Structural comparison of protein-RNA homologous interfaces reveals widespread overall conservation contrasted with versatility in polar contacts

Ikram Mahmoudi[1], Chloé Quignot[1], Carla Martins[1,#], Jessica Andreani[1,*]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

[#] Current address: MI-GSO PCUBED, Bâtiment Turcat II, 7 rue Alain Fournier, ZAC Saint Martin du Touche, 31300 Toulouse

[*] Corresponding author

Abstract

Protein-RNA interactions play a critical role in many cellular processes and pathologies. However, experimental determination of protein-RNA structures is still challenging, therefore computational tools are needed for the prediction of protein-RNA interfaces. Although evolutionary pressures can be exploited for structural prediction of protein-protein interfaces, and recent deep learning methods using protein multiple sequence alignments have radically improved the performance of protein-protein interface structural prediction, protein-RNA structural prediction is lagging behind, due to the scarcity of structural data and the flexibility involved in these complexes. To study the evolution of protein-RNA interface structures, we first identified a large and diverse dataset of 2,022 pairs of structurally homologous interfaces (termed structural interologs). We leveraged this unique dataset to analyze the conservation of interface contacts among structural interologs based on the properties of involved amino acids and nucleotides. We uncovered that 73% of distance-based contacts and 68% of apolar contacts are conserved on average, and the strong conservation of these contacts occurs even in distant homologs with sequence identity below 20%. Distance-based contacts are also much more conserved compared to what we had found in a previous study of homologous protein-protein interfaces. In contrast, hydrogen bonds, salt bridges, and π-stacking interactions are very versatile in pairs of protein-RNA interologs, even for close homologs with high interface sequence identity. We found that almost half of the non-conserved distance-based contacts are due to a small proportion of interface residues that no longer belong to the interface in the interolog, a phenomenon we term “interface switching out”. We also examined possible recovery mechanisms for non-conserved hydrogen bonds and salt bridges, uncovering diverse scenarii of switching out, change in amino acid chemical nature, intermolecular and intramolecular compensations. Our findings provide insights for integrating evolutionary signals into predictive protein-RNA structural modeling methods.

Introduction

Protein-RNA interactions are crucial in many cellular processes, such as RNA metabolism, translation, DNA damage repair, and gene regulation [1, 2]. They have also been implicated in numerous pathologies, such as cancers and neurological disorders [3]. Several studies of protein-RNA structures gave insights into possible pathological molecular mechanisms [4, 5]. However, for many pathologies, the underlying mechanisms remain unresolved, leading to limitations in the proposed treatments [3, 6, 7]. Therefore, understanding protein-RNA interactions presents a major challenge in molecular biology. Detailed comprehension of those interactions is a crucial goal for medical and pharmaceutical purposes like drug design [6, 7], which requires knowledge of the corresponding 3D atomic structures [2, 8].

Even if the number of available experimental structures for protein-RNA complexes has greatly increased in the last decades, experimentally solving protein-RNA structures is still very challenging [2]. Only approximately 5,300 structures of protein-RNA complexes were available in the Protein Data Bank (PDB) in October 2023, compared to more than 200,000 entries overall, mostly proteins and homomeric protein complexes [9]. Therefore, computational tools for protein-RNA structural prediction and interface characterization have been the subject of dedicated research for several decades [10].

One major strategy for computational prediction of macromolecular interfaces is template-based prediction, which provides high-quality predictions for protein-protein complexes [11]. For protein-RNA interfaces, one pioneering study searched for structural interface similarity within a dataset of 439 non-redundant protein-RNA binary interfaces and identified a threshold of 25% for the minimum among the protein-protein and RNA-RNA sequence identity, to identify structurally similar pairs of binary interfaces [12]. This study showed that above a sequence identity threshold of 30-35%, structural binding modes are similar and that many structurally similar complexes display low

sequence identity. However, the study did not investigate further how this conservation is enabled at the atomic scale and rather focused on template-based interface modeling.

When templates are not available or not detectable, interface modeling needs to resort to template-free docking. Free docking approaches most often consist of a sampling step generating many possible interface conformations, followed by a scoring step where these conformations are ranked [10]. Some of the most common scoring approaches rely on statistical potentials, deriving pairwise residue-ribonucleotide propensities from known structures of protein-RNA interfaces [13-15]. Other strategies involve coarse-grained force fields [16] or scores optimized by machine learning [17]. Protein-RNA free docking is more challenging than protein-protein docking due to the conformational flexibility of both protein and RNA partners [18] and the scarcity of high-resolution protein-RNA structures on which scoring methods can be trained. A hybrid strategy using both template-based and free docking improved protein-RNA interface structural predictions, especially in cases with low-homology templates [19]. Alternative methods, such as in the HADDOCK web server [20], include experimental restraints to guide the docking.

Evolutionary pressures apply to protein-protein interfaces to maintain interactions between partners [21]. Interface conservation and co-evolution signals between interface positions can be exploited to improve the structural prediction of protein-protein interactions in traditional docking [22], but also in global statistical methods exploiting covariation in multiple sequence alignments (MSAs) to derive the most likely direct contacts [23]. Many predictors of RNA-binding protein residues rely on machine learning using evolutionary information from Position Specific Scoring Matrices (PSSM) or homology transfer from structural templates [24, 25]. Co-evolutionary analysis can also indicate conserved RNA structures [26] and protein-RNA interfaces, with the caveat that these methods require large coupled MSAs and, therefore, are only applicable to a few bacterial protein-RNA complex families [27].

The recent release of powerful methods using deep learning algorithms to leverage information from MSAs, such as AlphaFold [28], AlphaFold-Multimer [29] and RoseTTAFold [30], has been a revolution

for the structural prediction of proteins and protein-protein interactions. These methods have increased applicability compared to previous covariation-based models, thanks to the ability to exploit relatively small MSAs, and the success rates and precision of models have greatly increased compared to traditional docking methods. Recently, deep learning methods have been extended to predict protein-nucleic acid interface structures, notably with RoseTTAFold2NA [31] and, most lately, RoseTTAFold All-Atom [32] and AlphaFold3 [33]. These new methods encouragingly demonstrate the ability to learn joint parameters for diverse macromolecular interactions despite scarce protein-nucleic acid structural data and report high prediction performance for protein-protein complexes. However, the reported performance for protein-RNA structure prediction is much lower; for instance, the reported interface local distance difference test (iLDDT) for a small test set of 25 protein-RNA complexes with low homology to PDB structures is 19% for RoseTTAFold2NA and 38% for AlphaFold3 [33]. Additionally, the AlphaFold3 code and weights have not yet been released to the general public.

An earlier study of the evolution of protein-protein interface structures [34] enabled us to identify conserved determinants that were subsequently useful to develop dedicated scoring functions that improved our predictive capacity [35-37]. Even in the current context of deep learning developments, understanding and leveraging evolutionary information remains crucial, and better protein-RNA interface structure prediction and scoring methods are still needed. Therefore, the structural analysis of protein-RNA interface evolution deserves special attention.

In the present study, we focused on the analysis of protein-RNA interface evolution, aiming to unravel relevant keys for structural modeling and prediction of protein-RNA complexes and for interpretation of multiple sequence alignments. We identified 2,022 pairs of homologous protein-RNA interfaces with structurally similar experimental 3D structures and we used this unique dataset to perform a detailed analysis of how interface contacts are conserved between interologs. Abundant literature over more than twenty years [38-40] has defined important contact types for the energetics and specificity of protein-RNA interfaces: beyond atomic proximity, we also considered hydrophobic interactions, hydrogen bonds (H-bonds), and salt bridges, as well as π-stacking contacts. In this study, we highlighted

the diverse conservation of these different contact types and the role of sequence divergence as a major determinant of contact conservation. We also explored the role of structural properties such as secondary structure and solvent accessibility in contact conservation.

Results

Identification of structural interologs

We first built a dataset of representative, high-resolution protein-RNA interface structures. From all experimental structures in the PDB, we retrieved the subset of entries containing at least one protein-RNA contact, where we defined contacts as amino acid/nucleotide pairs with a minimum heavy-atom distance below 5Å. From these PDB entries, we extracted binary interfaces containing one protein chain in contact with either one RNA chain, or two base-paired RNA chains that we merged into one double-stranded chain. Because we aim for good-quality structures with well-defined atomic contacts, we applied resolution and size criteria and obtained 3383 interfaces. Finally, we applied clustering to remove strictly identical protein-RNA interfaces to avoid bias in our dataset. This pipeline (Fig 1) resulted in 977 representative protein-RNA binary interfaces.

