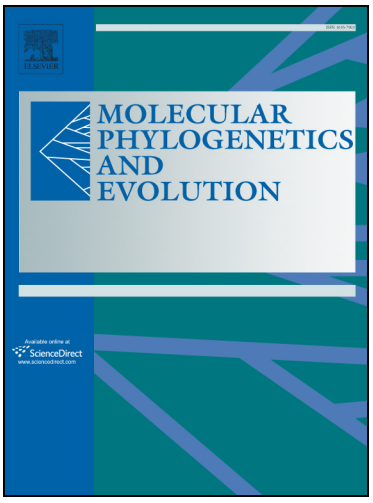


# How is structural divergence related to evolutionary information?

Diego Javier Zea, Alexander Miguel Monzon, Gustavo Parisi, Cristina Marino-Buslje

PII: S1055-7903(17)30618-8
DOI: https://doi.org/10.1016/j.ympev.2018.06.033
Reference: YMPEV 6216

To appear in: *Molecular Phylogenetics and Evolution*

Received Date: 18 August 2017
Revised Date: 1 June 2018
Accepted Date: 19 June 2018

Please cite this article as: Javier Zea, D., Miguel Monzon, A., Parisi, G., Marino-Buslje, C., How is structural divergence related to evolutionary information?, *Molecular Phylogenetics and Evolution* (2018), doi: https://doi.org/10.1016/j.ympev.2018.06.033

# How is structural divergence related to evolutionary information?

Diego Javier Zea[a], Alexander Miguel Monzon[b], Gustavo Parisi[b¶], Cristina Marino-Buslje[a¶*]

[a] Structural Bioinformatics Unit, Fundación Instituto Leloir, CONICET, C1405BWE, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina

[b] Departamento de Ciencia y Tecnología, CONICET, Universidad Nacional de Quilmes, Bernal, Argentina

[*]Corresponding author

E-mail: cmb@leloir.org.ar (CMB)

[¶]These authors contributed equally to this work.

# Abstract

The analysis of evolutionary information in a protein family, such as conservation and covariation, is often linked to its structural information. Multiple sequence alignments of distant homologous sequences are used to measure evolutive variables. Although high structural differences between proteins can be expected in such divergent alignments, most works linking evolutionary and structural information use a single structure ignoring the structural variability inside a protein family.

The goal of this work is to elucidate the relevance of structural divergence when sequence-based measures are integrated with structural information.

We found that inter-residue contacts and solvent accessibility undergo large variations in a protein family. Our results show that high covariation scores tend to reveal residue contacts that are conserved in the family, instead of protein or conformer specific contacts.

We also found that residue accessible surface area shows a high variability between structures of the same family. As a consequence, the mean relative solvent accessibility of multiple structures correlates better with the conservation pattern than the relative solvent accessibility of a single structure.

We conclude that the use of comprehensive structural information allows a more accurate interpretation of the information computed from sequence alignments. Therefore, considering structural divergence would lead to a better understanding of protein function, dynamics, and evolution.



# Keywords

conservation, solvent accessibility, coevolution, structural divergence, conformational diversity

# 1. Introduction

Since Anfinsen's experiments, it is known that protein structural information is codified in the protein sequence (Anfinsen, 1973). However, to derive evolutionary information such as structurally or functionally important positions, populated Multiple Sequence Alignments (MSAs) of homologous proteins are required (Ashkenazy et al., 2016; Marino Buslje et al., 2010; Panchenko et al., 2004).

Conserved positions in an MSA evolve under several constraints according to the neutral theory of molecular evolution. Most evolutionary changes result from mutations with minimal or minor functional impact that are fixed via random genetic drift (Kimura, 1983). However, mutations in biologically important residues (e.g., active sites) could undergo purifying selection or be compensated with other mutations elsewhere in the protein. The concept of position conservation, to denote its importance, was extensively used to infer structural and/or functional roles (Cooper and Brown, 2008; Guharoy and Chakrabarti, 2005; Karlin and Brocchieri, 1996; Lichtarge et al., 1996; Worth et al., 2009).

Another source of evolutionary information in a protein family is the coevolution (de Juan et al., 2013). Since structural and/or functional constraints involving multiple non-independent protein sites can lead to coevolutionary signals, covariation measures are used as a proxy for coevolution (Baldassi et al., 2014; Buslje et al., 2009), and have been applied to predict contacts (Buslje et al., 2009; Jones et al., 2012; Morcos et al., 2011), functional sites (Marino Buslje et al., 2010) and protein–protein interactions (Hopf et al., 2014; Ovchinnikov et al., 2014).

Most works linking evolutionary and structural information use a single structure to represent the structural space of a given protein family. Such information was used to analyse a particular protein of interest, to predict the location of binding and catalytic sites (Capra and Singh, 2007; Liang et al., 2006; Marino Buslje et al., 2010) and to predict tertiary contacts and protein folds (Sutto et al., 2015). It was found that the evolutionary rate of a site or its conservation (used also

as a proxy to the evolutionary rate) correlates with the residue Relative Solvent Accessible Surface Area (RASA) (Echave et al., 2016; Franzosa and Xia, 2009; S.-W. Yeh et al., 2014). Other works highlight the importance of local packing density as the main determinant of the evolutionary rate (S. W. Yeh et al., 2014). Those works use the RASA and contact density of a single structure without taking into account the structural variability inside a protein family. This is even more extreme in the case of covariation scores, where inter-residue contacts are used to optimise their parameters and to assess the performance of the methods (Buslje et al., 2009; Jones et al., 2012; Zea et al., 2016), ignoring that residue contacts may vary between protein conformers and, even more, between different protein structures.

It is well-known that to gain confidence in the covariation scores, MSAs with a high number of divergent homologous sequences are required (Martin et al., 2005; Morcos et al., 2011). So, Pfam alignments are a common choice for measuring these variables (Simonetti et al., 2013). A lot of structural divergence can be expected in such MSAs because the relationship between sequence identity and structural divergence is exponential (Chothia and Lesk, 1986; Flores et al., 1993; Russell et al., 1997). Moreover, structural differences between proteins of the same family (same MSA) may arise not only because of evolutionary divergence due to the accumulation of mutations, but also because of the conformational diversity of a single protein sequence (Monzon et al., 2017). The diference between conformers of the native state of a protein, can be as high as 23.4 Å of Root Mean Square Deviation (RMSD) of alpha carbons (Burra et al., 2009).

In this context, it is interesting to think about how the structural differences in an MSA are reflected in sequence-based evolutionary information such as conservation and coevolution.

In this work, we study how evolutionary information is related to structural information when all the available structures are considered. We have analysed the correlation between conservation and RASA and the performance of the inter-residue contact prediction by covariation methods when different structures from an MSA are considered.

