# Protein–protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces

**Elin Teppa, Diego Javier Zea, and Cristina Marino-Buslje***

Bioinformatics Unit, Fundación Instituto Leloir/IIBBA CONICET, Avda. Patricias Argentinas 435, CABA, Argentina

Received 4 May 2017; Accepted 2 October 2017
DOI: 10.1002/pro.3318
Published online 4 October 2017 proteinscience.org

**Abstract: Protein–protein interactions are essential to all aspects of life. Specific interactions result from evolutionary pressure at the interacting interfaces of partner proteins. However, evolutionary pressure is not homogeneous within the interface: for instance, each residue does not contribute equally to the binding energy of the complex. To understand functional differences between residues within the interface, we analyzed their properties in the core and rim regions. Here, we characterized protein interfaces with two evolutionary measures, conservation and coevolution, using a comprehensive dataset of 896 protein complexes. These scores can detect different selection pressures at a given position in a multiple sequence alignment. We also analyzed how the number of interactions in which a residue is involved influences those evolutionary signals. We found that the coevolutionary signal is higher in the interface core than in the interface rim region. Additionally, the difference in coevolution between core and rim regions is comparable to the known difference in conservation between those regions. Considering proteins with multiple interactions, we found that conservation and coevolution increase with the number of different interfaces in which a residue is involved, suggesting that more constraints (i.e., a residue that must satisfy a greater number of interactions) allow fewer sequence changes at those positions, resulting in higher conservation and coevolution values. These findings shed light on the evolution of protein interfaces and provide information useful for identifying protein interfaces and predicting protein–protein interactions.**

**Short Abstract: The objective of this study is to characterize the coevolutionary signal at the interfaces of protein–protein interactions. We used a large data set of protein complexes, finding that the coevolutionary and conservation signals are higher at the interface core than at the interface rim region. We also found that conservation and coevolution increase with the number of different interfaces in which a residue is involved. These findings shed light on the evolution of protein interfaces.**

**Keywords: protein–protein interaction; binding interface; interface core; interface rim; conservation; coevolution; mutual information**

## Introduction

Protein–protein interactions are involved in most cellular processes and play important roles in cell function, both in health and disease. The interaction interface is typically defined as the region composed by those residues that decrease their accessible solvent area (ASA) upon complex formation, while interacting residues can be defined as those residues

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Consejo Nacional de Investigaciones Científicas y Técnicas; Grant number: PIP 1057.

**Correspondence to:* [Cristina Marino-Buslje, Mailing address: Fundación Instituto Leloir, Avda. Patricias Argentinas 435, CABA, C1405BWE, Argentina]. E-mail: cmb@leloir.org.ar

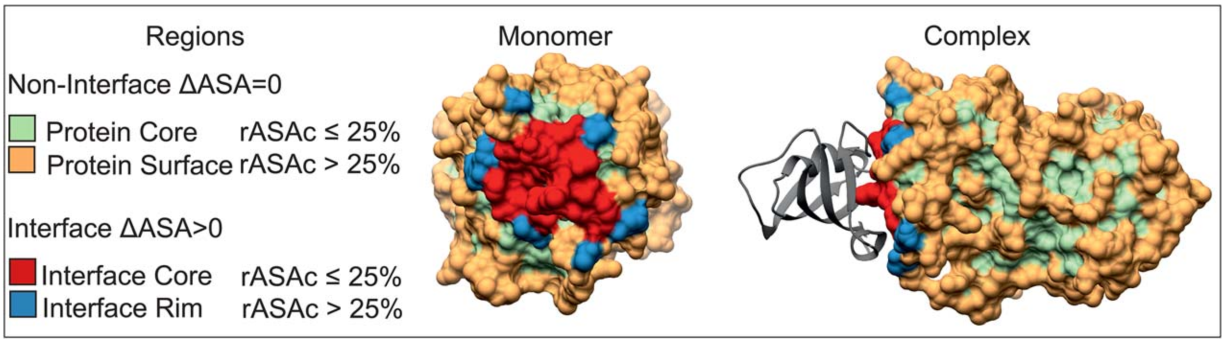
Regions
Monomer
Complex
Non-Interface ΔASA=0
🟩 Protein Core rASAc ≤ 25%
🟧 Protein Surface rASAc > 25%
Interface ΔASA>0
🟥 Interface Core rASAc ≤ 25%
🟦 Interface Rim rASAc > 25%