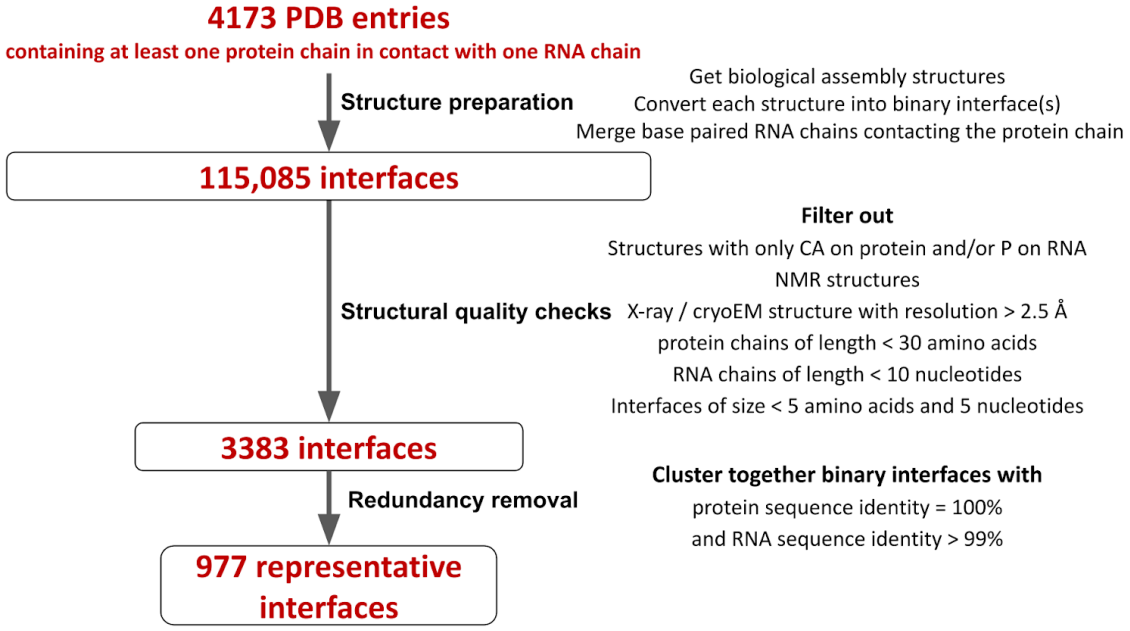

Fig 1: Pipeline for the construction of the interface dataset. The pipeline proceeds from all available protein-RNA complexes in the PDB (dated 21 February 2022) to the representative dataset of 977 interfaces used in this study.

Among these 977 representative interfaces, our goal was then to identify subsets of interface structural homologs (called structural interologs or interologs for short). We performed all-vs-all structural alignment of these 977 protein-RNA interfaces using TM-align, RNA-align, and MM-align respectively for protein alignment, RNA alignment, and protein-RNA interface alignment (see Methods and Fig 2A). Of note, this structural alignment step is computationally costly, further arguing in favor of the clustering performed to obtain the 977 representative interfaces, as we avoided unnecessary interface comparisons. Out of over 444,000 possible comparisons, MM-align succeeds in aligning simultaneously the protein and RNA molecules for around 207,000 pairs of protein-RNA interfaces. For those aligned pairs of interfaces, Fig 2B shows the relationship between the interface TM-score (measuring the interface structural similarity) and the minimum sequence identity based on a local alignment of the protein and RNA sequences (weighted by alignment coverage, see Methods). This graph shows that above a sequence identity threshold of 25%, all pairs of interfaces display interface TM-scores above 0.5 (the standard threshold above which TM-align results are indicative of

structurally similar folds), and a large fraction of those have interface TM-scores above 0.8. Conversely, Fig 2B also displays a densely populated region with interface TM-score above 0.5 but sequence identity below 25%. The red dots in Fig 2B represent 6,479 pairs of possible interologs.

With structural visualization, we observed that interfaces may not be similar even when the interface TM-score is high (Fig 2C). On the other hand, when the RNA TM-score is low, the interfaces may be structurally similar, but the flexibility of RNA molecules might result in structures that are not overall superimposable (S1A Fig). Taking into account these observations, we did not use the RNA TM-score to define interologs; we chose cutoff values of 0.5 and 0.6 for interface and protein TM-scores, respectively, and we complemented the TM-score criteria with interface coverage criteria, whereby the interface overlap had to exceed 40% in both protein and RNA chains within the pair of interologs (Fig 2A). This resulted in a final set of 2,022 pairs of confidently assigned structural interologs with interface TM-score above 0.5, spanning a wide range of possible interface sequence identities (S1B Fig). Among these 2,022 pairs, 515 pairs have an RNA TM-score lower than 0.5 (S1C Fig). We computed interface root mean square deviation (RMSD) between interologs and verified that the vast majority of interologs have low RMSD values (99.9% below 6Å and 95.5% below 4Å, S1D Fig). Notably, 765 out of the 977 initial representative interfaces (78%) have at least one identified structural interolog in this process.

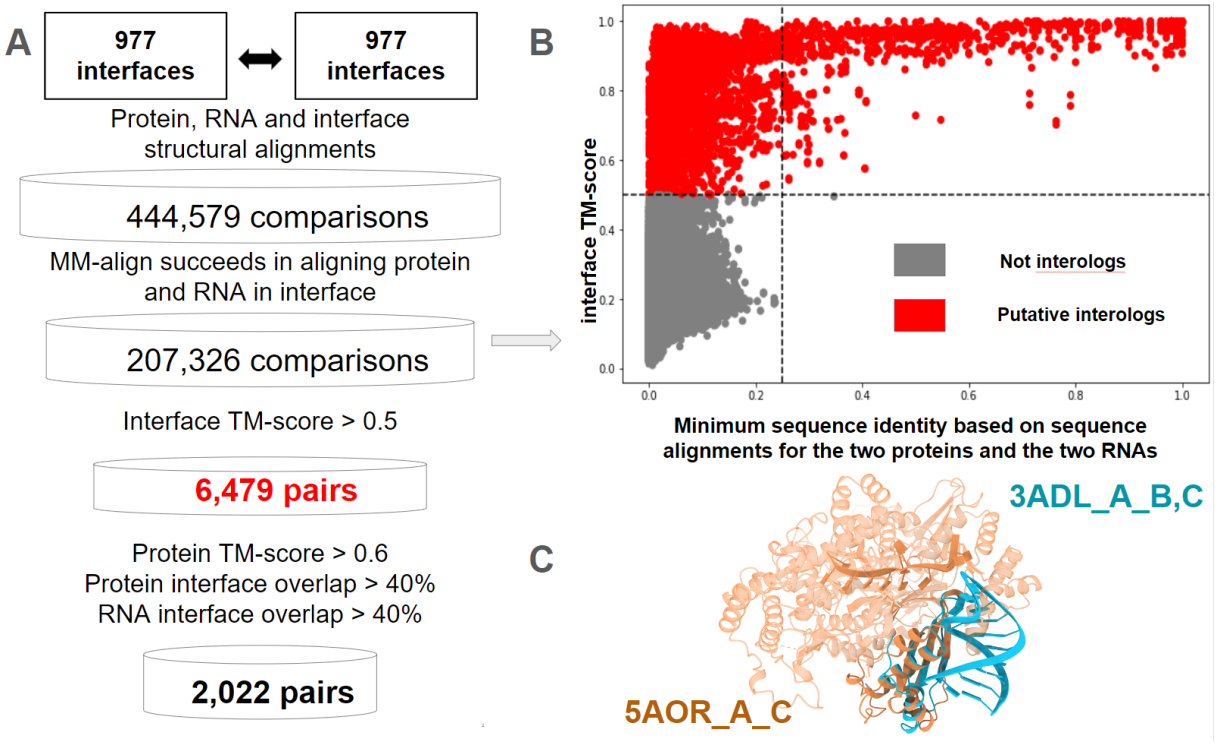

Fig 2: Identification of pairs of structural interologs among the 977 interfaces in our dataset. A: Pipeline for the identification of interologs from all-against-all structural alignment. B: Scatter plot of all-against-all interface TM-scores (y-axis) according to (x-axis) the minimum sequence identity within each pair of binary protein-RNA interfaces, weighted by alignment coverage. In the scatter plot, we excluded pairs of interfaces that MM-align failed to align. Gray dots have an interface TM-score lower than 0.5 (non-interologs). Red dots have an interface TM-score greater than 0.5 (putative interologs). C: An example pair of interfaces aligned with MM-align 5AOR_A_C; 3ADL_A_B, C associated protein TM-score 0.74, RNA TM-score 0.12, and interface TM-score 0.61; the two RNA chains do not bind in the same protein region so that the RNA interface coverage is 0 and this pair of interfaces is not retained as interologs.

Our definition of structural interologs was so far purely based on structural similarity criteria. To assess whether the 2,022 pairs of structural interologs actually correspond to evolutionary homologs, we used evolutionary classifications to annotate our interfaces, namely ECOD [41] for protein chains and Rfam [42] for RNA chains (see Methods). We then assessed whether for each pair of interologs, the ECOD and Rfam classifications were consistent for the two structurally similar interfaces. For protein

chains, 1,994 out of 2,022 pairs of structural interologs have ECOD annotations for both interfaces; out of these, 1,991 interolog pairs (99,8%) have at least one ECOD annotation in common at the T-group level (demonstrating homology and similar topological connections), and 1,803 interolog pairs (90%) have exactly the same ECOD T-group annotations. For RNA chains, 1,230 out of 2,022 pairs of interologs have Rfam clan annotation for both aligned interfaces and 100% of these have at least one clan in common (of which 99% have exactly the same clan annotation). 1,244 out of 2,022 pairs of interologs have Rfam family annotation for both aligned interfaces, 57% of which (703 pairs) have at least one common Rfam family. The set of interologs with a common clan but no common family corresponds to more remote interfaces, e.g. archaeal vs. eukaryotic ribosomal RNA. This analysis shows that beyond structural similarity, our dataset of interologs also contains a vast majority of evolutionarily related interfaces, despite the homology relationship being very distant in a large fraction of the interologs, as evidenced by low sequence identities.

