Article

Protein Science
DOI10.1002/pro.3996

Structural motifs in protein cores and at protein-protein interfaces are different

Anna Hadarovich[1,2], Devlina Chakravarty[1,3], Alexander V. Tuzikov[2], Nir Ben-Tal[4], Petras J. Kundrotas[1], and Ilya A. Vakser[1,5]

[1] Computational Biology Program, The University of Kansas, Lawrence, KS 66047, USA

[2] United Institute of Informatics Problems, National Academy of Sciences, 220012 Minsk, Belarus

[3] Department of Chemistry, Rutgers University, Camden, NJ 08102, USA

[4] Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Israel

[5] Department of Molecular Biosciences, The University of Kansas, Lawrence, KS 66045, USA

Correspondence: Ilya Vakser, vakser@ku.edu; Petras Kundrotas, pkundro@ku.edu.

Running title: Structural motifs in protein core vs interfaces

Received: Jul 17, 2020;Revised: Oct 30, 2020;Accepted: Oct 31, 2020

ABSTRACT

Structures of proteins and protein-protein complexes are determined by the same physical principles and thus share a number of similarities. At the same time, there could be differences because in order to function, proteins interact with other molecules, undergo conformations changes, etc., which might impose different restraints on the tertiary vs. quaternary structures. This study focuses on structural properties of protein-protein interfaces in comparison with the protein core, based on the wealth of currently available structural data and new structure-based approaches. The results showed that physicochemical characteristics, such as amino acid composition, residue-residue contact preferences, and hydrophilicity/hydrophobicity distributions, are similar in protein core and protein-protein interfaces. On the other hand, characteristics that reflect the evolutionary pressure, such as structural composition and packing, are largely different. The results provide important insight into fundamental properties of protein structure and function. At the same time, the results contribute to better understanding of the ways to dock proteins. Recent progress in predicting structures of individual proteins follows the advancement of deep learning techniques and new approaches to residue co-evolution data. Protein core could potentially provide large amounts of data for application of the deep learning to docking. However, our results showed that the core motifs are significantly different from those at protein-protein interfaces, and thus may not be directly useful for docking. At the same time, such difference may help to overcome a major obstacle in application of the co-evolutionary data to docking - discrimination of the intra-molecular information not directly relevant to docking.

***Keywords*:** protein recognition, protein docking, protein modeling, structure prediction

***Importance and impact*:** A systematic comparison of structural motifs at protein-protein interfaces with those in the protein core was based on the rapidly growing amount of structural data on proteins and their interactions. The results showed important differences in structural compositions of the interfaces and protein cores. The study contributes to better understanding of the fundamental properties of protein structure and function and has important implications for modeling of protein complexes.

INTRODUCTION

Protein tertiary and quaternary structures should have a lot in common because they are determined by the same physicochemical interactions.[1] However, functional restraints, such as proteins' need to undergo conformational changes and to associate with, and dissociate from, other macromolecules and ligands, may impose different constraints on the tertiary vs. quaternary structures. Earlier studies pointed out such similarities and differences in structural motifs and architecture,[2-4] residue-residue contact preferences,[5-7] including investigations specifically focusing on comparison of protein-protein and intra-chain domain-domain interfaces,[8] and structural diversity of the domain-domain interfaces.[9]

In this study we revisit this problem, inspired by the earlier pioneering structure-based studies,[2-4] by systematically investigating fundamental structural properties of protein-protein interfaces in comparison with the protein core, taking advantage of the multitude of structural data that have become available since then, combined with new structure-based approaches. The rapid growth of PDB allowed us to generate comprehensive datasets that include a broad spectrum of structural motifs extracted from the structure pool, which is by orders of magnitude larger the one used in earlier studies of this subject.

We studied protein-protein interfaces and protein cores from various perspectives, by extracting from protein cores pair-wise structural motifs (pseudo-interfaces or *intra*-faces) and systematically comparing them to protein-protein interfaces, as well as by direct alignment of the interfaces with the core. Our algorithm for generating *intra*-faces allowed us to sample the space of the fragments inside proteins, for the first time making possible a direct structural comparison of the interfaces with the pairwise interface-like structural motifs in the core. Along with the drastically larger amount of diverse structural data available for this study, and new approaches and analysis tools, we used a variety of structural and

physicochemical characteristics for the comprehensive comparison of the protein interfaces with the protein cores from different perspectives.