**Figure 1.** Definition of the four regions in an interacting unit (PDB: 1B6C_B). Where ΔASA is the difference in ASA upon complex formation and rASAc is the relative ASA in the complex. Residues that do not change their solvent accessibility upon complex formation (ΔASA = 0) are assigned to Protein Core (PC) or Protein Surface (PS) based on their rASAc. Residues that change their ASA upon binding (ΔASA > 0) belongs to Interface Core (IC) if are buried in the complex (rASAc ≤ 25%); or Interface Rim (IR) if they retain some solvent accessibility in the complex (rASAc > 25%). The four regions were colored in the chain B of the complex depicted in surface representation. The interacting region (composed of the core and rim) is shown in a front view (left) and side view (right) with the interaction partner in gray cartoon representation. Molecular graphics were performed with the UCSF Chimera package.[30]

involved in physical contact between two or more protein chains. Interface residues are expected to be subject to different evolutionary pressures, as they are heterogeneous in their roles and physicochemical characteristics.[1] The study of protein–protein interactions from an evolutionary perspective is challenging, since it is difficult to distinguish evolutionary constraints due to protein structure and function preservation from those that arise due to protein–protein interactions.

It has been demonstrated that residue conservation and amino acid composition only weakly distinguish the interacting interface from the rest of the surface; however, differences can be found by splitting the interfaces into core and rim regions.[2] The core-rim model has been proposed[3–5] whereby the core is defined as being composed of buried residues and is surrounded by a rim of residues that retain some solvent accessibility upon complex formation (Fig. 1).

The relevance of dissecting the protein interface into core and rim regions has been verified via analysis of human single amino acid variations in interacting proteins,[6] where it has been demonstrated that disease-causing mutations are preferentially located within the interface core rather than the rim.

Residue conservation in the core and rim regions has been explored by Guharoy et al. using Shannon entropy with a reduced alphabet of seven amino acids. They concluded that the mean entropy of the core is smaller (i.e., more conserved) than the corresponding value of the rim region in 68% of biological complexes.[7] Similar results were reported using the Voronoi description of protein–protein interfaces.[8]

A different kind of evolutionary information is given by amino acid coevolution, also called covariation or correlated mutation. The idea behind molecular coevolution is that whenever a functionally or structurally important residue changes, a compensatory mutation occurs elsewhere in the protein to preserve or restore activity. In general, coevolution analysis has proven useful in predicting 3D contact[9,10] and functionally important sites such as catalytic residues,[11] protein sectors,[12–14] and allosteric and ligand binding sites.[15] It has been demonstrated that catalytic residues have a network of residues with high mutual information (MI) in their structural proximity and this evolutionary signature allows its detection.[11] In that study, the cumulative mutual information (cMI) concept was introduced, a per-residue score that measures the degree of shared mutual information. The results suggest that other functionally important sites, such as interacting residues, could also be part of a given mutual information rich region.[11]

Molecular coevolution has never been explicitly analyzed under the core-rim model. Lovell and Robertson found that the interface core is conserved and suggested that, if present, coevolution must be sited in the rim.[16] This idea is to some extent supported by the study of Kann et al., which shows that residues spatially surrounding the binding site (binding neighborhoods) of interacting proteins are subject to stronger interprotein coevolution than the same number of randomly selected residues outside the binding neighborhood.[17] However, intraprotein coevolution at the interface has never before been studied. In the present study, we analyzed conservation and coevolution under the core-rim model using a large-scale dataset, and focused on answering the following questions: how is the coevolution signal distributed in the interface, and how strong is it? And how is it compared with the conservation signal? To answer those questions, we compared the level of coevolution with the conservation signal in the core and rim regions.