Contact conservation analysis

For the 2,022 pairs of structural interologs, we compared the specific positions involved in each interface contact. Our strict definition of interface contacts, based on a minimum heavy-atom distance of 5 Å, is adapted to our need to precisely assess whether the close neighbors of an interface position are conserved in the interolog interface. Corresponding residues between interologs were defined from the interface structural alignment, as illustrated in Fig 3A. We assessed whether each interface contact occurring in our dataset was conserved in the structural interolog, irrespective of whether the nature of the amino acid/nucleotide varied (Fig 3B). To better represent the conservation of atomic contacts in a situation where the residue nature can vary, we weighed the conservation of each amino acid/nucleotide contact by the number of atomic contacts it contains (see Methods). On average, 73% of these distance-based, atomic-weighted contacts were conserved among interologs.

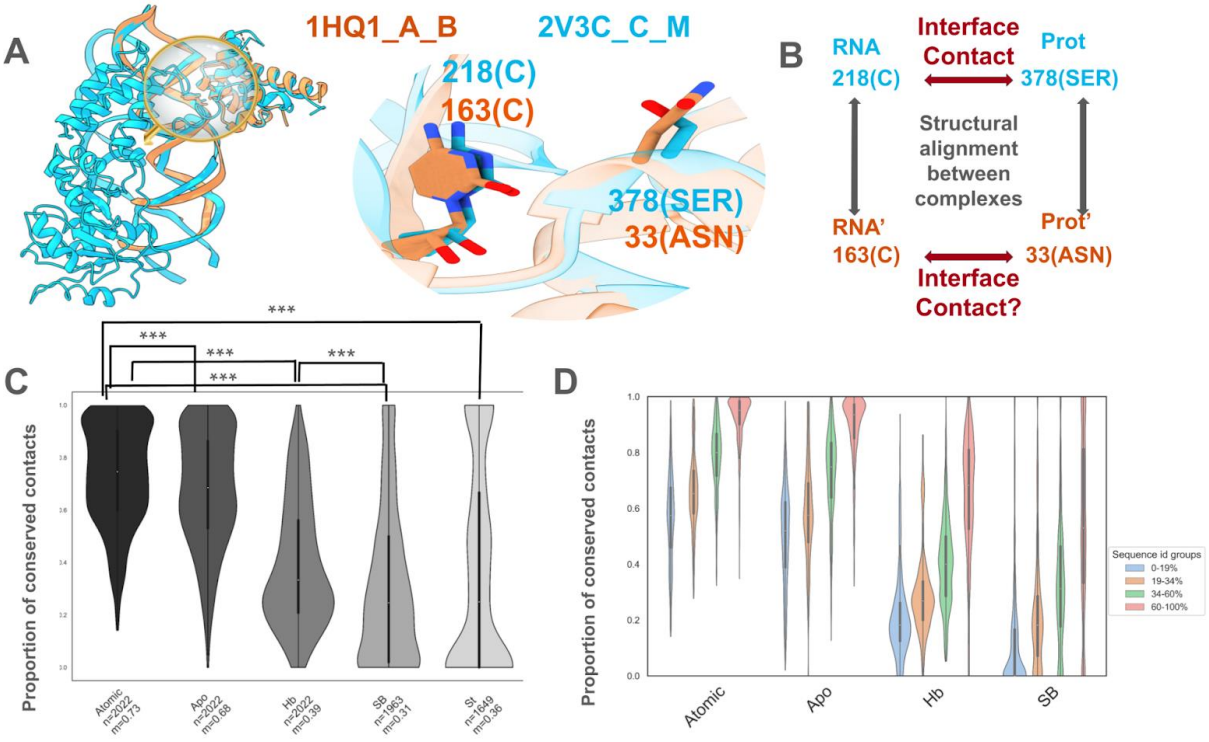

Fig 3: Interface contact conservation among interologs. A: Structural alignment illustration for two structural interologs (1HQ1_A_B *Escherichia coli* interface in cyan, aligned with 2V3C_C_M *Methanocaldococcus jannaschii* interface in orange) sharing 62% minimum interface sequence identity. The complexes were aligned with MM-align and the zoomed region shows alignment details for two structurally aligned amino acid/nucleotide pairs. B: Illustration of the way contact conservation between interologs was assessed, based on structurally aligned positions. C: Violin plot distribution of contact conservation for atomic contacts (Atomic), apolar contacts (Apo), H-bonds (Hb), salt bridges (SB), π-stacking contacts (St). n is the number of interolog pairs used in each violin plot (out of 2,022 pairs) and m is the mean conservation ratio. P-values between distributions denoted by *** in this panel are <2.2e-09 in a Wilcoxon rank sum test. D: Violin plot distribution of contact conservation for atomic contacts (Atomic), apolar contacts (Apo), H-bonds (Hb), and salt bridges (SB) for pairs of interologs separated into four groups of interface sequence identity (blue: 0-19%, brown: 19-34%, green: 34-60%, red: 60-100%). For any given type of contact, the differences between any two distributions of conservation among the four groups of sequence identities are statistically significant (p-value<8e-12 in Wilcoxon rank sum tests). For any given group of sequence identity, the differences between distributions of atomic and apolar contact conservation are also statistically significant (p-value<5e-6 in Wilcoxon rank sum tests).

We also assessed weighted conservation for atomic contacts restricted to pairs of C atoms (a subtype we called apolar contacts). Contact conservation was 68% on average for these apolar contacts, overall quite high, albeit significantly lower than for all atomic contacts (Fig 3C). However, only 39% of the H-bonds, 31% of the salt bridges, and 36% of the π-stacking contacts were conserved on average, with all distributions significantly lower than for atomic contacts (Fig 3C). Of note, H-bonds, salt bridges, and π-stacking interactions are also much less abundant compared to the atomic contacts (see supplementary results in S1 File), which might reflect that the interfaces are maintained in evolution through atomic and apolar contacts, while the specificity-driving H-bonds, salt bridges, and π-stacking interactions are more versatile.

The 2,022 pairs of interologs display a wide range of sequence identities. We created four equally populated groups of interologs according to the minimum interface sequence identity based on the structural alignment between the protein-RNA interfaces. Given that π-stacking is the least abundant type of contact in our protein interaction dataset (see supplementary results in S1 File), we decided to not further divide the π-stacking contact data into sequence identity groups, as these groups would contain very few data points. Henceforth, we will focus on atomic contacts, apolar contacts, H-bonds, and salt bridges. Fig 3D shows that the conservation of H-bonds and salt bridges is especially low for interologs with minimum interface sequence identity below 60%; even in the group of closest interologs (minimum interface sequence identity 60% to 100%), the average contact conservation is only 65% for H-bonds and 55% for salt bridges. On the contrary, atomic and apolar contacts are well conserved even in pairs of remote interologs; their respective average conservations are 56% and 50% in the group of interologs with minimum interface sequence identity below 19%.

We note that the more distant the interologs, the lower the interface overlap between them (S1E Fig). The 0-19% sequence identity group in particular has roughly one third fewer structurally aligned amino acid-nucleotide pairs despite similar interface sizes (see supplementary results in S1 File and S1F Fig),

hence contact conservation in this group might be moderately over-estimated compared to other groups because many amino acid-nucleotide pairs cannot be structurally aligned due to strong divergence. The challenges in structural alignment of such divergent interfaces will also translate into challenges for template-based modeling based on such remote templates. On the contrary, templates at 20% to 30% sequence identity do not seem to behave very differently from closer templates.

To verify that the observed trends were not due to our way of weighing the conservation metrics, we also analyzed distributions of contact conservation computed with two alternative metrics, either unweighted (i.e. where each amino acid-nucleotide pair counts for 1) or following an alternative definition of generalized Jaccard index (using the ratio of smaller to larger number of atomic contacts in each pair of aligned amino acid-nucleotide contacts, see supplementary methods in S1 File). Despite the values of conservation being overall lower with these alternative metrics, we observed similar trends of increasing contact conservation with increasing interface sequence identity (see S2 Fig and supplementary results in S1 File). In the rest of this manuscript, we performed all analyses using the Jaccard index weighted by the number of atomic contact pairs.

As our dataset of interologs contains a majority of interfaces involving a ribosomal protein (see supplementary methods in S1 File), we also verified the contact conservation behavior is mostly similar for ribosomal compared to non-ribosomal interologs, provided that we account for the different interface sequence identity composition of the two subsets (S3 Fig).

Interface determinants of atomic and apolar contact conservation

We then explored how different interface properties might influence the conservation of atomic and apolar contacts: secondary structure of interface amino acids and base pairing of interface nucleotides, interface regions assigned depending on solvent accessibility, and evolutionary conservation derived from a protein multiple sequence alignment.