The results showed that while characteristics that are largely based on physics (e.g. solubility) are similar, properties, such as structural composition and packing, that are related mostly to evolution/function are to a significant degree different. The results provide important insight into fundamental properties of protein structure and function and contribute to better understanding of the ways to model protein assemblies.

METHODS

Datasets

The following datasets were used in our study.

Domains set was used for analysis of structural motifs in the protein core. To exclude from consideration interfaces between protein domains, the core was defined as a single domain. For convenience of terminology in this study, in our definition, the core comprises the whole domain, including its surface. We used the full domain in order not to miss the cases where part of the aligned domain structure is on its surface. Removal of the surface residues would affect the partition of the protein domains into secondary structure elements (SSE), which could make the comparison with the *inter*-faces inconsistent. Our procedure was also able to identify alignments of the *inter*-faces with domain fragments that were partially on the surface. This allowed us to analyze cases which illustrate evolution of the domain fragments.

The dataset consisted of 13,807 representative domains, one per family, at the lowest level of the ECOD database hierarchy[10] (family/F-level, with domains in a family sharing significant sequence similarity). The representative domains were human/expert-selected by

the ECOD team (in case of multiple choices, the selected domain was the one with the best resolution).

Inter-faces were from an existing set of protein interfaces generated in an earlier study based on PDB December 2016 download. Binary interfaces were extracted from 60,945 biological units at 12 Å atom-atom distance cutoff between the interacting monomers. Such relatively large cut-off was used in order to preserve the integrity of SSE in the resulting interface fragments.[11] However, this also led to the inclusion of small non-interface parts and short fragments of the interface loops which could distort the structural alignment. These artifacts were purged by removing short (< 4 residues) fragments. The 4-residues fragments were kept only if all the residues were within 6 Å from the other protein. Finally, interfaces with buried surface area < 200 $Å^2$ per chain were excluded. This resulted in 137,083 interfaces, of which 36,929 were dimers and the rest were higher order oligomers.

The *inter*-faces were clustered by structure similarity according to TM-score,[12] which has values in the 0 to 1 range. Clustering with a stringent TM-score cutoff of 0.9 yielded a set consisting of 51,923 representative *inter*-faces. In addition, for direct comparison with the protein cores, we generated a smaller set of 23,878 representative *inter*-faces by clustering with a TM-score cut-off of 0.6. The representative structures were selected by resolution. If the structural resolution of *inter*-faces was similar, the larger *inter*-face was selected. In case of the same size, the one with the most recent release date was selected. In the resulting set, the *inter*-faces, on average, contained 6.7 SSE elements on one side. To compare SSE composition of the *inter*- and *intra*-faces, for computational efficiency, we used a pre-compiled set of protein *inter*-faces from the DOCKGROUND resource (5,936 *inter*-faces[13]).

Intra-faces were generated to analyze structural motifs inside the protein core. The outline of the procedure to generate the set from the Domains set (see above) is presented

in Figure 1. For SSE assignment, each domain was parsed by DSSP[14] (accessed through Biopython library). The eight states provided by DSSP were grouped into three classes: helix (G, H and I), strand (E and B) and loop (S, T, and C, where C sometimes is a blank space). An undirected graph was generated with SSE as nodes, and the weight of the edge between the nodes equal to the distance between corresponding SSE. The distance between two SSEs was calculated as the average of *n* shortest distances between C^{α} atoms, where *n* is the number of residues in the smaller SSE. The SSE were clustered by the hierarchal clustering algorithm based on the weight of the edges. At each iteration, a new graph was generated with clusters from the previous iteration as nodes. A pair of nodes formed a putative *intra*-face if more than half of the SSE in the smaller cluster/node were connected to the other cluster/node (the connection existed if an SSE was closer than 12 Å from an SSE in the other cluster). Nodes that satisfied this criterion were merged into a single node, so that the *intra*-faces would become parts of a joint/larger *intra*-face (Figure 1). The distances between the nodes were recalculated and the procedure was repeated for the new graph, with each iteration reducing the number of nodes. The procedure was iterated until the number of nodes did not change.

