA3-A6F70F0A9759%5Cnhttp://link.aps.org/doi/10.1103/RevModPhys.72.259.

[68] Chiara Cardelli, Francesca Nerattini, Luca Tubiana, Valentino Bianco, Christoph Dellago, Francesco Sciortino, and Ivan Coluzza. General Methodology to Identify the Minimum Alphabet Size for Heteropolymer Design. *Adv. Theory Simulations*, page 1900031, may 2019. ISSN 2513-0390. doi: 10.1002/adts.201900031. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adts.201900031.

[69] Chiara Cardelli, Valentino Bianco, Lorenzo Rovigatti, Francesca Nerattini, Luca Tubiana, Christoph Dellago, and Ivan Coluzza. The role of directional interactions in the designability of generalized heteropolymers. *Scientific Reports*, 7(1):4986, 12 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-04720-7. URL http://dx.doi.org/10.1038/s41598-017-04720-7http://www.nature.com/articles/s41598-017-04720-7.

[70] Francesca Nerattini, Luca Tubiana, Chiara Cardelli, Valentino Bianco, Christoph Dellago, and Ivan Coluzza. Protein design under competition for amino acids availability. *bioRxiv*, 1 2018. URL http://biorxiv.org/content/early/2018/05/25/331736.abstract.

[71] Valentino Bianco, Neus Pagès-Gelabert, Ivan Coluzza, and Giancarlo Franzese. How the stability of a folded protein depends on interfacial water properties and residue-residue interactions. *Journal of Molecular Liquids*, 245:129–139, 11 2017. ISSN 0167-7322. doi: 10.1016/j.molliq.2017.08.026. URL http:

//linkinghub.elsevier.com/retrieve/pii/S0167732217315416.

[72] A Maritan, C Micheletti, A Trovato, and J R Banavar. Optimal shapes of compact strings. *Nature*, 406(6793):287–90, 7 2000. ISSN 0028-0836. doi: 10.1038/35018538. URL http://www.ncbi.nlm.nih.gov/pubmed/10917526.

[73] George D Rose, Patrick J Fleming, Jayanth R Banavar, and Amos Maritan. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences*, 103(45):16623–16633, 11 2006. ISSN 0027-8424. doi: 10.1073/pnas.0606843103. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0606843103.

[74] Trinh X Hoang, Luca Marsella, Antonio Trovato, Flavio Seno, Jayanth R Banavar, and Amos Maritan. Common attributes of native-state structures of proteins, disordered proteins, and amyloid. *Proceedings of the National Academy of Sciences of the United States of America*, 103(18):6883–8, 5 2006. ISSN 0027-8424. doi: 10.1073/pnas.0601824103. URL papers2://publication/doi/10.1073/pnas.0601824103%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1458988&tool=pmcentrez&rendertype=abstract.

[75] Trinh Xuan Hoang, Antonio Trovato, Flavio Seno, Jayanth R Banavar, and Amos Maritan. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21):7960–4, 5 2004. ISSN 0027-8424. doi: 10.1073/pnas.0402525101. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=419539&tool=pmcentrez&rendertype=abstract.

[76] Predrag Kukic, Arvind Kannan, Maurits J J Dijkstra, Sanne Abeln, Carlo Camilloni, and Michele Vendruscolo. Mapping the Protein Fold Universe Using the CamTube Force Field in Molecular Dynamics Simulations. *PLOS Computational Biology*, 11(10):e1004435, 10 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004435. URL http://dx.plos.org/10.1371/journal.pcbi.1004435.

[77] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 10 1990. ISSN 00222836. doi: 10.1016/S0022-2836(05)80360-2. URL https://www-sciencedirect-com.uaccess.univie.ac.at/science/article/pii/S0022283605803602?via%3Dihubhttp://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

[78] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 6 2014. ISSN 0036-8075. doi: 10.1126/science.1242072. URL http://www.ncbi.nlm.nih.gov/pubmed/24970081http://www.sciencemag.org/cgi/doi/10.1126/science.1242072.

[79] Martin Vogtherr, Doris M. Jacobs, Tatjana N. Parac, Marcus Maurer, Andreas Pahl, Krishna Saxena, Heinz Rüterjans, Christian Griesinger, and Klaus M. Fiebig. NMR solution structure and dynamics of the peptidyl-prolyl cis-trans isomerase domain of the trigger factor from Mycoplasma genitalium compared to FK506-binding protein. *Journal of Molecular Biology*, 318(4):1097–1115, 2002. ISSN 00222836. doi: 10.1016/S0022-2836(02)00112-2.