We categorized contacts for each pair of interologs based on the type of secondary structure (helix, strand, or coil) in which amino acids and their structural equivalents are involved (see Methods). When amino acids have the same secondary structure type in both interologs, whether this secondary structure type is helix, strand or coil, the average proportion of conserved atomic contacts is 73%; in contrast, when amino acids change secondary structure, the average conservation drops to 55% (S4A Fig).

A similar but even stronger trend is observed in the analysis based on RNA secondary structure, i.e. whether nucleotides are base-paired or not in one or both interologs. The average contact conservation ratio drops from 90% (respectively 68%) when nucleotides remain base-paired (respectively, unpaired) in both interologs to 23% when they change base-pairing status (S4B Fig). Strikingly, the tendency is the same within pairs of remote interologs with low interface sequence identity (S4C Fig) and for apolar contacts (S4D Fig).

When nucleotides change base-pairing status between interologs, non-conserved contacts correspond to a vast majority of contacts made by a base-paired nucleotide, and lost in the interolog where the nucleotide is unpaired, rather than the reverse (see supplementary results in S1 File). A possible explanation for these non-conserved contacts might thus be that the loss of nucleotide base-pairing leads to increased flexibility, compared to base-paired nucleotides that have more limited possible contacts with the protein.

To further investigate determinants of atomic contact conservation, we classified protein interface amino acids into core and rim regions (see Methods). Numerous studies have previously emphasized the distinctive characteristics of these subregions concerning composition and evolutionary properties in the context of protein-protein interactions [34, 43, 44]. We found that for protein-RNA interologs, atomic contacts within the core exhibit significantly higher conservation compared to contacts in the rim region (S5A Fig). Specifically, the average conservation of atomic contacts involving amino acids

from the core regions in both interologs is 86%. In contrast, the average conservation of atomic contacts involving at least one residue from the rim region in any of the two interologs is 73 to 75% (p-value<1e-46 based on Wilcoxon rank sum tests). Note that this analysis is limited to contacts involving residues that remain at the interface in both interologs, therefore removing a major cause for non-conservation (see next section), which leads to rather high average conservation percentages.

Finally, we analyzed atomic contact conservation according to position-specific amino acid evolutionary conservation derived from a protein multiple sequence alignment (see Methods). This property is especially interesting in the perspective of predictive developments, since it can be derived without any knowledge of the protein-RNA interface structure. We defined four groups of evolutionary conservation according to the minimum amino acid conservation within a pair of structurally aligned contacts: 0-30%, 30-50%, 50-70%, and 70-100%, for which the average atomic contact conservation is 47%, 53%, 58%, and 68%, respectively (S5B Fig). This result confirms that the most evolutionarily conserved amino acids within a protein family are also the positions that conserve most often their protein-RNA structural contacts.

Analysis of non-conserved atomic contacts

Given the limitation to interface modeling that non-conserved contacts represent, we analyzed them in greater detail. We first examined the cases where either the amino acid or the nucleotide forming the non-conserved atomic contact no longer belongs to the interface. We previously identified this phenomenon, which we called “switching out of the interface”, as a major cause of non-conservation observed in protein-protein interface evolution [34]. In our dataset of 2,022 protein-RNA interolog pairs, on average, 11% of interface amino acids switch out of the interface in the interolog (Fig 5A) while the fraction is only 5% for interface nucleotides (S6A Fig). These fractions rise to 21% and 17%,

respectively, when switching out positions are not weighted by the number of atomic contacts they are involved in.

Despite the low proportions of switching out residues, we found that switching out was a major driver of contact non-conservation, with 47% of non-conserved atomic contacts linked to switching out on average. Strikingly, this proportion does not vary very strongly with sequence identity (S6B Fig). This means that in remote interologs, the fraction of non-conserved contacts is higher compared to close interologs, but so is the fraction of switching out residues, and the proportion of non-conserved contacts linked to switching out does not vary much. Among non-conserved contacts linked to switching out, 61% are cases where the amino acid switches out of the interface, 29% where the nucleotide switches out, and 10% where both switch out. These proportions are fully consistent with the proportions of interface amino acids and nucleotides observed to switch out of the interface in the interolog.

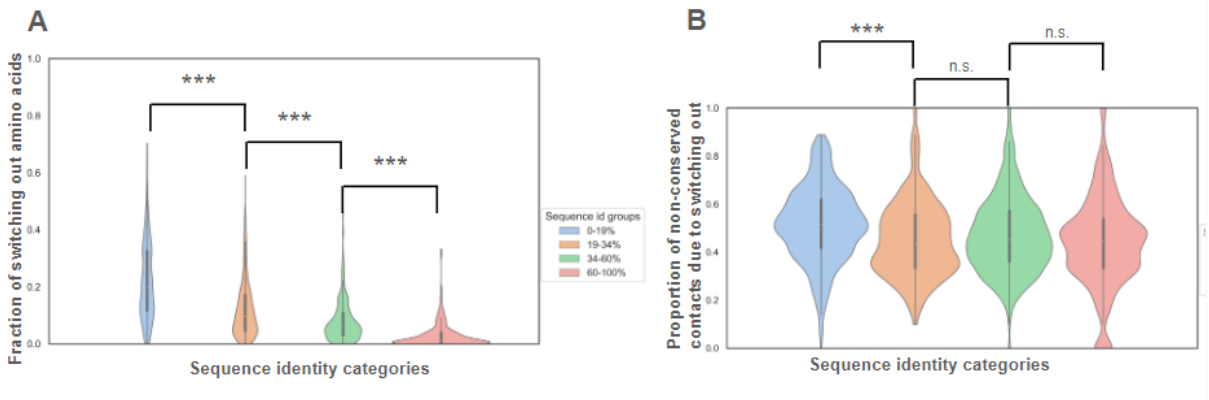

Fig 4: Analysis of switching out in contact non-conservation. A: Violin plot distribution of the percentage of switching out amino acids (weighted by the number of atomic contacts in which each amino acid is involved) across the four ranges of interface sequence identity. The differences between any two distributions of switching out among the four groups of sequence identities are statistically significant (p-value < 1.6e-13 in Wilcoxon rank sum tests). B: Violin plot distribution of the unweighted percentage of non-conserved contacts due to switching out across the four ranges of interface sequence identity. The difference between 0-19% and 19-34% distributions is significant (p-value = 8.4e-13 in Wilcoxon rank sum test), while the differences between 19-34% and 34-60%, and 34-60% and 60-100% are not (p-values = 0.035 and 0.015 in Wilcoxon rank sum tests).

Compensation mechanisms for non-conserved H-bonds and salt bridges

Numerous studies show that hydrogen bonds play a crucial role in protein-RNA interactions by contributing to the specificity and stability of the complexes and the function of RNA-protein complexes [45, 46]. As we observed a marked versatility of hydrogen bonds, even between interologs with high interface sequence identity (Fig 3D), we conducted an in-depth analysis of non-conserved H-bonds, aiming to investigate potential recovery mechanisms. In this analysis, we focused only on H-bonds involving the side chains of amino acids, as the ability to form H-bonds through the amino acid backbone does not vary with evolutionary mutations. This explains why the fraction of conserved H-bonds is different (light green slices in Fig 5A) compared to Fig 3.

Across the four groups of interface sequence identity, we accounted for a large part of the non-conserved H-bonds due to switching out or lost polarity of the sidechain (light blue and dark green slices in Fig 5A, illustrated by light blue and dark green boxes in Fig 5B). A large portion of the remaining non-conserved H-bonds (where the residues remain at the interface and the amino acid retains a polar sidechain) can be explained by recovery mechanisms involving other intermolecular H-bonds with a non-structurally aligned position in the interolog (dark blue slices in Fig 5A, illustrated by dark blue box in Fig 5B) or intramolecular H-bonds with a neighboring amino acid (dark purple slices in Fig 5A, illustrated by dark purple box in Fig 5B). A small fraction of the non-conserved H-bonds remains unexplained (purple slices in Fig 5A) and may correspond, for instance, to larger distance contacts mediated by water molecules. These findings provide guidelines for both template-based modeling and assessment of interfaces in protein-RNA docking poses.