To avoid bias in comparison with the *inter*-faces, the same procedure of filtration (removal of shorter fragments, structure resolution, release date, etc. - see above) was applied. Clustering with a TM-score cutoff of 0.9 yielded a dataset consisting of 22,339 *intra*-faces.

Residue-residue contact frequencies

The residue-residue contact preferences were calculated according to the methodology from our earlier study of protein-protein interfaces.[5] The residues were in contact if the C^{β} -C^{β}

distance (C^{α} for Gly) was < 6 Å. The normalized frequency of residue type i (i = 1, 2, ..., 20) was defined as $W_i = F_i / \sum_i F_i$, where F_i is the number of residues i (in the case of *inter*- and *intra*-face contacts, the residues were required to have at least one contact with any residue across the *inter*- or intra-face, respectively). The normalized number of contacts was calculated as $Q_{ij} = C_{ij} / \sum_{k,l} C_{kl}$, where C_{ij} is the total number of contacts between residue types i and j. The log odds of a contact between residues i and j was based on the ratio of the actual and the expected numbers of contacts $G_{ij} = 10 \log (Q_{ij} / (Wi\ W_j))$.

Procedure for detection of interface-like motifs in protein core

The following procedure was developed to detect interface-like structural motifs in the core of a protein domain. The part of an interface from one protein was aligned to the protein core by TM-align.[15] To maximize the detection of similarity, the TM-score was normalized by the length of the smaller component (the interface or, in rare cases of smaller domains, the domain), and designated as the score of the alignment, mTM-score. For alignments with mTM-score ≥ 0.5 we aimed to look for the other part of the interface in another region in the same domain. Thus, the aligned part of the core was deleted from the domain, and the remaining structure was aligned to the other part of the interface. The first aligned part of the core was deleted to avoid an overlap in case the second part of the interface would be aligned to it. Again, only alignments with mTM-score ≥ 0.5 were kept. The procedure was repeated for the *inter*-face parts in reverse order, and the alignment with the highest mTM-score was selected. Finally, if both parts of the interface were successfully aligned to the same domain, the two aligned parts of the core were extracted and structurally compared with the interfaces by MM-align[12] normalized by the smaller of the two structures. This was

done to ensure that the two core fragments are not only similar to the interface parts separately, but also have similar relative orientation. A threshold of MM-score 0.6 was used to define non-random structural similarity, which involves similar relative orientation of the fragments. All alignments of the structural motifs in the core with the *inter*-faces that satisfied the above conditions were kept. Thus, for a single *inter*-face there could be multiple aligned core motifs.

The procedure was validated on a dataset of 5,936 *inter*-faces from the DOCKGROUND resource (see above). For each *inter*-face from the set, the corresponding full complex was retrieved. The two chains of the complex were combined into one to represent the one-domain scenario. The above procedure was used to align the *inter*-faces from that set to determine if it detects the true *inter*-faces between the combined chains. The true *inter*-faces were consistently detected in all structures.

Prediction of inter-faces based on intra-faces

To evaluate the possibility of predicting protein-protein complexes based on the *intra*-faces, comparative modeling of the *inter*-faces (template-based docking) was performed by the partial structure alignment protocol.[16,17] To build a model of the target protein-protein complex the protocol uses TM-align to perform a systematic search of suitable templates in a library of *inter*-faces (in this study, *intra*-faces). Models with poor alignment to the templates (TM-scores < 0.4) were left out of the prediction pool.[18] Quality of the resulting model was assessed by the ligand C^{α} RMSD, with the receptor (the larger protein in the complex) optimally aligned. A model with the ligand RMSD ≤ 10 Å was considered correct (near-native).[19]

RESULTS AND DISCUSSIONS

Comparison of inter- and intra-faces

The *intra*-faces extracted from the protein core consisted of two structural components (see Methods) for an appropriate comparison with the *inter*-faces. The *intra*-faces were extracted from protein domains, rather than the entire proteins, in order to exclude domain-domain interfaces, which may resemble protein-protein interfaces. Different *intra*-faces were generated from the same protein domain, reflecting various structural configurations and sizes (two examples are shown in Figure 2).