We also evaluated conservation and coevolution values depending on the number of interfaces in

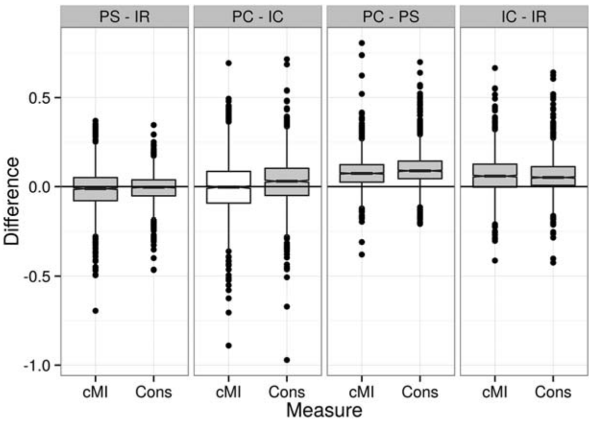

**Figure 2.** Distribution of the differences of conservation and cMI scores between regions. At the top of the boxplot are indicated the compared regions. PS: Protein Surface; PC: Protein Core; IC: Interface Core; IR: Interface Rim. Cons mean conservation and cMI means cumulative Mutual Information. A Wilcoxon rank sum test under the null hypothesis of median is equal to 0 was performed for each comparison, boxplot filled in gray indicate that the *P* value is smaller than 0.001 after a FDR correction for multiple testing.

which a residue is involved. We believe that our findings provide a residue-based understanding of evolutionary constraints at the interacting interface.

## Results

### *Interacting residue distribution within the core and rim regions*

The interaction interface is composed of residues in the core region that are buried within the complex, and residues of the rim region that retain some solvent accessibility after binding. However, different solvent accessibility cut-off values have been used in a number of studies to determine whether a residue is buried or exposed. One way to validate the chosen cut-off is to verify that the interacting residues (annotated at an atomic level in PICCOLO) truly belong to the interface.

The distribution of residues in the four regions shows that the core and rim contain almost all the interacting residues (99.42%), but they also contain some residues that are not interacting (Fig. S2). In our dataset, based on our region definition, 54,164 residues belong to the IC region, 17.09% of which are non-interacting, whereas 40,447 residues belong to the IR region, 41.05% of which are non-interacting. In our dataset, an interacting unit (a PDB chain) has on average 238 residues, of which approximately 10% are involved in protein–protein interactions.

### *Comparison of conservation and coevolution scores between regions*

To compare scores between the different regions, we calculated average scores (Conservation and cMI) per region for each protein and calculated the differences between the averages of the two compared regions. If compared regions have similar average scores, the difference will be close to zero. Figure 2 shows the distributions of the differences. The median and median absolute deviation are summarized in Table S1.

We found that the distribution of cMI differences between PS and IR (first box of Fig. 2) has a median close to zero (–0.0116), demonstrating that the coevolutionary signal at the interface rim is similar to that of the protein surface. Conservation is comparable at the interface rim and protein surface (the difference is close to zero, a median of $-0.0037$). The comparison between the protein and interface cores (second box of Fig. 2) shows that the median of cMI and C are also close to zero (–0.0032 and 0.0302, respectively). As expected, the comparison between the protein core and surface shows that cMI and conservation are greater in the former (both differences are $>0$, being 0.0887 and 0.0742, respectively). Finally, the comparison between the interface core and rim (fourth box of Fig. 2) demonstrates that the interface core has higher cMI and conservation scores than the rim (the median value is greater than zero, being 0.0589 and 0.0513, respectively). In general, we found that conservation and cMI scores show similar behavior throughout comparisons between regions.