# 2. Materials and methods

## 2.1. Data set construction

We performed all the analysis using ad-hoc scripts in Julia programming language (Bezanson et al., 2017), taking advantage of the MIToS toolkit (Zea et al., 2016). We selected Pfam protein families (Pfam version 30.0) (Finn et al., 2016) with at least 2 proteins with high resolution crystallographic structure (<3 Ångströms resolution) and a well formatted SIFTS file (Velankar et al., 2013). MSA columns belonging to insert or envelope regions were not used to avoid the effect of misalignments (Finn et al., 2010). Also, we filtered structures shorter than 30 residues, covering less than 80% of the MSA aligned columns and with differences between PDB ATOM and UniProt sequences. Additionally, columns with missing residues are not used in the analysis. After applying these criteria to the 16,306 families in Pfam, we ended up with 1,808 protein families suitable for the analysis.

MSAs often suffer from a high degree of sequence redundancy due to sequence database bias towards particular gene families or species. Thus, we employed the Hobohm I algorithm (Hobohm et al., 1992) to define sequence clusters, and to assign to each sequence within a given cluster a weight corresponding to one divided by the number of sequences in the cluster. Clusters were defined at a sequence identity threshold of 62% (Shackelford and Karplus, 2007) as in previous works (Buslje et al., 2009; Nielsen et al., 2004). To ensure a good sampling of the structural space of an MSA, we used the subset of families with at least 4 sequence clusters with at least one structure each. The minimum value of four clusters was a trade-off between retaining a sufficient number of families for the study and having enough structural information. That value allows the mean C-alpha RMSD values to be almost independent of the number of sequence clusters with structures (see supplementary information section 3 and Fig. S.1).

After applying all these filters, the final dataset has 817 Pfam families. This dataset was randomly divided into an exploration set (245 families) and a testing or confirmatory set (572 families) in order to avoid post hoc theorizing (Leung, 2011). The confirmatory dataset has a good coverage of the protein fold space using CATH classes and architectures (Sillitoe et al., 2015) (see supplementary information section 2.1). All the exploratory analysis had been performed on the exploration dataset. We later tested our hypothesis and validated our observations in the confirmatory subset. All the values and figures of this work comes from the confirmatory dataset. Data for each protein family and structure is available in the supplementary file Dataset.zip.

## 2.2. Structural and evolutionary analysis

We estimated the evolutionary variability of each column in an MSA using two measures from the information theory as implemented in Information module of MIToS (Zea et al., 2016): Kullback-Leibler divergence (KL) (Johansson and Toh, 2010) using BLOSUM62 matrix (Capra and Singh, 2007) and Shannon entropy (EN) (Durbin et al., 1998) (See section 1.2.1 in supplementary material).

Covariation scores between MSA positions were calculated with the Information module of MIToS (Buslje et al., 2009; Zea et al., 2016). We calculated mutual information scores (ZMIp) between column pairs as described in Buslje et. al. 2009 (Buslje et al., 2009; Zea et al., 2016) and direct information scores using GaussDCA (Baldassi et al., 2014) (see section 1.2.2 in supplementary material).

The RASA was calculated using NACCESS with default parameters. Briefly, as the solvent accessibility percent of a residue in the structure compared to the solvent accessibility of that residue type in an extended ALA-x-ALA tripeptide (Hubbard and Thornton, 1993). We considered a residue as exposed if its RASA is greater than 20% and buried otherwise

(Holbrook et al., 1990). Different RASA cutoffs were also explored (see section 4 in supplementary material). Other structural measures per position as Contact Number (CN) and Weighted Contact Number (WCN) were evaluated. Contacting residues were defined as those having any heavy atom at a distance <= 6.05 Ångströmg (Bickerton et al., 2011). The RMSD and the Root Mean Square Fluctuation (RMSF) values between any two structures within a protein family were calculated using the MSA as a guide for a rigid structural superimposition with the Kabsch algorithm (Kabsch, 1978, 1976) (see section 1.3 in supplementary material). Structures were hierarchically clustered at 0.4 Ångströms (Burra et al., 2009) with the complete-linkage algorithm (Olson, 1995) to reduce redundancy, and every structure was weighted as one over the number of structures in the cluster (detailed in the section 1.1 of supplementary material). We have used this cutoff since it is the experimental error in X-ray crystallography. We calculated the probability of two residues of being in contact as the number of structures where this pair of residues are in contact, divided by the total number of structures in the MSA. Similarly, for each residue we also calculated the probability of being exposed in the MSA as the number of structures where the residue is exposed divided by the total number of structures in the MSA. For one-structure calculations, we selected as reference the PDB of the sequence with the best E value against the HMM profile of the family.

## 2.3. Fraction of explained variance

Using the exploratory dataset of 245 Pfam families, we looked for linear correlations between the KL divergence and the RASA of a single structure of the MSA compared to the weighted mean RASA of each MSA column and the weighted probability of a given position of the MSA to be exposed. The last two variables include information from all the structures in the MSA. Each structure is weighted as the inverse of the number of structures in its cluster to avoid the effect of structural redundancy (Williamson et al., 2003).

To look for a linear correlation inside protein families between the three variables related to solvent exposure and KL, we measured how many protein families validates the linear model assumptions. The weighted mean RASA and exposure probability are not linearly correlated without transformations because only 6 families validate the linear assumptions. Applying the function log(x+0.05) to transform the KL values to logarithmic scale, 63 out of 245 families validate all the model assumptions (Peña and Slate, 2006) for the three variable pairs.

# 3. Results

## 3.1. Dataset description

MSAs of the confirmatory dataset (572 families) have 9730 sequences and a mean sequence identity of 26.76%, on average. The median of the mean percent sequence identity in the MSAs is 25.5%, with the first and third quartiles of 21.6% and 30.8%, respectively. The number of sequences and mean percent identity distributions are shown in supplementary Fig. S.2. The number of sequence clusters at 62% identity per family have a median of 1421, with the 50% of the families having between 653 and 2913.5 sequence clusters in the confirmatory dataset (see section 2.2. of supplementary material).

## 3.2. On structural divergence, residue exposure and contacts of different structures within a protein family.

We compared all against all structures within each protein family and found large structural variations. In the confirmatory dataset, the average value of the maximum RMSD between structures per family is 5.65 Å. The mean RMSD per family is 2.48 Å, on average.

Structural divergence increases as sequence identity decreases as shown in supplementary Fig.