Intra- and *inter*-faces are comparable to each other in size and buried surface area (Supplementary Materials Figure S1). Still, the *intra*-faces were somewhat tighter packed, which is likely due to the tighter packing of the protein core than that of the protein-protein interface (see the discussion of SSE packing below). A likely related observation was the greater presence of loops at the *inter*-faces (Figure S2, confirming earlier reports[3]), which are packed with lower density than the ordered secondary structure elements.

For further analysis, residue-residue contact propensities were calculated for domains, *intra*-faces, and *inter*-faces (see Methods). The patterns of the propensities for domains, *intra*-faces and *inter*-faces (Figure 3) resembled each other overall, although with certain differences. In general, the contacts between residues of the same type (the diagonal elements of the contact matrices) are underrepresented[20] due to combinatorics of the residue connections (Figure S3). Although the *intra*-faces were extracted from the domains, the residue-residue propensities in the *intra*-faces and the domains differ from each other because of all residue contacts in the domains only the ones across the *intra*-face are included in the *intra*-face count (Figure S3b). The preference of residues to connect to the ones of the same type was also stronger in the *inter*-faces than in the *intra*-faces. The likely

reason is the prevalence of homodimers in the *inter*-face set. As noted in earlier studies,[6] there is an evolutionary advantage of favoring pairs of the same residue types between identical chains, since in such pairs conservation of contacts between different residue types requires two neutral/beneficial mutations, whereas the contacts between identical residues need only one.

The preferences for interactions among hydrophobic amino acids for domains was stronger than that for the *intra*- and *inter*-faces, likely due to the tighter packing. One noticeable difference between the *intra*- and *inter*-faces were the histidine contacts. Histidine is more common at *inter*-faces, possibly because of its ability to change its protonation state upon small shifts of the pH values. This property could be useful for *inter*-faces, while not preferable (and thus, underrepresented) inside the protein core.

It is noteworthy that comparison to our previous analysis of pairwise preferences of the amino acids at protein-protein interfaces[5] shows quite a few differences. These could perhaps be attributed to the much larger, yet less redundant, dataset used here.

We also considered comparison of the *intra*-faces separately with obligate and non-obligate (transient) complexes. Due to the multiplicity and diversity of protein-protein interactions in living systems, the obligate/transient distinction is far from straightforward. This naturally limits the reliability of the methods that attempt to separate the two, especially when it comes to high-throughput applications needed for our very large set of 23,878 *inter*-faces. A more reliable approach to divide an *inter*-face set into obligate and transient *inter*-faces is manual curation by examining the existing literature, which obviously limits the size of the set. We calculated the residue-residue statistics for such set of 298 complexes[21] and compared it with the statistics for our *inter*-face and *intra*-face sets. For fair comparison, we applied the same procedure for extraction of the *inter*-faces from complexes to the 298-

complex set as we did for our set (see Methods). It resulted in 122 obligate and 176 transient dimeric interfaces. Due to the stark difference in the size of the full sets, by almost two orders of magnitude, the dispersion of the calculated values for the small set is far greater than that for the large set. Thus, although there are similar patterns in these sets, the visual comparison of the two (Figure S4) is not very useful. The analysis showed that in comparison with the *intra*-faces, the obligate interfaces have more contacts between histidines, but less histidine-proline ones. Obligate interfaces also have different patterns for arginine preferences, with a clear abundance of the arginine contacts with cysteine, glycine, proline and tryptophan. In contrast, in comparison of *intra*-faces and the transient interfaces, there is a clear distinction in the cysteine contacts, such as a larger number of cysteine contacts with positively charged residues and the lack of those with most other amino acids, especially the depletion of the cysteine bridges, in the transient interfaces. On the other hand, the *intra*-faces have more histidine-polar uncharged and negatively charged residue contacts.