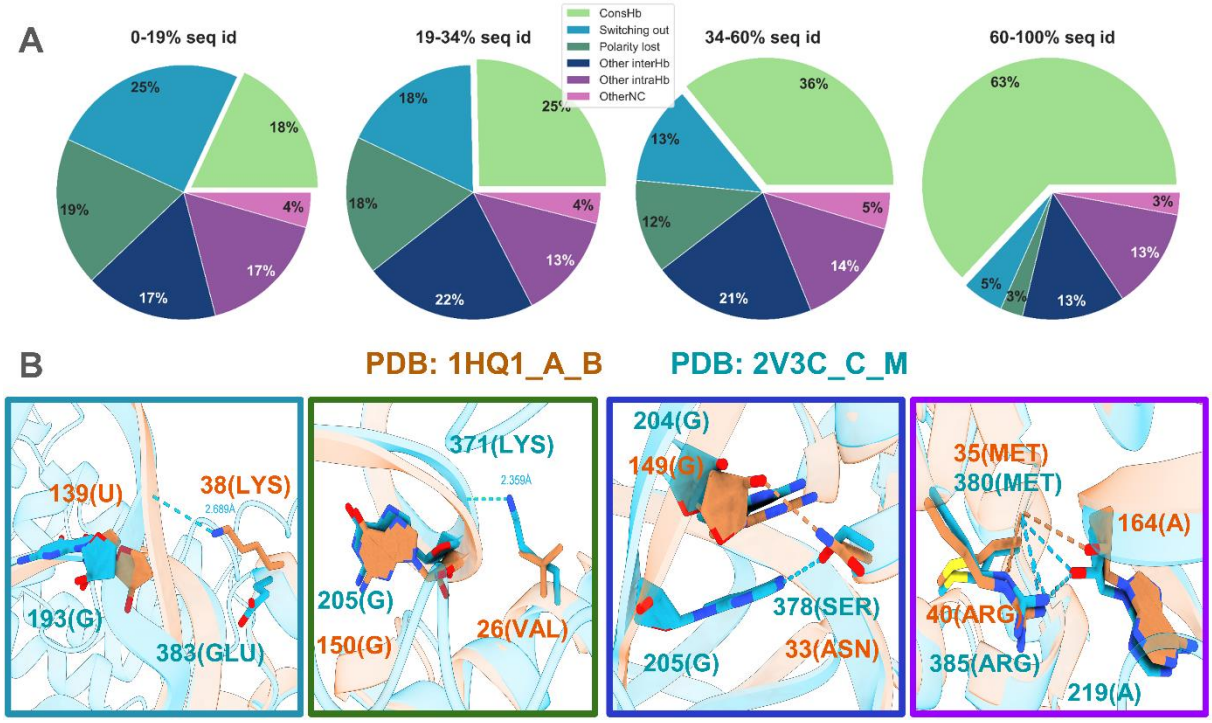

Fig 5: H-bond conservation and recovery mechanisms of non-conserved hydrogen bonds. (A) Pie plots representing the scenarios of H-bond conservation across the four groups of interface sequence identity. The six categories represent conserved hydrogen bonds (ConsHb), non-conserved H-bonds due to switching out (Switching out), non-conserved H-bonds due to non-polar character of the interface amino acid, leading to a loss of their ability to form H-bonds through their sidechain (Polarity lost), non-conserved H-bonds involving amino acids that retained the ability to form H-bonds with neighboring nucleotides (Other interHb), non-conserved H-bonds involving amino acids that retained the ability to form H-bonds but did not form them with nucleotides, rather with adjacent amino acids (Other intraHb) and other non-conserved H-bonds where the amino acid retained the ability to form H-bonds but did not form them either inter- or intra-molecularly (OtherNC). (B) Structural illustrations of four categories from the pie plot, from left to right: Switch (switching out of nucleotide 193(G)), Polarity Lost (26(VAL) lost the ability to form H-bonds through its sidechain), Inter (378(SER) forms a hydrogen bond with nucleotide 205(G), not structurally aligned with 149(G)) and Intra (40(ARG) no longer forms a H-bond with a nucleotide but forms an intramolecular H-bond with amino acid 35(MET)).

Salt bridges form a subset of hydrogen bonds involving short-distance ionic interactions. The conservation of salt bridges among protein-RNA interologs is even lower than for H-bonds (Fig 3D). We investigated recovery scenarios in a manner very similar to H-bonds. Similarly to H-bonds, a large part

of the non-conserved salt bridges can be explained by switching out or loss of the basic chemical property of the amino acid (light blue and dark green slices in Fig 6A, illustrated by light blue and dark green circles in Fig 6B). However, among the remaining non-conserved salt bridges, only a fraction can be explained by recovery mechanisms involving other intermolecular salt bridges with a non-structurally aligned position in the interolog (dark blue slices in Fig 6A) or intramolecular salt bridges (dark purple slices in Fig 6A). This leaves a fraction of unexplained non-conserved salt bridges (light purple slices in Fig 6A) that is larger compared to H-bonds. Amino acids which remain basic but no longer form salt bridges can instead form intermolecular H-bonds with neighboring nucleotides, but this scenario accounts for a very small fraction (1-2%) of the total (hence, not singled out in Fig 6 and included in the OtherNC category). These amino acids could also form longer distance electrostatic interactions or interactions mediated by solvent molecules.

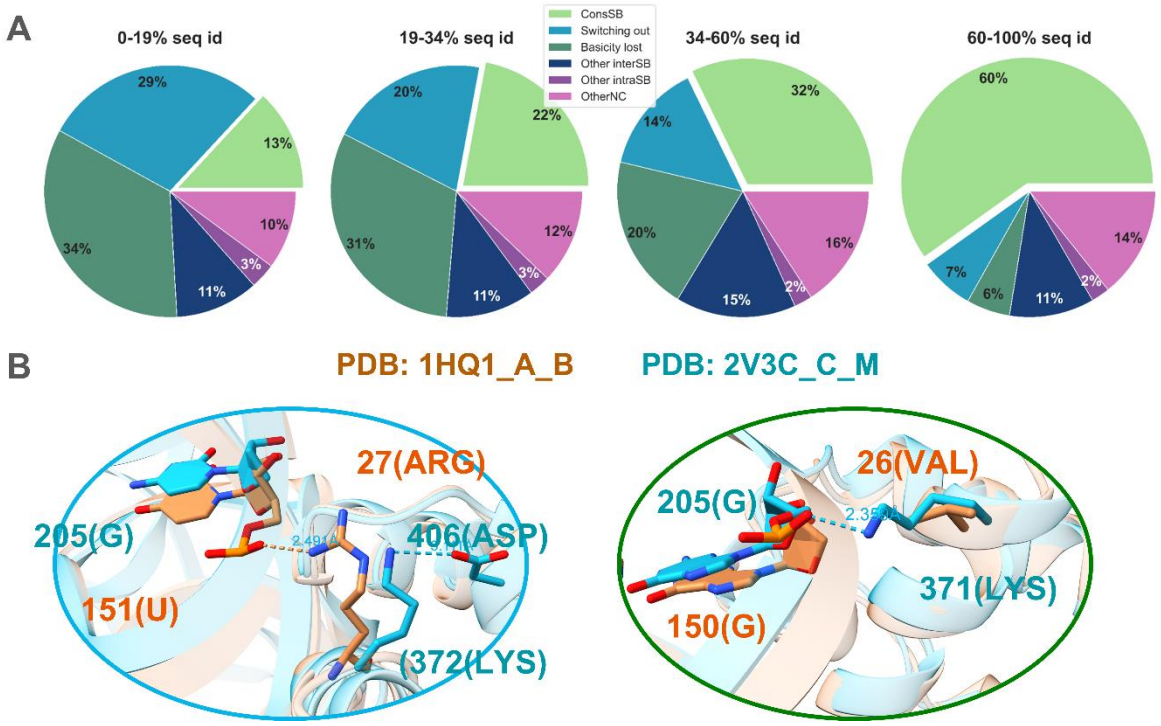

Fig 6: Salt bridge conservation and recovery mechanisms of non-conserved salt bridges. (A) Pie plots representing the scenarios of salt bridge conservation across the four groups of interface sequence identity. The six categories represent conserved salt bridges (ConsSB), non-conserved salt bridges due to switching out (Switching out), non-conserved salt bridges because the interface amino acid is not basic in the interolog (Basicity lost), non-conserved salt bridges involving basic amino acids that form

salt bridges with neighboring nucleotides (Other interSB), non-conserved salt bridges involving basic amino acids that form salt bridges with neighboring acidic amino acids (Other intraSB), and other non-conserved salt bridges (OtherNC). (B) Illustrations of non-conserved salt bridges due to switching out (left, light blue circle) and loss of basic character (right, dark green circle). In the left panel, the salt bridge between 27(ARG) and 151(U) in interface 1HQ1_A_B is not conserved due to switching out of amino acid 372(LYS) from interface 2V3C_C_M; this amino acid still forms an intramolecular salt bridge with neighboring amino-acid 406(ASP). In the right panel, salt bridge between 371(LYS) and 205(G) in interface 2V3C_C_M is not conserved in 1HQ1_A_B due to hydrophobicity of the 371(LYS) structural equivalent, 26(VAL).

Web interface for the exploration of interolog groups and contact conservation

To complement our study, we have developed a user-friendly web interface (see Methods) allowing for the dynamic exploration of groups and pairs of interologs and the interactive structural visualization of the 765 protein-RNA interfaces and 2,022 pairs of homologous interfaces in our dataset. Starting from a global view of interolog groups (Fig 7A), users can click on a group to display its network representation (Fig 7B) and details about the interolog pairs in this group. Users can then proceed to explore either a single interface (node of the network) or a pair of interologs (edge of the network). In the case of a single interface, contact details are displayed and the structure of the interface can be explored interactively (Fig 7C), including contacts involving any given amino acid or nucleotide. In the case of a pair of interologs, we display precomputed information about the conserved and non-conserved contacts and the aligned interface structures can also be explored interactively (Fig 7D). On both the interolog group page and the interface list page, users can search for specific keywords within PDB, ECOD and Rfam descriptions of the macromolecular components, to allow for the extraction of biologically relevant information from our data. Users can also download relevant information as tables for either the full datasets or a given interface or pair of interologs.