S.4. In the confirmatory dataset, the Spearman rank correlation coefficient (rho) between the sequence identity and the RMSD comparing all possible structure pairs within each protein family is -0.80. A reasonable expectation is that the more sequence divergence in the alignment, the more structural divergence. The structure divergence of an MSA can be described by the mean RMSD between its structures. Similarly, its sequence divergence can be described by the mean sequence identity between its sequences. Supplementary Fig. S.5. shows the distribution of these variables in the confirmatory dataset. The Spearman's rho between these two variables is -0.48, which agreed with our expectation.

As a consequence of the structural divergence inside a protein family, the average mean change in the RASA per position between structures is 37.39% and 50.57% of MSA columns changed their exposed/buried status on average (see supplementary information section 4). Furthermore, on average, 53.32% of the column pairs that are in physical contact in at least one structure, are not in contact in another structure from the same MSA. These results show that the RASA and residue contacts change considerably between different structures of the same protein family. It is important to note that while the conformation of the backbone is reasonably well represented with four or more sequence clusters with known structures, this may not be enough to represent lateral chain changes between divergent structures (see supplementary information section 3).

## 3.3. Correlation between residue solvent accessibility and conservation

For each MSA we calculated the correlations between the conservation (KL) and the RASA values of each protein structure. We observed that the Spearman's rho between these two variables varies over a wide range, from -0.82 to 0.23, with a mean of -0.37 (see distribution in Fig. 1A) and being -0.37 also the mean of the mean rho value per family. This means that for some structures, the RASA is strongly and negatively correlated with their family conservation

while for others, it is not correlated or is even positively correlated. Indeed, the difference between the highest and the lowest correlation coefficient values (Δrho) within a family varies from 0.02 to 0.76, with a mean of 0.21 (Fig. 1B). This means that in some families, different structures show very similar correlations between the RASA and conservation (i.e. the difference is close to 0.02), whilst in other families, different structures may show very different correlation values (i.e the difference is close 0.74). The mean change of 0.21 can seem small, but is proportionally important (57%) relative to the mean correlation value of -0.37. The same was done with the Shannon entropy and results are consistent (see supplementary information section 6.1). In conclusion, the correlation value between conservation and solvent accessibility depends on the selected structure.

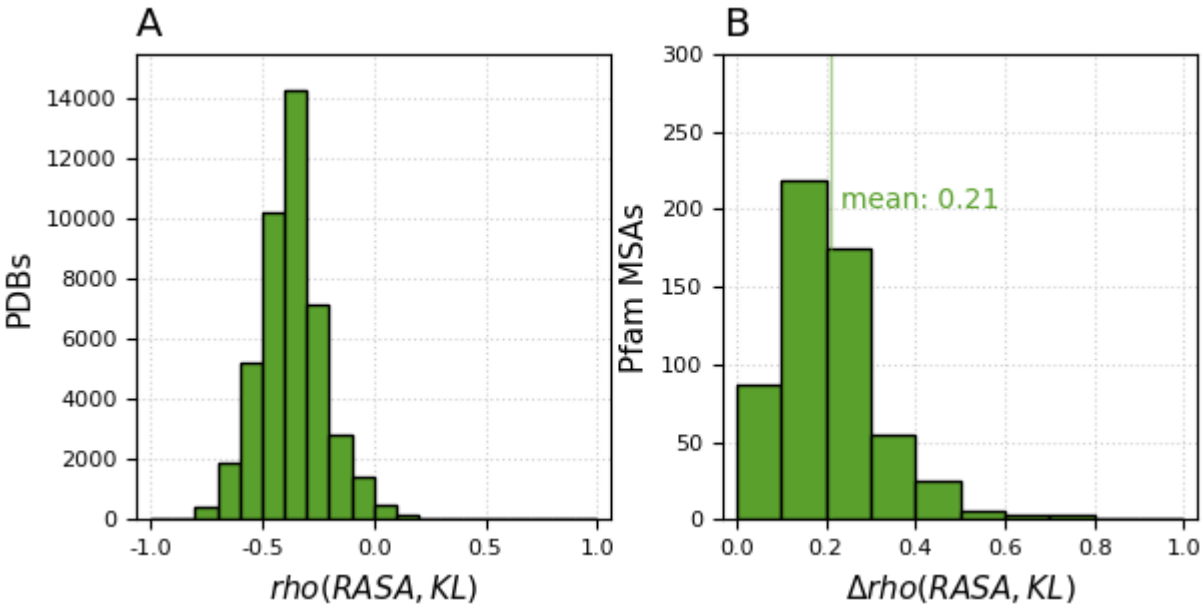


**Fig. 1. Spearman rank correlation coefficient between solvent accessibility (RASA) and conservation (KL). A)** For each PDB structure in our dataset we estimated the Spearman's rho between the residue RASA and the conservation of its MSA column. The histogram shows the number of structures per bin of rho. **B)** A protein family has many values of rho as available PDBs. This histogram shows the number of Pfam MSAs with a particular rho change (measuring the change as the difference between the highest and the lowest rho within the family).

## 3.4. Can more structures explain more of the conservation variance?

We then investigated whether the conservation pattern is better explained taking into account the solvent accessibility from several structures rather than from a single one. We used two scores to measure the solvent exposure of an MSA column taking into account all the available structures. These scores are the weighted mean RASA value and the weighted probability of a position to be exposed (see Materials and methods).

We measured the association strength between variables using the Spearman's rho because a linear correlation can not be assumed *a priori*. The rho between the KL of each MSA column and the RASA of a single structure per family is -0.36 on average (see Fig. 1A). This correlation value is a significantly lower ($P << 0.001$ after a Wilcoxon signed rank test) than the rho of -0.40 between KL and the weighted probability of being exposed or the rho of -0.42 between KL and the weighted mean RASA of the MSA column (Fig. 2). A similar tendency was found using EN instead of KL (see supplementary information section 6.2).