The structural similarity of *intra*- and *inter*-faces was determined by MM-align all-to-all structural comparison in a combined dataset of *intra*- and *inter*-faces together, with subsequent hierarchal clustering based on the MM-scores. Figure 4 shows the distribution of clusters depending on the clustering threshold. At a non-random structural similarity threshold of MM-score = 0.6[22], as well as at less stringent MM-score thresholds 0.5 and 0.4, only a few mixed clusters, comprised of both *inter*- and *intra*-faces were present. At MM-score threshold of 0.6, among the non-mixed clusters, 23,910 were *inter*-face only (19,922 at MM-score 0.5 and 14,368 at MM-score 0.4) and 19,299 *intra*-face only (16,205 at MM-score 0.5 and 11,497 at MM-score 0.4). This indicates that structural similarity of *intra*- and *inter*-

faces is rare. Most clusters obtained with the non-random structure similarity threshold were singletons (consisted of a single member; Figure S5).

Analysis of the structures in the mixed clusters revealed that most were simple/trivial, such as: both sides are β-sheets; both sides are α-helices; and simple combinations of β-sheets and α-helices (Figures 5). We further analyzed the SSE composition of the *intra*- and *inter*-faces. For each structure, two SSE (one from each side) with the shortest distance between them were selected, in all possible combinations of the SSE types (Figure S6). Most SSE distance distributions for *intra*- and *inter*-faces were similar. However, the *intra*-face β - β distances were significantly shorter than the ones in the *inter*-faces. We further analyzed the residue composition of such nearest SSE. The overall composition of residues in the *intra*- and *inter*-faces was similar (Figure S7). However, the composition of the nearest SSE (Figure 6) had certain differences, especially in the β - β case: the *intra*-faces had larger representation of some of the hydrophobic residues (Ile, Val and Phe), whereas some of the hydrophilic residues (Asp, Glu, and Gln) were more common in the *inter*-faces. A possible reason for that is a higher hydrophobicity and tighter packing of the protein core compared to the *inter*-faces having looser packing and more loops to perform their function.

We did not perform structural comparison of the obligate and transient *inter*-faces (see above) to protein core separately, because the comparison of the combined *inter*-face set with the core did not yield structural similarities except the trivial ones.

Comparison of inter-faces with protein core

To extend the comparison of structural motifs in protein domains and protein-protein interfaces we performed analysis that bypasses generation of the *intra*-face library and directly compares protein *inter*-faces to protein domains by structural alignment (see

Methods). We found multiple cases where one component of the protein-protein interface was structurally well-aligned to part of a protein core at the domain-domain interface. However, the cases when both parts were identified within the same domain were rare and simple/trivial such as the described above helixes. Thus, overall, the results were in agreement with those obtained from the comparison of *intra*- and *inter*-faces.

One can expect that some structural motifs at the domain-domain interfaces end up inside domains due to domain fusion. However, once inside the domain core they can evolve into motifs that are likely to be found in the core. An example, in Figure 7, illustrates this scenario. In this example, both parts of an interface consisting of β-strands were found within the same domain but in a different relative orientation than at the interface. Interestingly, these parts were closer inside the domain than at the interface, likely due to the evolutionary pressure for tighter packing inside the protein core.

Implications for modeling of protein complexes

The template-based approach for structural modeling of protein complexes relies on experimentally determined complexes of proteins similar to the targets. To detect an appropriate template, a search against a diverse library of protein-protein complexes is performed according to some measure of target/template similarity, which is often based on structure.[16,17,23] Availability of good templates is a major factor affecting the quality of the prediction. Thus, we asked a question: whether it is possible to use *intra*-faces as templates for protein-protein docking (prediction of *inter*-faces).

To answer this question, we utilized template-based docking protocol (see Methods), using *intra*-faces as templates. To assess the docking performance, the success rate was defined as a fraction of targets, for which at least one in top *N* models was correct (see

Methods). The results had an extremely low success rate (1-3%, Figure S8), far smaller than the one based on *inter*-face templates, which are typically in the 40-60% range.[17]

Co-evolution of residues provides structural information by inferring distances between co-evolving residues, propagating beyond the first layer of residue-residue contacts to more distant layers of co-evolutionary relationships. In recent years, there has been major progress in utilizing this information for predicting protein structures, based on new ideas on how to use it, combined with advances in machine learning (deep learning).[24-27] A major obstacle to application of this approach to protein-protein docking is a perceived lack of sufficient amount of data on protein-protein interfaces needed for the deep learning. Thus, it may be tempting to utilize interface-like structural motifs from protein core as a source of such information for docking. However, our study showed that such *intra*-faces are significantly different from the actual interfaces. Thus, a direct application of such data from protein cores to protein docking may not be productive.