It is well-known that the number of non-redundant sequences in a MSA is critical for calculating mutual information. In particular, it has been demonstrated that MSAs with at least 400 clusters of 62% identity perform well for contact prediction. In our data set, approximately 50% of the MSAs fulfill that criterion (1129 MSAs). We performed the same analysis shown in Figure 2 with the subset of MSAs with at least 400 clusters of sequences giving the same tendency (summarized in Table S1 and Fig. S3).

We also investigated to what extent residues with high conservation scores are the same as residues with high coevolution scores in the IC. First, we investigated whether the two variables are correlated over all regions. We found that the correlation is very weak (Spearman's rank correlation coefficient: 0.123), indicating that each variable provides different information. Next, we analyzed the score distributions of the interface core residues (32,934 residues in total). Figure 3 shows that 67.48% of IC residues (22,224 residues show low cMI and conservation scores (both scores $< 0.5$), 9.44% (3109 residues) show low cMI and high conservation (cMI $<0.5$ and Cons $\geq 0.5$), and 19.97% (6577 residues) show high cMI and low conservation (cMI $\geq 0.5$ and Cons $< 0.5$). Only 3.09% of residues (1018 residues) have high cMI and conservation (cMI $\geq 0.5$ and Cons $\geq 0.5$). These results demonstrate that the majority of residues in the interface core have low levels of conservation and coevolution, despite being interacting residues. Moreover, the probability of a given residue having both scores $\geq 0.5$ given that at least

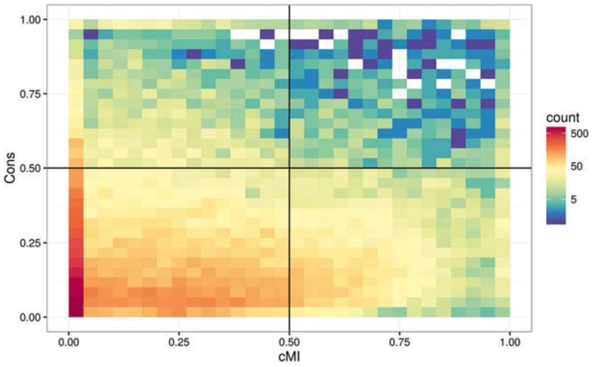

**Figure 3.** Score distribution of the Interface Core residues. Conservation (Cons) and cMI scores were normalized to fall in the range [0–1]. The plot shows the score distribution of Interface Core residues in the dataset (32,934 residues). We considered 0.5 as the threshold between high and low scores (black lines). The majority of points (67.48% residues) shows low scores (cMI $<$0.5 and Cons $<$ 0.5). Only 3.09% of the residues have high cMI and Conservation scores.

one score is $\geq 0.5$, is 0.095. This shows that residues with high conservation and cMI scores are not the same in 90.5% of the cases.

### *Conservation and coevolution of residues involved in several interactions*

Protein chains can have multiple interactions with one or more partners, thus defining different interacting interfaces. Either distinct binding sites may be used for different interactions in a single protein, or the same binding site may be used for all interactions. Similarly, a particular residue can participate in one or several interactions (illustrated in Fig. 4). We analyzed conservation and coevolution scores with respect to the number of interactions in which a protein is involved in and the number of times that a particular residue is annotated as an "interacting residue" in the PICCOLO database. A subset of 322 interacting units with multiple interactions contains 249 proteins that participate in two interactions and 73 involved in three interactions. The number of times a residue was annotated as an interacting residue was also calculated. For proteins having two interactions, there are 5296 residues involved in only one and 6519 residues involved in both interactions. Considering proteins involved in three interactions, there are 1613, 1371, and 1528 residues involved in one, two and three interactions, respectively.

We found that conservation and coevolution scores increase with the number of interactions in which a residue is involved. All comparisons were statistically significant after Mann–Whitney $U$ test ($P$ value $<$ 0.05; Fig. 5).