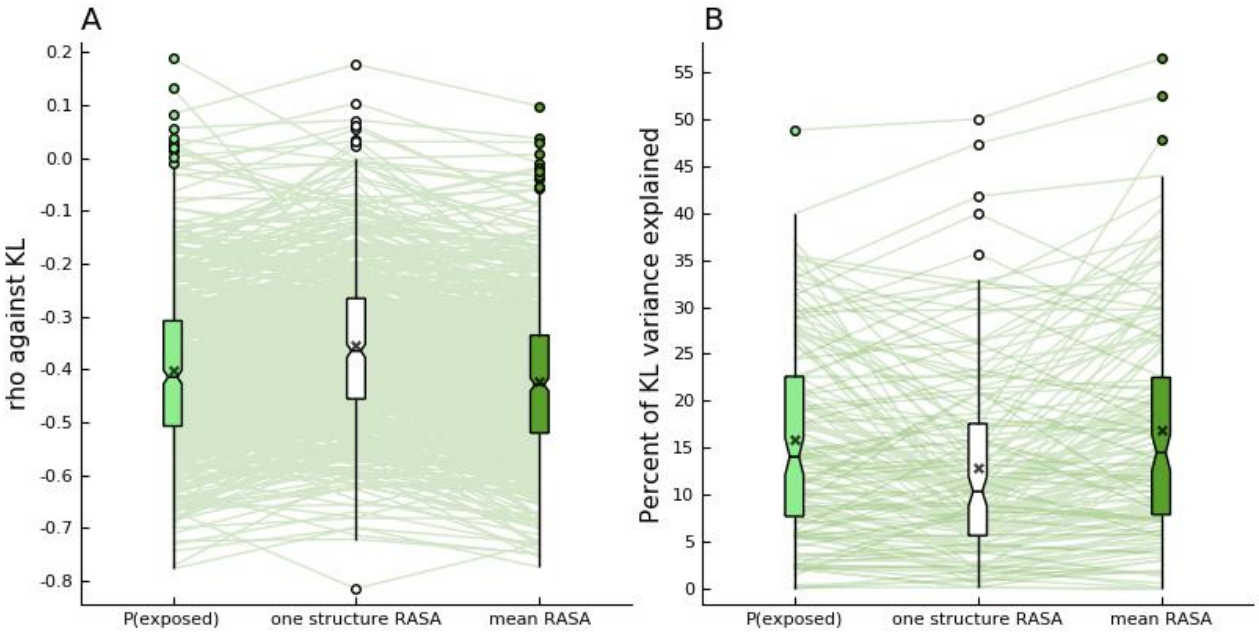


**Fig. 2: KL is better explained using the solvent accessibility of more than one structure.** The boxplots show the distribution of the correlation scores between KL and P(exposed) (light green) or weighted mean RASA (dark green) compared with the distribution of the correlations between KL and the RASA of one structure from the MSA (white). Each green line connects the values of a given protein family. The crosses inside the boxplots indicate the mean values. A) The distribution of the Spearman rank correlation rho show stronger correlation values when more than one structure is used. B) The percentage of KL variance explained is higher when more than one structure are used.

The relationship between conservation and the RASA of a single structure, the weighted mean RASA or the probability of being exposed, is almost linear when the function *log(x+0.05)* is applied to transform KL values. We found that the RASA coming from a single structure for each MSA explains 12.84% of the conservation variance (using the squared value of the Pearson correlation coefficient), on average (median of 10.36%). What's more, the mean percentage of explained variance using the weighted probability of a residue to be exposed is 15.82% (median of 14.07%). However, the variable that explains better the conservation variance is the weighted

mean RASA of each MSA column that explains 16.86% of the variance, on average (median of 14.49%; Fig. 2B). The last two, integrate data from many structures. Therefore, using more than one structure is possible to explain between 2.98% to 4.02% more of the variance, on average. Information is gained by including more structures, even taking into account the information lost due to mean and probability calculations. In fact, the fraction of KL variance explained using the weighted probabilities of being exposed and the weighted mean RASA values is greater than using the RASA of a single structure for 65.07% and 75.34% of the analysed families respectively.

The gain of conservation variance explained using all the structures show a weak correlation with the number of available structures (see section 5 in supplementary material).

## 3.5. Covariation and protein contacts

It is a common practice to evaluate the performance of coevolution methods in terms of their ability to predict protein contacts in a single structure as a reference for a whole alignment (Buslje et al., 2009; Jones et al., 2012). But, how will it perform on different structures having in mind that ~53.32% of the residue contacts are different in proteins within a given protein family?

In order to address this question, we measured the Area Under the ROC Curve (AUC) to predict residue contacts for each structure in our dataset. The mean AUC for contact prediction was 0.74 for ZMIp and 0.76 for DCA. Those AUC distributions are shown in Fig. 3A. Surprisingly, the mean change in the AUC per family, defining change as the difference between the AUC with the PDB that performed best and the one that performed worst, is only 0.029 for ZMIp and 0.031 for Gaussian DCA (shown in Fig. 3B.). Why is there almost no change in the predictive performance if almost half of the contacts change from one structure to the other?

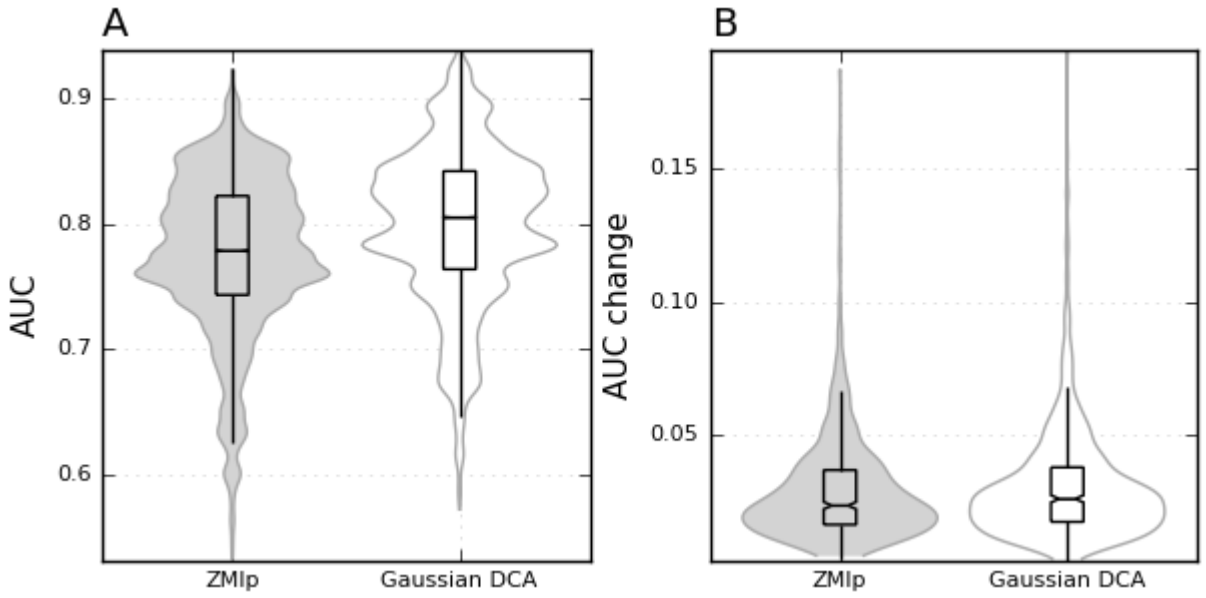


**Fig. 3: AUC for contact prediction by covariation methods. A)** AUC distribution of the totality of protein contacts prediction for all PDB in our dataset. **B)** Distribution of the difference between the higher and the lowest AUC per family.