On the other hand, the results of our study are promising with regards to the other major problem in application of co-evolutionary information to docking - a lack of clarity on how to distinguish the *inter*-molecular co-evolutionary information, directly related to docking, from the not directly relevant *intra*-molecular one. The recent progress in structure prediction of individual proteins takes into account the covariation of residues that are not necessarily in contact, to predict distances between them, rather than simple contacts. The non-contact covariations are weaker, but many, which compensates for their weakness. In docking, such approach means that one has to go deeper into the protein core to take into account these non-contact co-variations. At that point it becomes difficult to distinguish the *inter*-molecular covariations from the *intra*-molecular ones (also because the *inter*-molecular covariations are

generally weaker than the *intra*-molecular ones). For that matter, the fact that the *intra*-faces are significantly different from the *inter*-faces should be helpful in solving this problem.

CONCLUSIONS

A systematic comparison of structural motifs at protein-protein interfaces and inside protein cores was performed based on various structural alignment strategies and similarity metrics. A set of structural fragments that are conceptually similar to protein-protein interfaces were generated from protein cores (pseudo-interfaces or *intra*-faces) and systematically compared to the protein-protein interfaces. The protein interfaces were also directly aligned with the protein cores for additional analysis. The results showed that the propensities largely based on physical properties (such as charge-based and hydrophilicity/hydrophobicity) are similar at the interfaces and inside the core. On the other hand, the propensities significantly influenced by evolutionary pressure, such as structural composition and packing, are largely different. The results complement earlier studies, by taking advantage of the vast amount of currently available structural data and new analysis tools.

Overall, the study provides insights into fundamental properties of protein structure and interaction. At the same time, it contributes to better understanding of the ways to model protein complexes. Recent progress in predicting protein structures to a significant extent has been based on the ability to utilize residue co-evolution data and the advance of the deep learning techniques [24-27]. A similar development in the modeling of protein complexes (protein docking) has not occurred because of the supposed lack of data on protein-protein interfaces required for the deep learning. Structural motifs in the protein core could potentially complement such data. However, our results showed that these motifs are significantly different from those at protein-protein interfaces. Thus, in this regard, the direct

use of structural information from protein core may not be beneficial for docking. However, the revealed difference between the *intra*- and *inter*-faces may be useful in solving another major obstacle to application of co-evolutionary data in docking - discrimination of the *intra*-molecular information, which is not directly relevant to the prediction of protein complexes.

SUPPLEMENTARY MATERIAL

The file Supplementary Material.pdf contains figures describing additional details of the study. The data and the in-house code to generate the results are available at https://gitlab.ku.edu/vakser-lab-public/intra-faces.

ACKNOWLEDGMENTS

This study was supported by NIH grant R01GM074255 and NSF grants DBI1565107 and DBI1917263. N.B.-T.'s research is supported in part by the Abraham E. Kazan Chair in Structural Biology, Tel Aviv University, and grant 450/16 from the Israel Science Foundation (ISF). The clustering procedure used in the study was written by Ivan Anishchenko. A.H. is grateful to Alexander Kalinouski for helpful discussions and suggestions.