This web interface is freely accessible at https://bioi2.i2bc.paris-saclay.fr/django/rnaprotdb/.

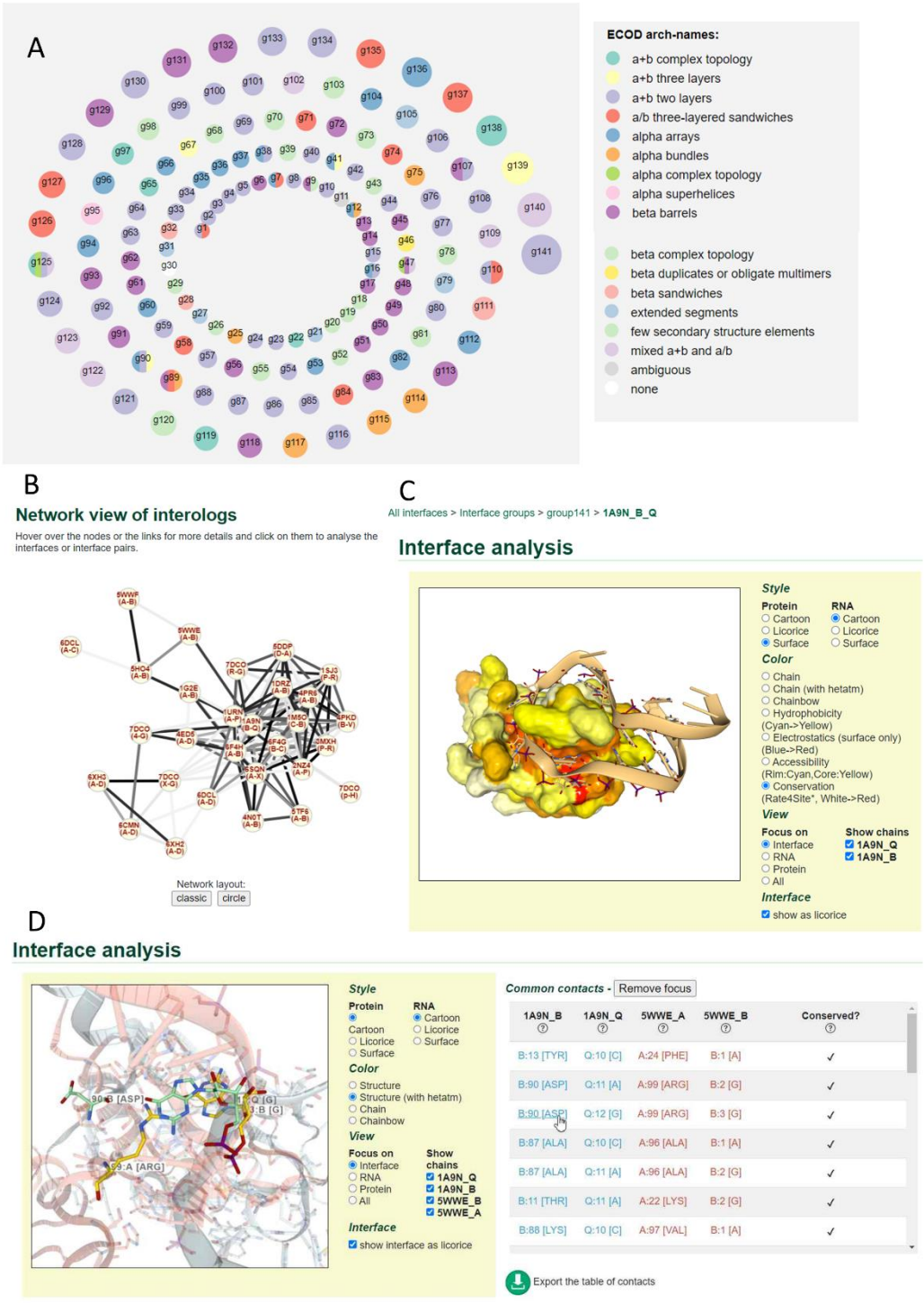

Fig 7: Web interface for the exploration of protein-RNA interfaces and interologs in our datasets (https://bioi2.i2bc.paris-saclay.fr/django/rnaprotdb/) (A) Overall view of the 141 interolog groups. Nodes are colored by ECOD architecture level information from the common ECOD domains in each group. Nodes can be clicked to explore each group. (B) Network view of the largest group (g141). Nodes

are interfaces and edges are homology relationships between interfaces. Edge colors reflect interface sequence identity (gray color scale - the darker, the higher the identity). Nodes can be reorganized according to predefined layouts or dragged manually. Nodes can be clicked to go to a single interface 3D structural view (panel C). Edges can be clicked to go to interface comparison 3D structural view (panel D). (C) 3D structural visualization of a single interface. Various modes of representation can be chosen. The illustration shows a surface view for the protein colored by evolutionary conservation (from white to red, most variable to most conserved) and a cartoon view for the RNA. Interface residues are highlighted as licorice. (D) 3D structural comparison of two homologous interfaces. The table of contacts shown to the right can be clicked to focus the visualization on any given position (here, ASP 90 of 1A9N_B). The structurally aligned position is shown (here, ARG 99 in 5WWE_A) as well as the corresponding contacting nucleotides (G 12 in 1A9N_Q and G 3 in 5WWE_B).

Discussion

Protein-RNA interactions are critical for many biological processes, including gene expression, RNA processing, and translation. Understanding the mechanisms underlying these interactions is pivotal for improving structural prediction methods and for further drug design targeting these interactions. Our in-depth analysis of structural interologs provides crucial insights into the evolution of protein-RNA interactions. We first highlighted that when protein-RNA interfaces share a minimum of 25% sequence identity, they share similar overall interface structures provided they display a similar protein fold. This is especially important for structural modeling of protein-RNA complexes, when aiming to identify templates or to retrieve relevant homologous sequence data. We then built a large dataset of over 2,000 pairs of well-resolved protein-RNA structural interologs, which enabled us to study the conservation of protein-RNA interface structures in great atomic detail. The high conservation of distance-based atomic contacts, and especially of apolar contacts involving base-paired nucleic acids, offers a reliable foundation for transferring these structural features of protein-RNA complexes in

template-based modeling. In parallel, the conservation of atomic and apolar contacts suggest that they are key for interface stability and underscores their potential as robust features for the development of machine learning predictors and for better identification of correct poses in template-free docking (e.g. through propensity-based scoring functions taking evolutionary information into account, a strategy we previously found successful for protein-protein interactions [35]). We further investigated the importance of interface sub-region and secondary structure conservation, and we evidenced increased contact conservation for evolutionarily more conserved interface regions. This latter feature is especially interesting since it does not require prior experimental knowledge of a protein-RNA interface structure. Conversely, the versatility of H-bonds and salt bridges, even between interfaces with high sequence identity, prompts cautious consideration in template-based modeling, and may offer insights into modulating the specificity of protein-RNA complexes, especially by considering the possible recovery mechanisms involving neighboring nucleotides or amino acids. Additionally, the insights we offer on protein-RNA interface evolution for a diverse range of protein and RNA families open new possibilities for tailoring scoring metrics to further improve protein-RNA specificity predictions; indeed, a recent study used carefully curated sequence and structure data to gain insights into the recognition code for RRM-RNA interactions and derive a specificity prediction score [47].

In a previous study, we analyzed a dataset of around 1,000 pairs of protein-protein structural interologs [34], which provides us with a basis for comparing protein-protein and protein-RNA interface evolution. Protein-RNA interface evolution shows distinctive features, with protein-RNA interfaces displaying a lower proportion of non-conserved contacts for all contact types, even though our protein-RNA interface dataset contains many remote interologs. The proportion of non-conserved contacts explained by switching out is slightly higher in protein-RNA interologs compared to protein-protein interologs, despite the proportion of switching out residues being much lower. Altogether, this would suggest that protein-RNA interfaces are less versatile than protein-protein interactions, even if previous studies show greater flexibility for both protein and RNA in protein-RNA interfaces compared

to protein-protein interfaces [18]. These observations help us to identify the most relevant keys to guide the modeling of protein-RNA interactions.

Other possible implications of protein-RNA interface evolution include insights into prebiotic world interactions. Those were previously investigated through the analysis of small datasets of protein-RNA complexes resulting from in vitro evolution [48, 49]. However, in the present study we rather perform a statistically robust analysis of a large number of homologous interfaces to derive general principles of protein-RNA interface evolution.

In conclusion, our study offers insights to advance understanding of protein-RNA interactions and their evolution. It lays the groundwork for developing propensity-based scoring functions and refining structural prediction methods. Finally, it opens perspectives into protein-RNA interface modulation and possible design of drugs targeting protein-RNA interfaces. Future research should leverage these evolutionary insights in advanced protein-RNA modeling approaches building upon the latest deep-learning advances, such as AlphaFold3 [33] and RosettaFoldNA [31]. Our insights into conservation of protein-RNA interfaces suggest that these methods may be leveraging structural information from remote interologs in the training dataset, raising the as yet unanswered question of how to fairly assess their generalizability.