Subsequently, instead of using a contact matrix for each PDB, we defined contact matrices for each MSA. We considered that a pair of columns are in contact if those positions are in physical contact in at least one structure; otherwise that column pair is considered not in contact. MSA columns in contact are then classified as conserved if they are present in all the structures, and non-conserved or changing contact if the contact is absent in at least one structure.

The AUC to predict conserved contacts was 0.79 for ZMIp and 0.82 for DCA. This indicates a good prediction performance for these methods in predicting conserved contacts through evolution. However, the AUC used to predict contacts specific for some structures (non-conserved) in the MSA is 0.66 for ZMIp and 0.67 for DCA. This indicates a poor performance in the prediction of the conformer, protein or subfamily specific residue contacts. Fig. 4 shows that the AUC to predict conserved contacts is always greater than the AUC to predict the changing ones by ZMIp (Fig. 4A.) and DCA (Fig. 4B.). In fact, conserved contacts reach larger mean

covariation values than changing contacts or residue pairs not in contact as shown in Fig. S.6. Also, Fig. S.3. shows the distribution of ZMIp AUC values for conserved, total and changing contacts as a function of the number of sequence clusters.

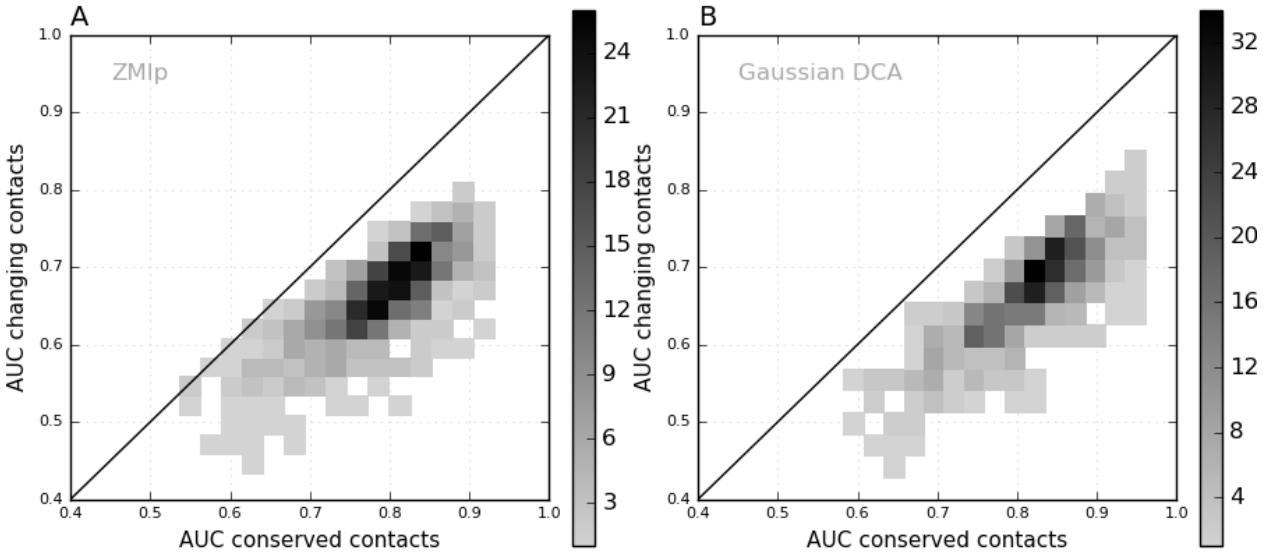


**Fig. 4. Covariation methods predict better the conserved contacts through the MSA.** We calculated the AUC for conserved and changing contacts prediction for each Pfam family in our dataset. Each plot show a bidimensional histogram of the AUC to predict conserved contacts vs the AUC to predict changing contacts. Note that the performance to predict conserved contacts is higher since the cloud of points lies under the line y=x. **A)** Using ZMIp. **B)** Using DCA.

## 3.6. Which is the main structural determinant of residue conservation when multiple structures are used?

Next, we studied the relationships between conservation and structural variables when tested in large Pfam alignments containing multiple and diverse protein structures. We calculated the correlation score between position conservation and structural derived information. To avoid problems derived from non-linear relationships, we used Spearman rank and partial correlation scores. Conservation correlates with a rho of 0.3 with CN and with a rho of 0.33 with WCN,

respectively. Supplementary Fig. S.7 shows the joint distributions between the correlation of KL and RASA and the correlation of KL and CN (panel A) or KL and WCN (panel B). To avoid being misled due to data redundancy, since different families contribute different numbers of structures, we sampled one structure per family 100 times and measured the average of the mean correlation values. The tendency did not change, neither for the correlation between KL and RASA, nor for the correlation between KL and CN or WCN. The average mean correlation values were -0.37, 0.31 and 0.34, respectively. Results are consistent using EN instead of KL (see supplementary material section 6.3).

We found that the RASA correlates better with the conservation than with variables related to local packing density (CN and WCN). In order to know if the correlation between the RASA and conservation could be explained in terms of the local packing density scores, we measured partial correlations between the RASA and KL by controlling for the CN and WCN. Results in Table 1 show that the correlation between the RASA and KL did not disappear after controlling for the CN and WCN; they only show a small decrease (mean decrease was 0.15) in absolute value. Results are consistent when using EN instead of KL (see supplementary Table 1).

Previously, we showed that the weighted mean RASA value across MSA columns showed the best correlation with the conservation of an MSA column (-0.42). In that sense, we tested if the mean CN and WCN also perform better than a randomly selected structure from the MSA. We found a Spearman rho of 0.32 between KL and the weighted mean CN and, 0.35 using the weighted mean WCN. These two correlation values were only 0.01 unit, on average, above the average correlation values from sampling only one random structure for each MSA 100 times. However, the correlation value of -0.37 between the RASA and KL (Table 1) is lower than the -0.43 obtained between the weighted mean RASA and KL. The same tendency is observed with EN instead of KL (see supplementary material section 6.3.). We can conclude from these results that the RASA scores are a better determinant of residue conservation than local packing density scores.

Other research groups highlight the importance of local flexibility as a determinant of evolution rates (Huang et al., 2014). RMSF is an estimation of structural flexibility that takes into account all the available structural information in the MSA. The correlation between KL and the weighted mean RASA per MSA column is -0.42, while the correlation between KL and the RMSF of an MSA column is -0.33. In order to see the authenticity of the last correlation, we calculated the partial correlation between KL and RMSF controlling for the weighted mean RASA. The correlation was 0.16 showing that a considerable part of the correlation between conservation and local flexibility can be explained using the weighted mean RASA.