REFERENCES

1. Kessel A, Ben-Tal N. Introduction to proteins: Structure, function and motion. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC (Taylor & Francis Group); 2018.
2. Tsai C-J, Lin SL, Wolfson HJ, Nussinov R. Protein-protein interfaces: Architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. Crit Rev Biochem Mol Biol. 1996;31:127-152.
3. Tsai C-J, Xu D, Nussinov R. Structural motifs at protein-protein interfaces: Protein cores versus two-state and three-state model complexes. Protein Sci. 1997;6:1797-1809.
4. Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O. Architectures and functional coverage of protein–protein interfaces. J Mol Biol. 2008;381:785-802.
5. Glaser F, Steinberg D, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins. 2001;43:89-102.
6. Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Biol. 2003;325:377-387.
7. Bickerton GR, Higueruelo AP, Blundell TL. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: The PICCOLO database. BMC Bioinformatics. 2011;12:313.
8. Jones S, Marin A, Thornton JM. Protein domain interfaces: Characterization and comparison with oligomeric protein interfaces. Protein Eng. 2000;13:77-82.
9. Verma R, Pandit SB. Unraveling the structural landscape of intra-chain domain interfaces: Implication in the evolution of domain-domain interactions. PloS One. 2019;14:e0220336.
10. Cheng H, Schaeffer RD, Liao Y, et al. ECOD: An evolutionary classification of protein domains. PLoS Comp Biol. 2014;10:e1003926.
11. Sinha R, Kundrotas PJ, Vakser IA. Protein docking by the interface structure similarity: How much structure is needed? PloS One. 2012;7:e31349.
12. Mukherjee S, Zhang Y. MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucl Acids Res. 2009;37:e83.
13. Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. Structural templates for comparative protein docking. Proteins. 2015;83:1563–1570.
14. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577-2637.

15. Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score. Nucl Acid Res. 2005;33:2302-2309.
16. Sinha R, Kundrotas PJ, Vakser IA. Docking by structural similarity at protein-protein interfaces. Proteins. 2010;78:3235-3241.
17. Chakravarty D, McElfresh GW, Kundrotas PJ, Vakser IA. How to choose templates for modeling of protein complexes: Insights from benchmarking template-based docking. Proteins. 2020;88:1070-1081.
18. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci USA. 2012;109:9438–9441.
19. Hunjan J, Tovchigrechko A, Gao Y, Vakser IA. The size of the intermolecular energy funnel in protein-protein interactions. Proteins. 2008;72:344–352.
20. Jha AN, Vishveshwara S, Banavar JR. Amino acid interaction preferences in proteins. Protein Sci. 2010;19:603—616.
21. Soner S, Ozbek P, Garzon JI, Ben-Tal N, Haliloglu T. DynaFace: Discrimination between obligatory and non-obligatory protein-protein interactions based on the complex's dynamics. PLoS Comp Biol. 2015;11:e1004461.
22. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010;26:889-895.
23. Cukuroglu E, Gursoy A, Nussinov R, Keskin O. Non-redundant unique interface structures as templates for modeling protein interactions. PloS One. 2014;9:e86738.
24. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577:706-710.
25. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins. 2019;87:1011-1020.
26. Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D. Origins of coevolution between residues distant in protein 3D structures. Proc Natl Acad Sci USA. 2017;114:9122–9127.
27. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. Science. 2017;355:294-298.

FIGURE LEGENDS

Figure 1. *Pipeline for extraction of intra-faces*.

Figure 2. *Examples of intra-faces.* ECOD domain e1a8dA2 of tetanus neurotoxin protein. The domain is in green, and the two parts of the *intra*-faces are in blue and red. (A) The full domain. (B) A small *intra*-face. (C) A large *intra*-face.