[80] Thomas Tradler, Gerlind Stoller, Karl P. Rücknagel, Angelika Schierhorn, Jens U. Rahfeld, and Gunter Fischer. Comparative mutational analysis of peptidyl prolyl cis/trans isomerases: Active sites of Escherichia coli trigger factor and human FKBP12. *FEBS Letters*, 407(2):184–190, 1997. ISSN 00145793. doi: 10.1016/S0014-5793(97)00345-1.

[81] Maria Raffaella Martina, Eleonora Tenori, Marco Bizzarri, Stefano Menichetti, Gabriella Caminati, and Piero Procacci. The precise chemical-physical nature of the pharmacore in FK506 binding protein inhibition: ElteX, a new class of nanomolar FKBP12 ligands. *Journal of Medicinal Chemistry*, 56(3):1041–1051, 2013. ISSN 00222623. doi: 10.1021/jm3015052.

[82] Pulak Ranjan Nath and Noah Isakov. Insights into peptidyl-prolyl cis-trans isomerase structure and function in immunocytes. *Immunology Letters*, 163(1):120–131, 2015. ISSN 18790542. doi: 10.1016/j.imlet.2014.11.002. URL http://dx.doi.org/10.1016/j.imlet.2014.11.002.

[83] Albert C. Pan, Huafeng Xu, Timothy Palpant, and David E. Shaw. Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 13(7):3372–3377, 2017. ISSN 15499626. doi: 10.1021/acs.jctc.7b00172.

[84] Antnio J M Ribeiro, Gemma L. Holliday, Nicholas Furnham, Jonathan D. Tyzack, Katherine Ferris, and Janet M. Thornton. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research*, 46(D1):D618–D623, 1 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1012. URL http://academic.oup.com/nar/article/46/D1/D618/4584620.

[85] Ambrish Roy, Jianyi Yang, and Yang Zhang. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, 40(W1):W471–W477, 7 2012. ISSN 0305-1048. doi: 10.1093/nar/gks372. URL http://www.ncbi.nlm.nih.gov/pubmed/22570420http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3394312https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks372.

[86] Ambrish Roy, Alper Kucukural, and Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 4 2010. ISSN 1754-2189. doi: 10.1038/nprot.2010.5. URL http://www.ncbi.nlm.nih.gov/pubmed/20360767http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2849174http://www.nature.com/articles/nprot.2010.5.

[87] Lin Zhang, Diannan Lu, and Zheng Liu. How native proteins aggregate in solution: A dynamic Monte Carlo simulation. *Biophysical Chemistry*, 133(1-3):71–80, 2008. ISSN 03014622. doi: 10.1016/j.bpc.2007.12.008.

[88] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12(1):7–8, 1 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3213. URL http://www.ncbi.nlm.nih.gov/pubmed/25549265http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4428668http://www.nature.com/articles/nmeth.3213.

[89] Adel Golovin and Kim Henrick. MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, 9(1):312, 7 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-312. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-312.

[90] Chengxin Zhang, Peter L. Freddolino, and Yang Zhang. COFACTOR: Improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.*, 45(W1):W291–W299, 2017. ISSN 13624962. doi: 10.1093/nar/gkx366.

[91] Andrea Possenti, Michele Vendruscolo, Carlo Camilloni, and Guido Tiana. A method for partitioning the information contained in a protein sequence between its structure and function. *Proteins: Structure, Function, and Bioinformatics*, 5 2018. ISSN 08873585. doi: 10.1002/prot.25527. URL http://doi.wiley.com/10.1002/prot.25527.

[92] Thomas M Cover and Joy a Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York, USA, 1991. ISBN 0471062596. doi: 10.1002/0471200611. URL http://doi.wiley.com/10.1002/0471200611.

[93] Richard R. Stein, Debora S. Marks, and Chris Sander. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Computational Biology*, 11(7):1–22, 2015. ISSN 15537358. doi: 10.1371/journal.pcbi.1004182.