## Discussion

We employed the core-rim model to study evolutionary signals in four protein regions and to better define their differences. We calculated the level of coevolution in terms of cMI score and conservation as Kullback–Leibler divergence.

The definitions of core and rim regions used in this study are similar to those previously used,[1] where 25% rASAc is used to distinguish between the core and rim. Levy demonstrated that a threshold of 25% rASAc most effectively segregated amino acids in the protein core from those at the surface.[3]

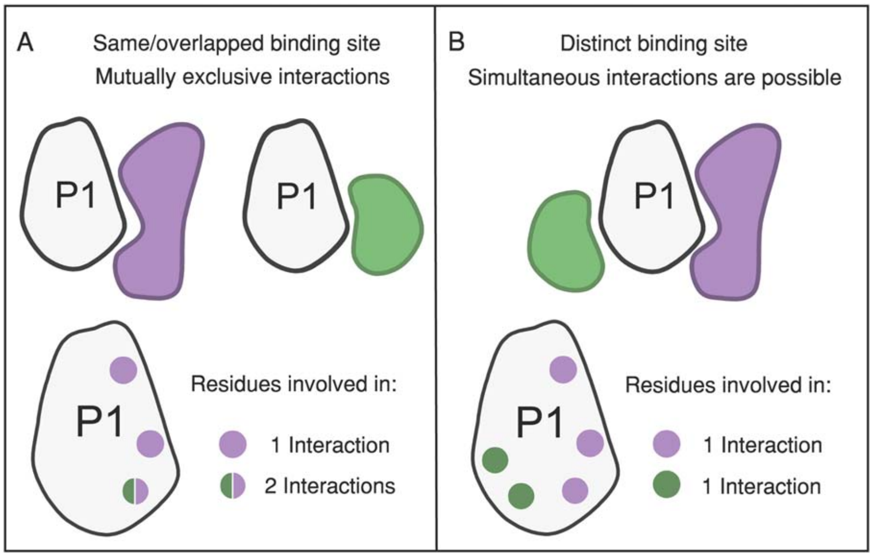

**Figure 4.** Number of interactions in which a residue is involved. Schematic representation of two possible scenarios for a protein involved in two interactions. In the first one (A) the protein P1 uses the same binding site or an overlapped binding site to bind different partners colored in green and violet. The interactions are mutually exclusive, allowing a given residue to participate in both interactions. (B) The protein P1 has two distinct binding sites and both interactions can exist at the same time. Even though the protein participates in two interactions, the residues are involved in only one interaction.