Methods

Interface database construction

Fig 1 depicts the pipeline used to obtain the interface dataset. We collected 3D atomic structures of protein-RNA complexes from the Protein Data Bank [9] that contained at least one protein polymer entity and one RNA polymer entity, resulting in a total of 4,173 structures as of 21 February 2022. For each PDB entry, we used Gemmi [50] to generate coordinates of the biological assembly, which better

represents the complexes in their functional form compared to the asymmetric unit, and to assess pairwise heavy-atom distances between each amino acid and each nucleotide. We defined a contact between an amino acid and a nucleotide if the minimum heavy-atom distance between them was less than 5Å. The interface is defined as a set of all amino acids and nucleotides involved in such inter-chain contacts. We also processed all generated biological assembly CIF files using the software tool x3DNA v2.4 to assign base pairs in RNA structures [51, 52]. The complexes were divided into binary interfaces consisting of one protein chain contacting either one RNA chain or two base-paired RNA chains if at least one amino acid from the protein chain is in contact with two base-paired nucleotides, one from each RNA chain. A total of 115,085 binary interfaces were identified in the complete interface dataset.

As we aimed at a high-resolution study of interface contact conservation, we filtered out NMR structures and structures with a resolution worse than 2.5 Å, as well as protein chains shorter than 30 amino acids or containing only CA atoms, and RNA chains shorter than 10 nucleotides or containing only P atoms. We also excluded interfaces with fewer than 5 protein or 5 RNA interface residues. Clustering was performed on the resulting 3,383 protein-RNA (binary or ternary) interfaces, to group strictly redundant interfaces containing protein chains with 100% sequence identity and RNA chains with 99% sequence identity or more. We used MMseqs2 [53] and CD-HIT [54] for clustering protein and RNA sequences, respectively. The resulting interface dataset contains a representative (chosen as the interface with the best resolution) for each of the 977 interface clusters.

Interface analysis

We then generated detailed structural information for the dataset of 977 representative protein-RNA interfaces. Using Gemmi [50], we recorded the number of interacting atomic pairs (at a minimum distance of 5Å) involved in each amino acid-nucleotide contact and the number of apolar atomic pairs (focusing on carbon atoms). To characterize contacts of different natures [55, 56], we used x3DNA v2.4 [51, 52] to identify π-stacking interactions, salt bridges and hydrogen bonds (H-bonds), including

details about whether each H-bond involves sidechain/backbone for the amino acid and sugar/base/phosphate for the nucleotide. We also assigned secondary structures for the protein amino acids using the DSSP algorithm [57] through the Biopython Bio.PDB module [58]. We converted the DSSP output into three classes: helix (H, including DSSP H, G, and I categories), strand (E, including DSSP B and E categories) and coil (C, including DSSP T, S and - categories).

To determine the evolutionary divergence of each protein chain in our dataset, we generated multiple sequence alignments (MSAs) using 1 iteration of HHblits version 3.0.0 [59] against the Uniref30 [60] version of February 2022. We filtered the obtained MSAs with HHfilter [59] using a 30% minimum sequence identity with the query sequence and a diff parameter to 80, limiting the number of sequences while ensuring sequence diversity in the MSA. The Rate4Site software package was then used to calculate a conservation score for each position in the protein sequence [61, 62]. The calculated values were rescaled from 0 to 100, with higher values associated with more conserved residues likely indicative of functional importance. Rate4Site calculations failed for a subset of 38 interfaces in our dataset, and the corresponding interologs are therefore excluded from evolutionary conservation analysis.

We divided the protein interface according to the core-rim model [43]. We calculated the relative accessible surface area of the complex (rASAc) for the amino acids in each interface using the Python module of freesasa, using a probe radius of 1.4 Å. Interface rim protein residues are those that have rASAc > 25% and all other protein interface residues were assigned to the interface core region following previous work [43, 44]. Freesasa calculations failed for a subset of 141 interfaces in our dataset, and the corresponding interologs are therefore excluded from core/rim conservation analysis.

Structural alignment

We generated coordinates in the PDB file format for each of the 977 representative binary interfaces. Using PDBFixer from the openMM Python module [63], we converted non-standard amino

acids/nucleic acids to their standard counterparts (using the Chemical Component Dictionary from the PDB) and we added missing atoms. We then performed all-against-structural comparisons between the 977 interfaces, excluding pairs of interfaces that belong to the same PDB entry and have one chain in common. For the remaining approx. 445,000 pairs of interfaces, we used TM-align (Version 20190822) for protein structural alignment [64], RNA-align (Version 20191021) for RNA structural alignment [65], and the updated version of MM-align (Version 20191021) for protein-RNA interface structural alignment [66]. For each pair of compared interfaces, the software provided protein, RNA, and interface TM-scores normalized by each interface separately; we retained only the TM-scores normalized by the smallest molecule/interface.

Structural comparison metrics

We converted the alignment file from MM-align into a dictionary of structural correspondence listing structurally aligned pairs of amino acids and nucleotides from each pair of interfaces. The interface coverage was computed for protein and RNA as the number of interface amino acids or nucleotides that had a structural correspondent normalized by the size of the smallest (protein or RNA) interface. For any given pair of structurally compared interfaces, all following analyses of contact conservation were restricted to amino acids and nucleotides with structural correspondents in both interfaces.

After superimposing interfaces with MM-align, we calculated the interface RMSD (Root Mean Square Deviation), which measures the structural similarity between two protein-RNA complexes at their interface, using coordinates of the P atoms for interface nucleotides and CA atoms for interface amino acids.

In this study, we are using structural alignment as a gold standard since all considered interfaces have good resolution structural coordinates. Structure-based sequence identity was computed for two aligned protein chains or RNA chains based on structural interface alignment results, as the number of identical positions divided by the number of aligned positions. Interface sequence identity was

computed by considering only interface positions (as defined by the minimum heavy-atom 5Å distance criterion). In general, e.g. with the goal of interface template-based modeling or MSA-based predictions, interface structures might not be known in advance. Hence, as a reference, we also computed a sequence-based sequence identity by extracting reference sequences from the PDB (canonical sequences, where non-standard amino acids and nucleotides are converted to their standard counterpart as much as possible) and aligning separately protein sequences and RNA sequences for each pair of compared interfaces, using the FASTA36 program [67] as was done in the previous study of template-based protein-RNA interface modeling [12]. Because this can lead to very short aligned segments for dissimilar molecules, we weighted the sequence-based sequence identity by the sequence-based alignment coverage (number of aligned positions divided by the length of the largest molecule).

Evolutionary classification of RNA-binding domains and RNA chains

We organized the 2,022 pairs of interologs into groups. Using the networkx [68] Python package, we defined a graph where each of the 765 interfaces is a node and we added edges corresponding to pairs of structural interologs. Groups of interologs were defined as the connected components of this interolog graph. We obtained 141 groups containing between 2 and 29 interfaces.

ECOD (Evolutionary Classification of Protein Domains) [41] version 20230309 (develop288) was used to annotate protein chains with its hierarchical classification system based on evolutionary relationships and structural similarities [41, 69]. We annotated each protein chain involved in a binary protein-RNA interface according to all ECOD domains containing amino acids within 5 Å of the RNA chain. Interfaces with more than one RNA chain were labeled with the union of all ECOD domains including residues from all pairwise protein-RNA interfaces. For each pair of interologs, we checked whether the sets of ECOD domain labels for the two protein chains (when available) were identical or overlapping. We used the ECOD T-group level for comparison (groups of homologs with similar

topological connections). Each group of interologs was labeled with the most represented (or set of most represented) ECOD domains among all interfaces in the group; if no ECOD annotation was available or ECOD labeling was not unambiguously possible, then the group was labeled ambiguous.

Rfam [42] is an evolutionary classification of RNA families. Rfam families are grouped into clans when they are remotely homologous or when they can be aligned, but have distinct functions. We retrieved the mapping of Rfam (version 14.8) families and clans to all RNA chains belonging to the 765 interfaces in our dataset. We annotated pairs and groups of interologs with Rfam labels in the same manner as described for ECOD labeling.

Calculation of interface contact conservation

The analysis of protein-RNA interface contact conservation was conducted similarly to our previous study of protein-protein interface evolution [34]. Interface contact conservation between a pair of structural interologs (homologous interfaces) was computed only for contacts involving amino acids and nucleotides with structural correspondents in both interfaces. The conservation of each type of contact was calculated using the Jaccard index (similarity coefficient), which is the ratio of the number of conserved contacts (formed in both homologous interfaces) to the total number of contacts formed in at least one interface. For atomic and apolar contacts, this ratio was weighted by the average number of atomic contacts between two positions (only for C atoms in the case of apolar contacts) if the contact existed in both interologs. If it only existed in one interolog, the ratio was weighted by the number of atomic contacts in the interolog where it existed. We also computed the conservation of H-bonds, salt bridges and π-stacking contacts assigned by SNAP by using the Jaccard index (no weighting).

For each pair of interologs, we computed the fraction of amino acids (respectively, nucleotides) switching out of the interface, by weighing each amino acid (respectively nucleotide) by the sum of all atomic contacts it makes within its respective interface. For each pair of interologs, the fraction of non-conserved contacts linked to switching out was computed as the unweighted ratio of the number of

non-conserved contacts where the amino acid and/or the nucleotide switch out of the interface compared to all non-conserved contacts.