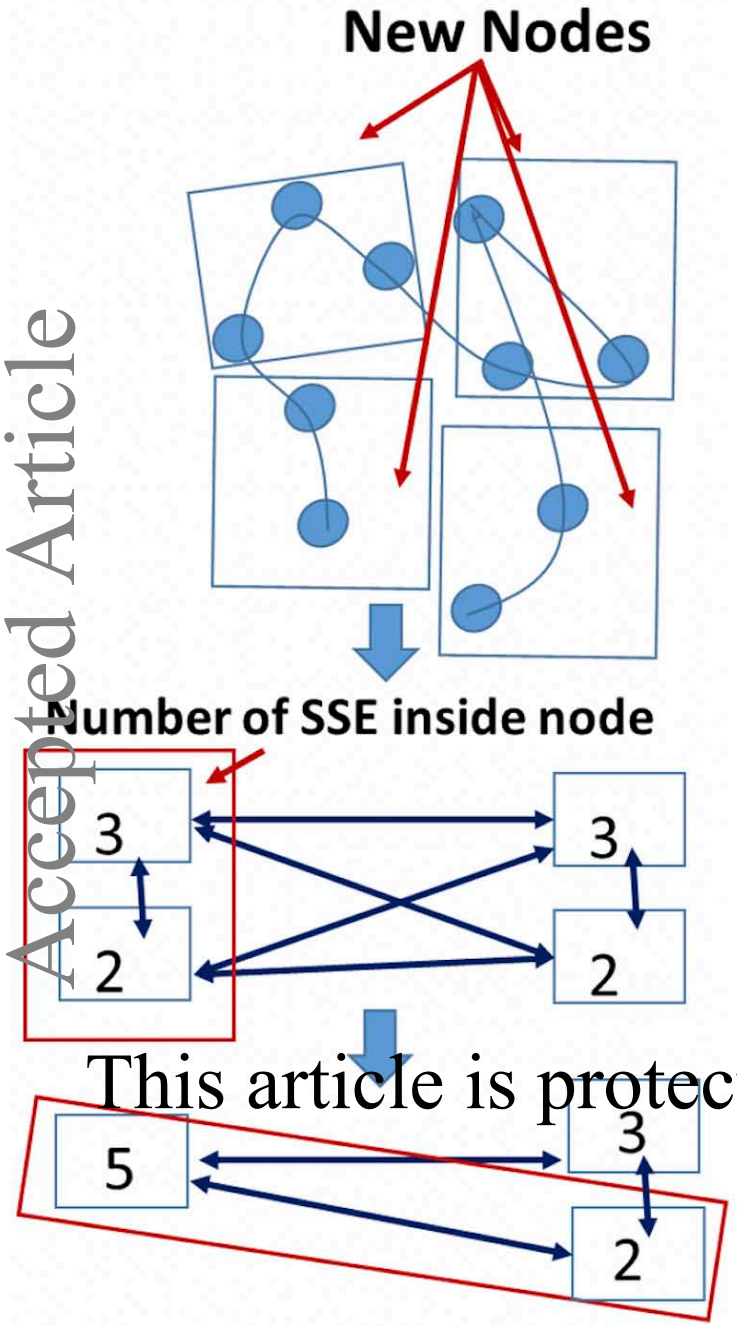
Figure 3. *Amino acids contact preferences.* Color is according to the log odds of the contact (G_{ij}, see Methods) shown by the color bar. (A) Full domains. (B) *Intra*-faces. (C) *Inter*-faces.

Figure 4. *Number of clusters in combined intra- and inter*-face *sets.* The number of mixed clusters, consisting of *intra*- and *inter*-faces, is negligible compared to the number of clusters that include either *intra*- or *inter*-faces only.

Figure 5. *Examples of intra- and inter-faces within one cluster. Inter*-faces are in cyan and green and *intra*-faces are in yellow and magenta. Top panel shows *inter*- and *intra*-faces separate and aligned. (A) β-sheets, (B) α-helices, (C) α/β motifs. Bottom panel shows D. melanogaster Pur-alpha repeat III DNA binding protein 5fgo, with *intra*-face aligned to the *inter*-face between chains.

Figure 6. *Residue composition of the nearest SSE in intra- and inter-faces.* Other combinations (helix-loop and beta-helix) had similar *intra*- and *inter*-face distributions and thus are not shown. Overall, in close contacts there is preference for hydrophobic residues in *intra*-faces, and for titratable and (some of the) polar residues and glycine in *inter*-faces.

Figure 7. *Example of inter-face with both sides found in the same domain.* The *inter*-face parts (green and cyan) are aligned to the domain (red). The alignments are shown (A) separately and (B) together. The relative orientation of the *inter*-face parts aligned to the domain are different than that within the *inter*-face.



For each protein domain:

Build graph of secondary structure elements (node = SSE)

Cluster nodes based on distance between them

Build graph with new nodes (node = cluster of SSE)

Combine pairs of nodes into *intra*-face

Build graph with new nodes (node = combined pairs of old nodes)

Filter