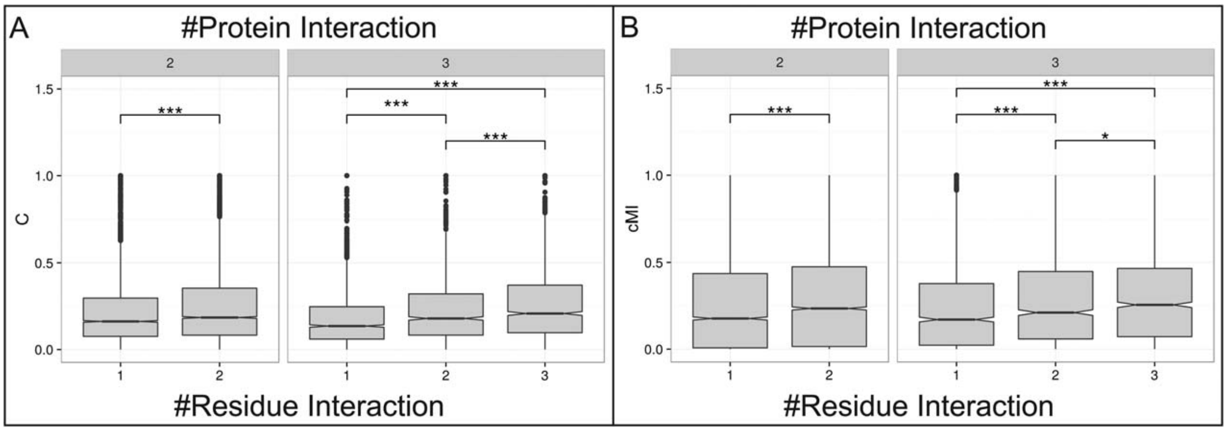

**Figure 5.** Distribution of Conservation and Coevolution according to the number of times a residue is an interacting residue in the subset of proteins with multiple interactions. At the top of the boxplot the number of interactions in which a protein is involved is indicated; at the bottom of the boxplot is shown the number of times a residue is an interacting residue. The distribution were compared using Mann–Whitney *U* test (****P* value < 0.001; **P* value < 0.05). Sample size, median, and median absolute deviation of conservation and cMI scores are shown in Table S2.

We found that residues in the interface core show a higher level of conservation and coevolution than those in the rim. Higher conservation within the core as compared to the rim has also been demonstrated in prior studies.[7] However, a higher level of covariation at the interface core as compared to the rim is a novel finding of this study. This result is somewhat unexpected, since it has been speculated that a coevolutionary signal should be observed in the rim region due to the higher conservation of the core.[16] Here, we demonstrate that both signals coexist at the interface core. This result can be explained by: (i) highly (but not perfectly) conserved residues can yield high levels of cumulative MI; (ii) the interface core can be composed by some residues that are highly conserved and others that are highly coevolved. To distinguish between these options, we calculated the correlation between conservation and cMI score and analyzed the score distribution of the interface core region. We found a very weak correlation between scores and that the interface core residues with high cMI are not the same as those with high conservation scores. These results suggest that the second option is the more likely scenario. We also found that 83% of residues in the interface core are interacting, the majority of which have low levels of conservation and coevolution. Therefore, we expect that none of these scores, without using additional information, would demonstrate good predictive performance for protein–protein interacting residues. It is worth mentioning that the use of other covariation method such as mfDCA[18] or Gremlin[19] may give different results. However the used MI-based method proved to be good at detecting functionally important residues.[11]

The coevolutionary signals in the four regions support the idea that the protein core is similar to the interface core, and the protein surface is similar to the interface rim. This similarity between regions has been reported based on amino acid composition and conservation;[7] this work adds that coevolution also follows the same trend.

It has been reported that interface residues are more conserved than other surface residues. This difference is more evident when the classification of residues (between the interface and the rest of the surface) takes into account the multiple interactions a protein can possess.[20] In this study, we demonstrate that the increase in net conservation occurs specifically due to the conservation of residues involved in interactions (not to all the interface residues). Furthermore, this conservation increases with the number of interactions in which each residue is involved. We also demonstrate that the level of coevolution is greater in those residues that are involved in multiple interactions; to our knowledge, this relationship has not been previously reported.

As a general conclusion, considering conservation and coevolution, our results demonstrate that when more constraints are present at a given position (a residue has to satisfy multiple interactions), fewer sequence changes are allowed resulting in higher residue conservation and coevolution values.

It is worth mentioning that the number of interactions that a protein has can be underestimated, as we only analyzed those complexes with solved structures. Another possible approach to estimate the number of interactions is to extract the information from specialized databases (i.e., String, IntAct Molecular Interaction Database) that integrate experimental information from different experimental systems such as co-purification, yeast two-hybrid experiment, or genetic interaction.[21,22] However, it is not possible to identify the interacting residues using these methods, and the distinction between redundant and non-redundant interfaces that avoid sampling bias would not be possible.

We believe that our findings contribute to a better understanding of interface evolution and provide information that can be used to develop methods for predicting protein–protein interactions.