[94] https://bitbucket.org/ivan_coluzza/vienna-protein-simulator/src/Optmized/

[95] http://dca.rice.edu/portal/dca/

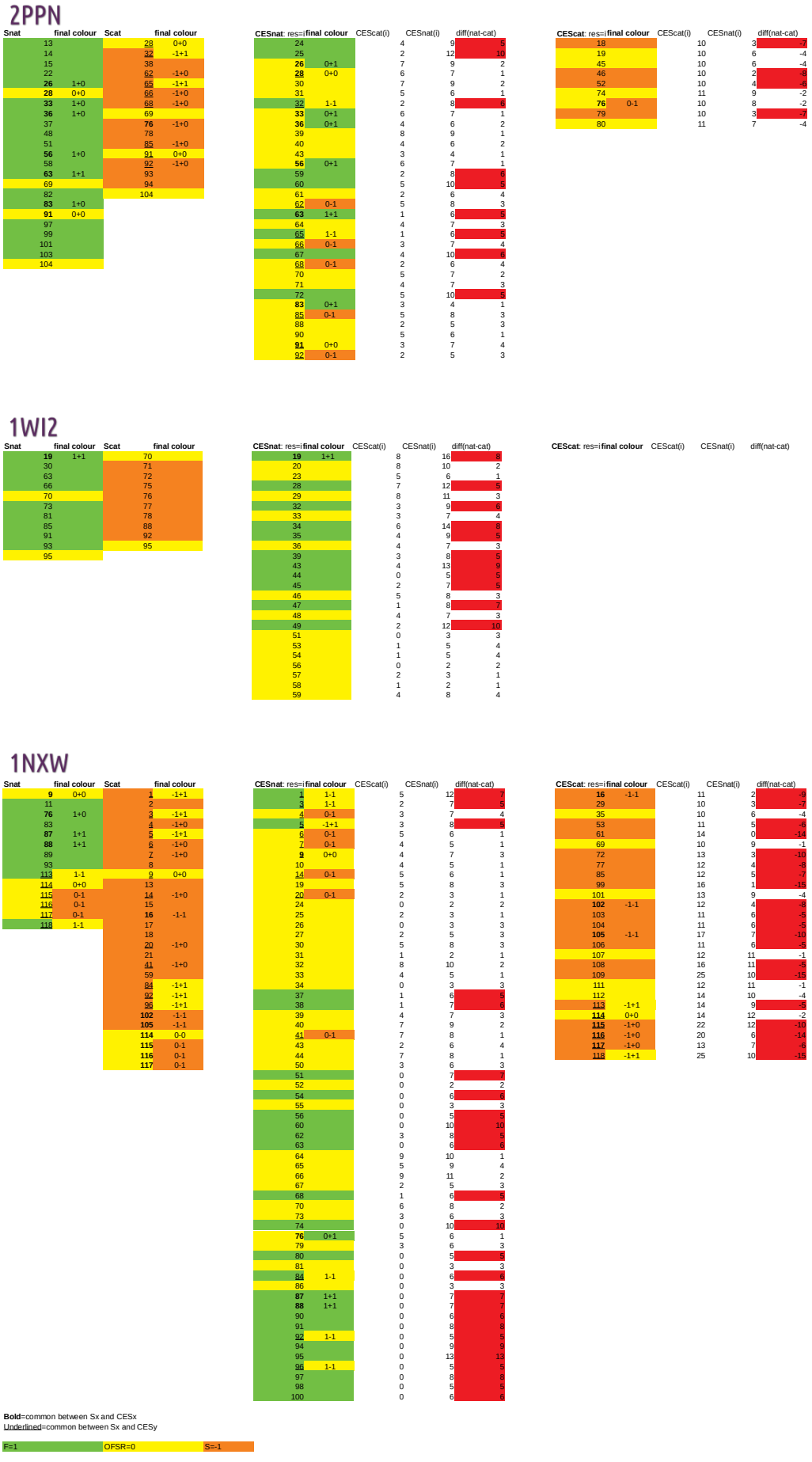
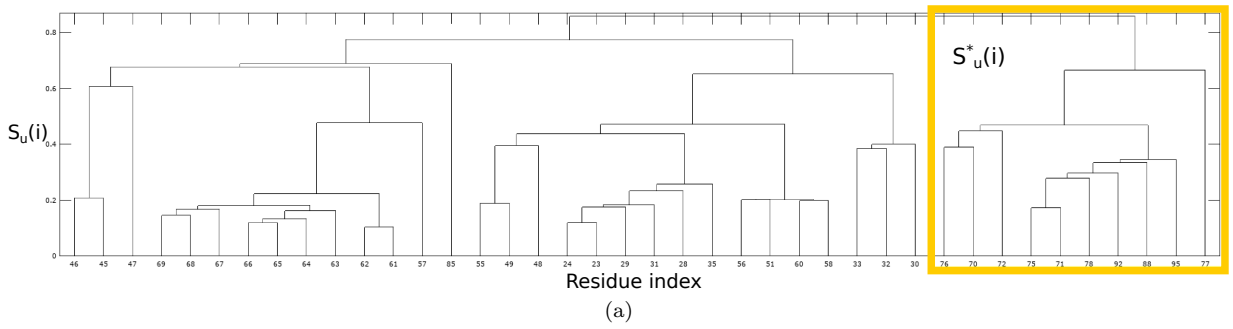
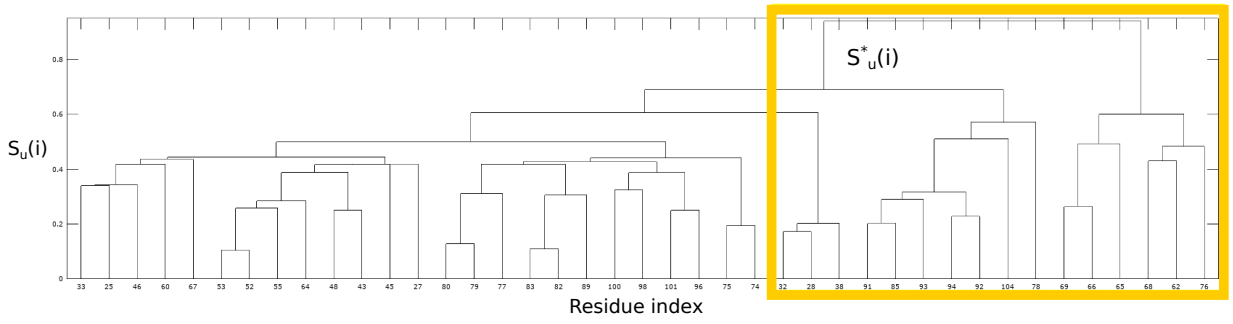