We hypothesize that solvent exposure is a better determinant of conservation than flexibility because the correlation between conservation and flexibility is lower than the correlation between conservation and the weighted mean RASA. Also, almost a half of the correlation between conservation and flexibility can be explained by the weighted mean RASA.

| | **rho** | **rho controlling for CN** | **rho controlling for WCN** |
|---|---|---|---|
| **RASA vs. KL** | -0.37 (-0.37) | -0.25 (-0.24) | -0.21 (-0.2) |

**Table 1. Spearman correlation scores between RASA and KL, controlling for CN and WCN.** The partial correlation score is calculated as *rho(RASA, KL, column)*. So, the first column is simply the correlation coefficient without controlling for any variable. Numbers between parentheses are the mean correlation scores after sampling 100 time only one structure per family in order to avoid bias by redundancy.

## 3.7. Case example: Structural change of a protein kinase domain

The serine/threonine-protein kinase PIM-1 (UniProt: PIM1_HUMAN) is a proto-oncogene

involved in cell survival and proliferation (Tursynbay et al., 2016). Its biological activity is determined by its protein kinase domain (Pfam: PF00069). We analysed 615 structures (only the columns with solved residues in all structures). To avoid redundancy, the structures were clustered at 0.4 Å of RMSD giving 335 clusters.

The maximum structural divergence for this family was 12.05 Å, achieved between two protein domains with 34.55% sequence identity (see Fig. S.8.). The conformational diversity of its proteins is high, but lower than its structural divergence. For example, we had 80 conformers available for PIM-1, showing a maximum RMSD between conformers of 1.31 Å. The human Aurora kinase A, has the maximum conformational diversity within the family, 8.37 Å RMSD between conformers (see Fig. S.8).

The structural alignment implicit in the Pfam sequence alignment is shown in Fig. 5. The mean RMSD between the structures in our dataset was 3.39 Å. This structural divergence in the C alpha backbone is accompanied by large changes in the residue lateral chain positions. We found that 85.45% of the MSA columns have residues that change their buried or exposed status in at least one structure. Fig. S.9. shows how the PIM-1 4N70 structure residue RASA are related to the weighted mean RASA of each column of the MSA. We also found that 81.26% of protein contacts (where 100% is the number of contacts present in at least one structure), disappeared in at least one structure of the MSA. Fig. S.10. shows the top 5% DCA scores and the probability of each pair of positions being in contact through the alignment. The AUC to predict conserved contacts (pairs with a 1.0 probability of being in contact) by DCA score in this particular family was 0.924, while the score to predict contacts that change through the alignment was 0.742. Fig. S.11. shows the probability of a position pair being in contact as a function of the DCA score. Notably, the top 1% DCA scores pointed out position pairs with a high probability of being in contact.

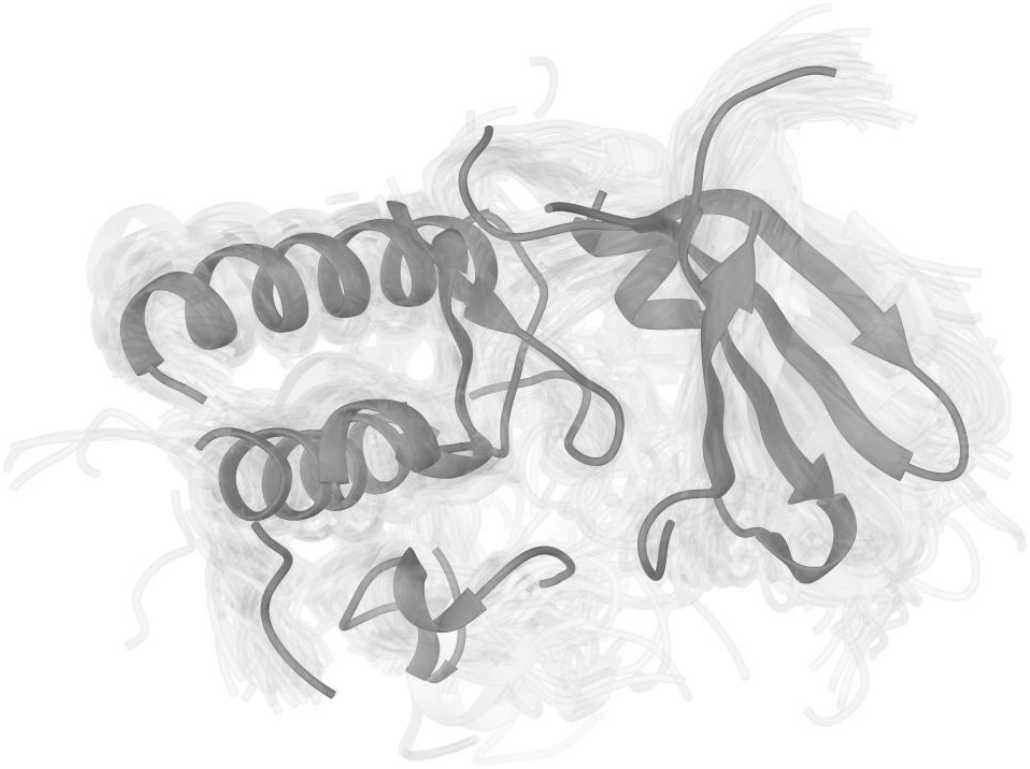

**Fig. 5. Structural superimposition of protein kinase domains.** One PDB for each of the 77 protein domains with known structures in the protein kinase family was selected for this visualization. The structure in dark grey corresponds to the human PIM1 (PDB 4N70). The other structures are shown in light grey.

# 4. Discussion

It is a common practice to map sequence information derived from homologous proteins into one structure taken as a reference for the protein family. However, a number of potential problems occur that lead to the following questions:

1. Can a single conformer of a single protein represent the structural space of the protein family?
2. Is the native state of each protein in the alignment contributing the same structural

constraints during protein evolution?

In this study, we have shed some light on those questions. In particular, we studied how the structural divergence is related to the evolutionary information.

We analysed two types of evolutionary information that could be extracted from a multiple alignment of homologous sequences: conservation and coevolution. Many distant homologous are needed for some of these methods, particularly covariation methods used to estimate coevolution. Consequently, Pfam alignments are a common choice given their high quality and their large number of distantly related homologous sequences.

Pfam alignments show a large structural divergence, rendering high RMSD values when comparing their protein backbones and also, residue contacts and solvent accessibility change considerably between structures. These changes are evident in the case example, where structures have the same fold, nevertheless many structural changes are observed, even at the conformational diversity level.