## Materials and Methods

### Dataset of non-redundant interacting interfaces

We used the PICCOLO database[23] of non-redundant protein–protein interactions at atomic level, the version build with PISA generated assemblies,[24] which are more likely to reflect the biologically relevant oligomeric assembly. This database defines the interaction unit as the polypeptide PDB chain. All PDB entries containing more than one polypeptide chain were identified and every unique pair of non-identical chains was examined, that is, for a PDB entry with four chains A, B, C, and D, six comparisons are performed (AB, AC, AD, BC, BD, and CD). The database also provides annotations of the interacting residues, taking into account the physicochemical properties of the interaction at atomic level. In particular, distance and angle terms were used to distinguish 12 different interaction types, including van der Waals contacts, hydrogen bonds and hydrophobic contacts. It is worth noting that this definition provides a more specific set of interactions than a simple distance cut-off.

All the interfaces in PICCOLO are first grouped by the unique combination of UniProt identifiers of both components, then all the UniProt pairs having >75% of residue in common at both sides of the interfaces are clustered. The highest quality structure for each cluster was chosen as the representative complex. We leave out of the analysis the homocomplexes as it would not be possible to distinguish whether the covariation between two positions is due to intra or interprotein cause. We end up with a set of 1792 PDBs that form binary complexes. The corresponding set of 'sanitized' PDB files was provided by PICCOLO's authors. The structures included in the analysis cover the four CATH classes:[25] 29% Mainly Alpha (four architectures), 31% Mainly Beta (eight architectures), 37% Alpha Beta (nine architectures), and 3% Few Secondary Structures (one architecture).

The number of different interactions in which a protein is involved was estimated as the number of times a UniProt accession appears in the non-redundant PICCOLO database. We found 1336 proteins involved in one interaction, 249 proteins involved in two interactions, and 73 in three interactions.

### Construction of the multiple sequence alignments of homologous proteins

We extracted the sequences of each member of the heterocomplexes (representative structure) of the dataset. For each sequence a search for homologous sequences was carried out using phmmer (http://hmmer.org/) against UniRef 90 database.[26] Multiple sequence alignments (MSAs) were gap trimmed to remove positions with gaps in the reference sequence. We discarded from the analysis: (i) sequence with length <50 residues, (ii) shorter than 50% of the reference sequence (coverage < 50%), and (iii) families with <50 sequences in the MSA. We ended up with a dataset of 2255 MSAs containing 42,284 interacting residues. Next, sequences in each MSA were clustered at 62% identity.

The distribution of the number of clusters is given in Figure S1.

### Scores calculation

Sequence conservation was calculated for each MSA position as the Kullback–Leibler relative entropy[27] using the UniProt background frequency distribution of amino acids (http://www.uniprot.org/)

The MI score was calculated as described in Buslje et al.,[28] that includes APC correction,[29] sequence weighting, low count correction and Z-score permutation.

A derived Mutual Information score per residue was calculated as the cumulative Mutual Information (cMI) that measures the degree of shared mutual information of a given residue.[11]

Shortly, the cMI score for each residue is calculated as the sum of the MI values above a certain threshold where the particular residue appears, as shown in the following equation:

$$cMI_i = \sum_{j, MI_{i,j} > t} MI_{ij}$$

Where $\mathrm{MI}_{i,j}$ is the mutual information value between the positions *i* and *j;* and *t* is the MI threshold. The MI threshold was optimized for contact prediction ($t = 6.5 \pm 2.5$) at a sensitivity of 0.4 and a specificity of 0.95.[28]

For each MSA, the cMI and conservation scores were normalized so that the values fall in the range [–1] as defined in the equation below.

$$S_{\mathrm{norm}} = S - S_{\mathrm{min}} / (S_{\mathrm{max}} - S_{\mathrm{min}})$$

where $S_{\mathrm{norm}}$ is the normalized score, $S$ is the measured score and $S_{\mathrm{min}}$ and $S_{\mathrm{max}}$ are the lowest and largest values in a given MSA.