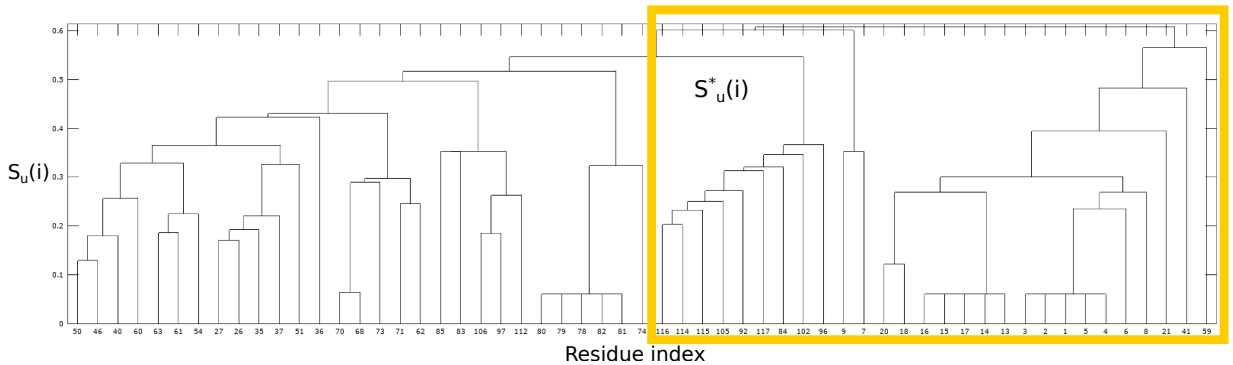
FIG. 6: Residues S_{NAT}, S_{CAT}, CES_{NAT} and CES_{CAT}.



(a)



(b)



(c)

FIG. 7: Dendrogram of the entropy $S_u(i)$ calculated on the artificial sequences for: a) the PDZ, b) FKBP and c) Response Regulator domains. In yellow we have highlighted $S_c^*(i) = S_{CAT}$, that is the outliers that do not belong to the largest cluster.