There are structural constraints to the evolutionary process in order to maintain the native state of proteins (Zea et al., 2013) In particular, residue solvent accessibility has been related to their evolutionary rate; results show that residues on the protein surface tend to vary more than residues in the core (Franzosa and Xia, 2009). Also, residue contact number has been correlated with residue conservation (S. W. Yeh et al., 2014). In addition, inter-residue contacts constrains position pairs variation, leaving a covariation signal (Dunn et al., 2008).

To understand how the structure constrains protein evolution, we measured the correlations between solvent accessibility and conservation. Also, we measured the performance of covariation scores to predict residue contacts, when different structures from the same family are used. We found that the correlation between solvent accessibility and conservation/variation scores greatly changed depending on the selected structure. However, the predictive performance of covariation scores to predict contacts almost did not change between structures because they mainly predict contacts present in all the structures. For each multiple sequence

alignment, we have an unique vector of conservation or variation values for a given protein family. However, for each structure in the family, we have an RASA vector with an RASA value per residue. A correlation value can be measured between the conservation vector and each RASA vector per structure. Previous studies used a single structure to explore this structural feature, so they had a single correlation coefficient per family. The approach of these previous studies can be thought of as stratified sampling, where a single random structure is selected from each stratum (a protein family in this case). In fact, the reported mean correlation values of those studies are very similar to mean correlation values in our dataset using all the available structures.

However, an interesting fact emerge from this study, that is that the correlation values between solvent accessibility and conservation can vary greatly inside a protein family. Actually, positions that are buried in one structure of the family can be totally exposed in another structure. If solvent accessibility determines conservation, we may expect that a measure taking into account all known structures will give a better picture of the structural constraints. Indeed, we found that measures which resumed the solvent accessibility of all known structures are better at explaining the conservation calculated from the multiple sequence alignment.

To capture the solvent accessibility diversity in a protein family in a single vector, we calculated two measures: the probability of a residue of being exposed to the solvent and the mean RASA value of the position. Using these parameters, we found that the conservation variance explained was 16.34% on average, 3.5% above the value found when only one structure is considered per family (12.84%). The small increment in the explained conservation variance may be due to the fact that structural factors are not the main constraints to sequence divergence during evolution (Bloom et al., 2006) or that the structural space of these protein families are still incomplete (Marino-Buslje et al., 2017).

Additionally, we also analysed how other structural variables correlate with conservation. In particular, contact number, weighted contact number and RMSF. These variables, show lower

correlation values than the relative solvent accessibility area, even when multiple structures are used. Therefore, contact density and flexibility variables may be poorer determinants of residue conservation than solvent accessibility in our dataset.

It is known that residue contacts can lead to a covariation signal, i.e. coevolution (Dunn et al., 2008). We observed that many inter-residue contacts changed between structures through evolution and, consequently, we expected predictive performance changes when different structures were used. However, our results show that the predictive performance of the covariation scores remains the same. That can be explained by the fact that these scores predict conserved contacts, i.e. inter-residue contacts common to all the known structures in the alignment. This result is in concordance with recent findings (Rodriguez-Rivas et al., 2016) where evolutionary conserved contacts between proteins were predicted using covariation methods.

As a result of a covariation measure ranking first conserved contacts, it is difficult to predict conformer or subfamily specific contacts (i.e. functional important contacts that change between structures). Considering the large structural space implicit in diverse multiple sequence alignments, could be useful to develop new contact prediction methods oriented to predict them.

We show that residue contacts and solvent accessibility can greatly change, even between conformers of the same protein. Because of this, we believe that evolutionary information derived from protein alignments, even for smaller and curated alignments, deserve to be matched with all known structural information in order to achieve a better understanding of the studied system, particularly when a single protein or protein family is analysed.

When no or little structural information is available, we must bear in mind that conservation and covariation measures take into account the implicit structural space of the protein family. In particular, residue conservation mapped on the surface of a single structure could potentially lead to misinterpretations. In the case of covariation scores, high values tend to indicate conserved residue contacts in the family. Therefore, interpreting these values as predictors of

protein or conformer specific contacts has to be taken with caution.

# 5. Conclusions

It is a common practice to correlate evolutionary information derived from protein family alignments with a single structure taken as reference. We have shown that matching evolutionary information with all the available structural information provides a better understanding of the studied system. This is particularly important when a single protein or protein family is analysed.

We found large structural changes between proteins within a Pfam family at the residue level. In particular, residue solvent exposed area and inter residue contacts change greatly between structures within a protein family. As a result, linking evolutionary information as conservation and coevolution to variables calculated from a single structure, can be misleading. The use of solvent accessibility data from multiple structures permits a better understanding of the conservation pattern than a single structure. Also, solvent accessibility is a better determinant of residue conservation when multiple structures are taken into account, compared to flexibility or residue contacts.

Covariation methods, a proxy to coevolution, are less sensitive to protein or conformer specific residue contacts, they predict with higher performance contacts that are conserved through the alignment (common to all the structures).

In general, our results show that the use of multiple structures from a protein family provide a more comprehensive view of the structural constraints affecting protein evolution.

# Acknowledgements

We thank the members of the Structural Bioinformatics Unit of the Fundación Instituto Leloir and members of the Structural Bioinformatics Group of the Universidad Nacional de Quilmes for their multiple comments and useful discussions. We also thank the anonymous reviewers of that manuscript whose recommendations considerably improved it. All the authors are researcher of the Argentinean National Research Council (CONICET). This work was supported by Universidad Nacional de Quilmes [grant UNQ 1402/15] and CONICET [grant PIP 1087]. We also thank the anonymous reviewers as their exhaustive corrections improved considerably this manuscript.

# References

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181, 223–230.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., Ben-Tal, N., 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 44, W344–W350.

Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., Pagnani, A., 2014. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PLoS One 9, e92721.

Bezanson, J., Edelman, A., Karpinski, S., Shah, V., 2017. Julia: A Fresh Approach to Numerical Computing. SIAM Rev. 59, 65–98.

Bickerton, G.R., Higueruelo, A.P., Blundell, T.L., 2011. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. BMC Bioinformatics 12, 313.

Bloom, J.D., Drummond, D.A., Arnold, F.H., Wilke, C.O., 2006. Structural determinants of the rate of protein evolution in yeast. Mol. Biol. Evol. 23, 1751–1761.

Burra, P.V., Zhang, Y., Godzik, A., Stec, B., 2009. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. Proc. Natl. Acad. Sci. U. S. A. 106, 10505–10510.

Buslje, C.M., Santos, J., Delfino, J.M., 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.

Capra, J.A., Singh, M., 2007. Predicting functionally important residues from sequence conservation. Bioinformatics 23, 1875–1882.

Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in

proteins. EMBO J. 5, 823–826.
Cooper, G.M., Brown, C.D., 2008. Qualifying the relationship between sequence conservation and molecular function. Genome Res. 18, 201–205.
de Juan, D., Pazos, F., Valencia, A., 2013. Emerging methods in protein co-evolution. Nat. Rev. Genet. 14, 249–261.
Dunn, S.D., Wahl, L.M., Gloor, G.B., 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24, 333–340.
Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.
Echave, J., Spielman, S.J., Wilke, C.O., 2016. Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17, 109–121.
Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–85.
Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., 2010. The Pfam protein families database. Nucleic Acids Res. 38, D211–22.
Flores, T.P., Orengo, C.A., Moss, D.S., Thornton, J.M., 1993. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 2, 1811–1826.
Franzosa, E.A., Xia, Y., 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26, 2387–2395.
Guharoy, M., Chakrabarti, P., 2005. Conservation and relative importance of residues across protein-protein interfaces. Proc. Natl. Acad. Sci. U. S. A. 102, 15447–15452.
Hobohm, U., Scharf, M., Schneider, R., Sander, C., 1992. Selection of representative protein data sets. Protein Sci. 1, 409–417.
Holbrook, S.R., Muskal, S.M., Kim, S.H., 1990. Predicting surface exposure of amino acids from protein sequence. Protein Eng. 3, 659–665.
Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M.J.J., Marks, D.S., 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 3. https://doi.org/10.7554/eLife.03430
Huang, T.-T., del Valle Marcos, M.L., Hwang, J.-K., Echave, J., 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol. Biol. 14, 78.
Hubbard, S.J., Thornton, J.M., 1993. NACCESS, 2.1. 1. Dept of Biochemistry and Molecular Biology: University.
Johansson, F., Toh, H., 2010. A comparative study of conservation and variation scores. BMC Bioinformatics 11, 388.
Jones, D.T., Buchan, D.W.A., Cozzetto, D., Pontil, M., 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28, 184–190.
Kabsch, W., 1978. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A 34, 827–828.
Kabsch, W., 1976. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A 32, 922–923.
Karlin, S., Brocchieri, L., 1996. Evolutionary conservation of RecA genes in relation to protein structure and function. J. Bacteriol. 178, 1881–1894.
Kimura, M., 1983. The Neutral Theory of Molecular Evolution.
Leung, K., 2011. Presenting Post Hoc Hypotheses as A Priori: Ethical and Theoretical Issues. Management and Organization Review 7, 471–479.
Liang, S., Zhang, C., Liu, S., Zhou, Y., 2006. Protein binding site prediction using an empirical

scoring function. Nucleic Acids Res. 34, 3698–3707.
Lichtarge, O., Bourne, H.R., Cohen, F.E., 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257, 342–358.
Marino-Buslje, C., Monzon, A.M., Zea, D.J., Fornasari, M.S., Parisi, G., 2017. On the dynamical incompleteness of the Protein Data Bank. Brief. Bioinform. https://doi.org/10.1093/bib/bbx084
Marino Buslje, C., Teppa, E., Di Doménico, T., Delfino, J.M., 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.
Martin, L.C., Gloor, G.B., Dunn, S.D., Wahl, L.M., 2005. Using information theory to search for co-evolving residues in proteins. Bioinformatics 21, 4116–4124.
Monzon, A.M., Zea, D.J., Marino- Buslje, C., Parisi, G., 2017. Homology modeling in a dynamical world. Protein Sci. 26, 2195–2206.
Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., Weigt, M., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108, E1293–E1301.
Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., Lamberth, K., Buus, S., Brunak, S., Lund, O., 2004. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics 20, 1388–1397.
Olson, C.F., 1995. Parallel algorithms for hierarchical clustering. Parallel Comput. 21, 1313–1325.
Ovchinnikov, S., Kamisetty, H., Baker, D., 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife 3, e02030.
Panchenko, A.R., Kondrashov, F., Bryant, S., 2004. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci. 13, 884–892.
Peña, E.A., Slate, E.H., 2006. Global Validation of Linear Model Assumptions. J. Am. Stat. Assoc. 101, 341.
Rodriguez-Rivas, J., Marsili, S., Juan, D., Valencia, A., 2016. Conservation of coevolving protein interfaces bridges prokaryote-eukaryote homologies in the twilight zone. Proc. Natl. Acad. Sci. U. S. A. 113, 15018–15023.
Russell, R.B., Saqi, M.A., Sayle, R.A., Bates, P.A., Sternberg, M.J., 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J. Mol. Biol. 269, 423–439.
Shackelford, G., Karplus, K., 2007. Contact prediction using mutual information and neural nets. Proteins 69 Suppl 8, 159–164.
Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., Lehtinen, S., Studer, R.A., Thornton, J., Orengo, C.A., 2015. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43, D376–81.
Simonetti, F.L., Teppa, E., Chernomoretz, A., Nielsen, M., Marino Buslje, C., 2013. MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res. 41, W8–14.
Sutto, L., Marsili, S., Valencia, A., Gervasio, F.L., 2015. From residue coevolution to protein conformational ensembles and functional dynamics. Proc. Natl. Acad. Sci. U. S. A. 112, 13567–13572.
Tursynbay, Y., Zhang, J., Li, Z., Tokay, T., Zhumadilov, Z., Wu, D., Xie, Y., 2016. Pim-1 kinase as cancer drug target: An update (Review). Biomedical Reports 4, 140–146.
Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.-J., Kleywegt, G.J., 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 41, D483–9.

Williamson, J.M., Datta, S., Satten, G.A., 2003. Marginal analyses of clustered data when cluster size is informative. Biometrics 59, 36–42.
Worth, C.L., Gong, S., Blundell, T.L., 2009. Structural and functional constraints in the evolution of protein families. Nat. Rev. Mol. Cell Biol. https://doi.org/10.1038/nrm2762
Yeh, S.-W., Huang, T.-T., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., Echave, J., 2014. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. Biomed Res. Int. 2014, 572409.
Yeh, S.W., Liu, J.W., Yu, S.H., Shih, C.H., Hwang, J.K., Echave, J., 2014. Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. Mol. Biol. Evol. 31, 135–139.
Zea, D.J., Anfossi, D., Nielsen, M., Marino-Buslje, C., 2016. MIToS.jl: Mutual Information Tools for protein Sequence analysis in the Julia language. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw646
Zea, D.J., Miguel Monzon, A., Fornasari, M.S., Marino-Buslje, C., Parisi, G., 2013. Protein conformational diversity correlates with evolutionary rate. Mol. Biol. Evol. 30, 1500–1503.

# Highlights

- 51% of residues change their exposed/buried status between structures of the family.
- Inter residue contacts change commonly between structures of the family.
- Conservation is better explained by the solvent accessibility of multiple structures.
- Covariation methods predict evolutionary conserved inter residue contacts.