### Protein region definitions

We dissected each interacting unit into four non-overlapping regions: protein core (PC), protein surface (PS), interface core (IC), and interface rim (IR) according to the relative solvent accessibility in the complex (rASAc) and delta solvent accessibility upon complex formation. Our definition, described below, is based on that used by Levy and Chakrabarti.[3,4]

The **interacting interface** is formed by those residues that change their Accessible Surface Area (ASA) upon complex formation. The interacting interface is divided into interface core and interface rim regions. **Interface rim** residues are those that have $\Delta\mathrm{ASA} > 0$ upon complex formation and $\mathrm{rASAc} > 25\%$ in the complexed state (i.e., they retain certain solvent accessibility in the complex). All other interface residues were assigned to the **interface core** region ($\Delta\mathrm{ASA} > 0$ upon complex formation and rASAc ≤ 25% in the complexed state). **Protein surface** residues are those with ΔASA = 0 upon complex formation and rASAc > 25%. All other residues with ΔASA = 0, were assigned to the **protein core** region (illustrated in Fig. 1).

For a given amino acid, ΔASA upon complex formation and rASAc values were extracted from PICCOLO database. Relative accessibility represents the accessible surface of each residue *X* relative to that observed in an Alanine-*X*-Alanine tripeptide.

## Acknowledgments

E.T., D.J.Z., and C.M.B. are researchers at the Argentinean national research council (CONICET). This project was partially founded by PIP 1057 (CONICET). The authors declare no competing financial interests.

## References

1. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V (2008) Characterization of protein–protein interfaces. Protein J 27:59–70.
2. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface?. Protein Sci 13:190–202.
3. Levy ED (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol 403:660–670.
4. Chakrabarti P, Janin J (2002) Dissecting protein–protein recognition sites. Proteins 47:334–343.
5. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. Proteins 53:708–719.
6. David A, Sternberg MJE (2015) The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. J Mol Biol 427:2886–2898.
7. Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein–protein interfaces. Proc Natl Acad Sci USA 102:15447–15452.
8. Bouvier B, Grünberg R, Nilges M, Cazals F (2009) Shelling the Voronoi interface of protein–protein complexes reveals patterns of residue conservation, dynamics, and composition. Proteins 76:677–692.
9. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 108:E1293–E1301.
10. Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31:999–1006.
11. Marino Buslje C, Teppa E, Di Doménico T, Delfino JM, Nielsen M (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. PLoS Comput Biol 6:e1000978.
12. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–299.
13. Aguilar D, Oliva B, Marino Buslje C (2012) Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features. PLoS One 7:e41430.
14. Rivoire O, Reynolds KA, Ranganathan R (2016) Evolution-based functional decomposition of proteins. PLoS Comput Biol 12:e1004817.
15. Kolesov G, Mirny LA (2009) Using evolutionary information to find specificity-determining and co-evolving residues. Methods Mol Biol 541:421–448.
16. Lovell SC, Robertson DL (2010) An integrated view of molecular coevolution in protein–protein interactions. Mol Biol Evol 27:2567–2575.
17. Kann MG, Shoemaker BA, Panchenko AR, Przytycka TM (2009) Correlated evolution of interacting proteins: looking behind the mirrortree. J Mol Biol 385:91–98.
18. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6:e28766.
19. Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 3:e02030.
20. Choi YS, Yang J-S, Choi Y, Ryu SH, Kim S (2009) Evolutionary conservation in multiple faces of protein interaction. Proteins 77:14–25.
21. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43:D447–D452.
22. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40:D841–D846.
23. Bickerton GR, Higueruelo AP, Blundell TL (2011) Comprehensive, atomic-level characterization of structurally characterized protein–protein interactions: the PICCOLO database. BMC Bioinformat 12:313.
24. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797.
25. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Nicholas Furnham, Roman A Laskowski, David Lee, Jonathan G. Lees, Sonja Lehtinen, Romain A. Studer, Janet Thornton, and Christine A. Orengo. (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381.
26. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926–932.
27. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Statist 22:79–86.
28. Buslje CM, Santos J, Delfino JM, Nielsen M (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 25:1125–1131.
29. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333–340.
30. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. (2004) UCSF Chimera: a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612.