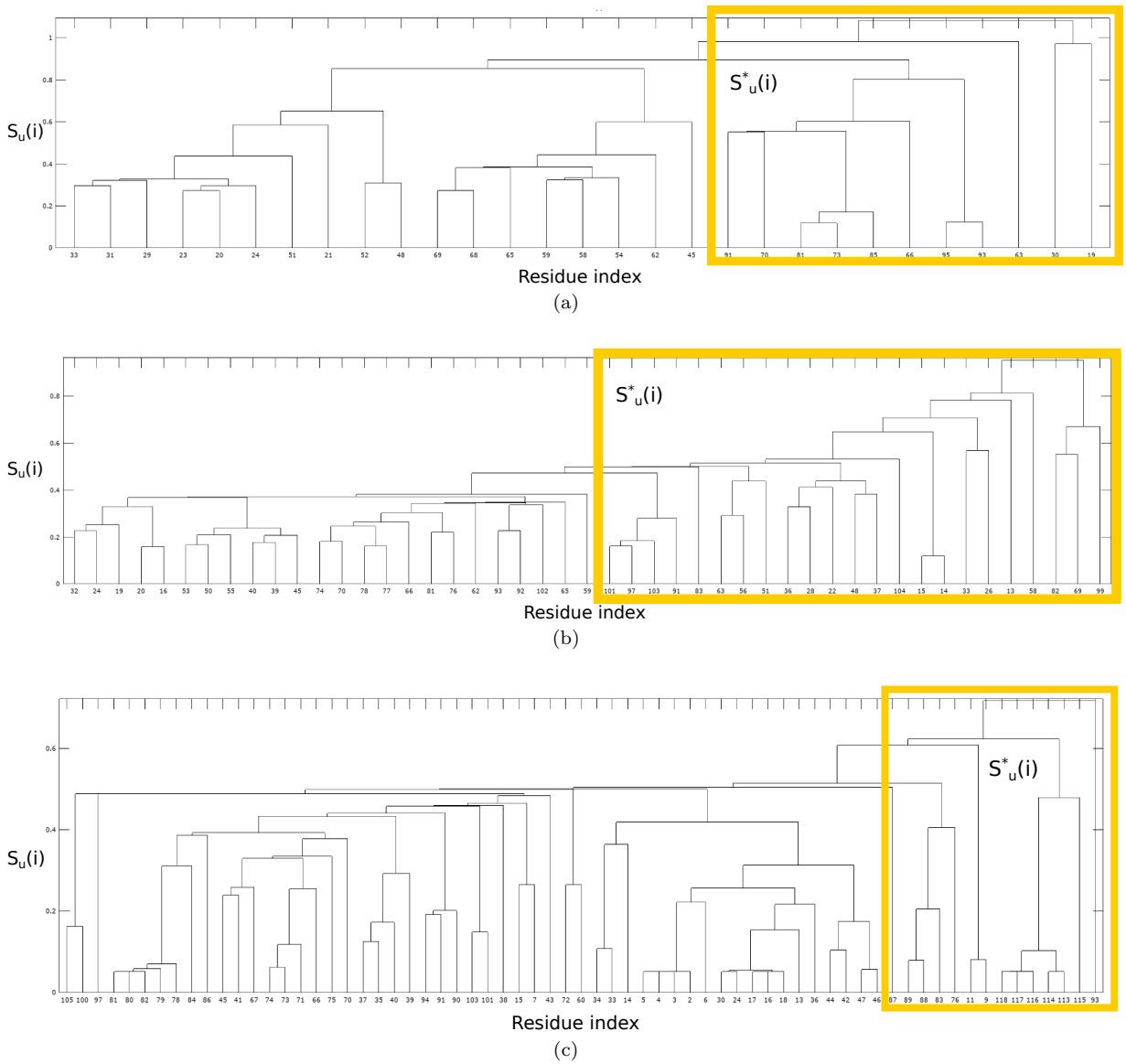


FIG. 8: Dendrogram of the entropy $S_u(i)$ calculated on the natural sequences for: a) the PDZ, b) FKBP and c) Response Regulator domains. In yellow we have highlighted $S_c^*(i) = S_{NAT}$, that is the outliers that do not belong to the largest cluster.