For all structural visualization aspects in this manuscript, we used the ChimeraX software [70].

Web interface for interolog exploration and visualization

The web server hosted at https://bioi2.paris-saclay.fr/django/rnaprotdb/ provides an interactive and comparative view of all 2,022 pairs of interologs in our database. It was generated using Django version 4.1.7 [71] and uses NGL Viewer [72, 73] to allow users to freely manipulate and explore the interolog structures and their pre-calculated features. Additionally, users can interactively explore interolog group networks thanks to the Cytoscape.js plugin version 3.28.1 [74].

Acknowledgments

We thank R. Guerois and D. J. Zea for fruitful discussions and feedback on the manuscript. We thank E. Frezza and A. Taly for early feedback on the project. We thank M. Wojdyr for help with Gemmi and R. D. Schaeffer for help with ECOD. The initial code for the web interface to visualize interolog structures was inspired from the visualization interface developed by J. Rey together with P. Tufféry for the InterEvDock3 web server.

References

1. Krüger DM, Neubacher S, Grossmann TN. Protein–RNA interactions: structural characteristics and hotspot amino acids. RNA. 2018;24(11):1457-65. doi: 10.1261/rna.066464.118.
2. Jones S. Protein-RNA interactions: structural biology and computational modeling techniques. Biophys Rev. 2016;8(4):359-67. doi: 10.1007/s12551-016-0223-9.
3. Salta E, De Strooper B. Noncoding RNAs in neurodegeneration. Nature Reviews Neuroscience. 2017;18(10):627-40. doi: 10.1038/nrn.2017.90.
4. Daubner GM, Clery A, Jayne S, Stevenin J, Allain FH. A syn-anti conformational difference allows SRSF2 to recognize guanines and cytosines equally well. EMBO J. 2012;31(1):162-74. doi: 10.1038/emboj.2011.367.

5. Zhang J, Lieu YK, Ali AM, Penson A, Reggio KS, Rabadan R, et al. Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. Proc Natl Acad Sci U S A. 2015;112(34):E4726-34. doi: 10.1073/pnas.1514105112.
6. Warner KD, Hajdin CE, Weeks KM. Principles for targeting RNA with drug-like small molecules. Nature Reviews Drug Discovery. 2018;17(8):547-58. doi: 10.1038/nrd.2018.93.
7. Childs-Disney JL, Yang X, Gibaut QMR, Tong Y, Batey RT, Disney MD. Targeting RNA structures with small molecules. Nature Reviews Drug Discovery. 2022;21(10):736-62. doi: 10.1038/s41573-022-00521-4.
8. Miao Z, Westhof E. RNA Structure: Advances and Assessment of 3D Structure Prediction. Annu Rev Biophys. 2017;46:483-503. doi: 10.1146/annurev-biophys-070816-034125.
9. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Research. 2023;51(D1):D488-D508. doi: 10.1093/nar/gkac1077.
10. Nithin C, Ghosh P, Bujnicki JM. Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes. Genes (Basel). 2018;9(9). doi: 10.3390/genes9090432.
11. Liu X, Duan Y, Hong X, Xie J, Liu S. Challenges in structural modeling of RNA-protein interactions. Current Opinion in Structural Biology. 2023;81:102623. doi: 10.1016/j.sbi.2023.102623.
12. Zheng J, Kundrotas PJ, Vakser IA, Liu S. Template-Based Modeling of Protein-RNA Interactions. PLoS Comput Biol. 2016;12(9):e1005120. doi: 10.1371/journal.pcbi.1005120.
13. Li CH, Cao LB, Su JG, Yang YX, Wang CX. A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins. 2012;80(1):14-24. doi: 10.1002/prot.23117.
14. Huang S-Y, Zou X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucleic Acids Research. 2014;42(7):e55-e. doi: 10.1093/nar/gku077.
15. Pérez-Cano L, Solernou A, Pons C, Fernández-Recio J. STRUCTURAL PREDICTION OF PROTEIN-RNA INTERACTION BY COMPUTATIONAL DOCKING WITH PROPENSITY-BASED STATISTICAL POTENTIALS. Biocomputing 2010: WORLD SCIENTIFIC; 2009. p. 293-301.
16. Setny P, Zacharias M. A coarse-grained force field for Protein-RNA docking. Nucleic Acids Research. 2011;39(21):9118-29. doi: 10.1093/nar/gkr636.
17. Guilhot-Gaudeffroy A, Froidevaux C, Azé J, Bernauer J. Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities. PLoS ONE. 2014;9(9):e108928. doi: 10.1371/journal.pone.0108928.
18. Perez-Cano L, Romero-Durana M, Fernandez-Recio J. Structural and energy determinants in protein-RNA docking. Methods. 2017;118-119:163-70. doi: 10.1016/j.ymeth.2016.11.001.
19. Yan Y, Zhang D, Zhou P, Li B, Huang SY. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. 2017;45(W1):W365-W73. doi: 10.1093/nar/gkx407.
20. van Zundert GCP, Rodrigues J, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. 2016;428(4):720-5. doi: 10.1016/j.jmb.2015.09.014.
21. Teichmann SA. The constraints protein-protein interactions place on sequence divergence. J Mol Biol. 2002;324(3):399-407. doi: 10.1016/s0022-2836(02)01144-0.
22. Andreani J, Quignot C, Guerois R. Structural prediction of protein interactions and docking using conservation and coevolution. WIREs Computational Molecular Science. 2020;10(6):e1470. doi: 10.1002/wcms.1470.
23. Durham J, Zhang J, Humphreys IR, Pei J, Cong Q. Recent advances in predicting and modeling protein–protein interactions. Trends in Biochemical Sciences. 2023;48(6):527-38. doi: 10.1016/j.tibs.2023.03.003.

24. Wang K, Hu G, Wu Z, Su H, Yang J, Kurgan L. Comprehensive Survey and Comparative Assessment of RNA-Binding Residue Predictions with Analysis by RNA Type. Int J Mol Sci. 2020;21(18). doi: 10.3390/ijms21186879. PubMed PMID: 32961749; PubMed Central PMCID: PMCPMC7554811.
25. Walia RR, Xue LC, Wilkins K, El-Manzalawy Y, Dobbs D, Honavar V. RNABindRPlus: A Predictor that Combines Machine Learning and Sequence Homology-Based Methods to Improve the Reliability of Predicted RNA-Binding Residues in Proteins. PLOS ONE. 2014;9(5):e97725. doi: 10.1371/journal.pone.0097725.
26. Rivas E. Evolutionary conservation of RNA sequence and structure. Wiley Interdiscip Rev RNA. 2021;12(5):e1649. doi: 10.1002/wrna.1649.
27. Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS. 3D RNA and Functional Interactions from Evolutionary Couplings. Cell. 2016;165(4):963-75. doi: 10.1016/j.cell.2016.03.030.
28. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. doi: 10.1038/s41586-021-03819-2.
29. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2022:2021.10.04.463034. doi: 10.1101/2021.10.04.463034.
30. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871-6. doi: 10.1126/science.abj8754.
31. Baek M, McHugh R, Anishchenko I, Baker D, DiMaio F. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. bioRxiv. 2022:2022.09.09.507333. doi: 10.1101/2022.09.09.507333.
32. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science. 2024;384(6693):eadl2528. doi: 10.1126/science.adl2528. PubMed PMID: 38452047.
33. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024. doi: 10.1038/s41586-024-07487-w. PubMed PMID: 38718835.
34. Andreani J, Faure G, Guerois R. Versatility and invariance in the evolution of homologous heteromeric interfaces. PLoS Comput Biol. 2012;8(8):e1002677. doi: 10.1371/journal.pcbi.1002677.
35. Andreani J, Faure G, Guerois R. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. Bioinformatics. 2013;29(14):1742-9. doi: 10.1093/bioinformatics/btt260.
36. Quignot C, Granger P, Chacon P, Guerois R, Andreani J. Atomic-level evolutionary information improves protein-protein interface scoring. Bioinformatics. 2021. doi: 10.1093/bioinformatics/btab254.
37. Quignot C, Postic G, Bret H, Rey J, Granger P, Murail S, et al. InterEvDock3: a combined template-based and free docking server with increased performance through explicit modeling of complex homologs and integration of covariation-based contact maps. Nucleic Acids Res. 2021;49(W1):W277-W84. doi: 10.1093/nar/gkab358.
38. Treger M, Westhof E. Statistical analysis of atomic contacts at RNA-protein interfaces. Journal of molecular recognition: JMR. 2001;14(4):199-214. doi: 10.1002/jmr.534.
39. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein-RNA interactions: a structural analysis. Nucleic Acids Research. 2001;29(4):943-54. doi: 10.1093/nar/29.4.943.
40. Corley M, Burns MC, Yeo GW. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Mol Cell. 2020;78(1):9-29. doi: 10.1016/j.molcel.2020.03.011.
41. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: An Evolutionary Classification of Protein Domains. PLoS Computational Biology. 2014;10(12):e1003926. doi: 10.1371/journal.pcbi.1003926.
42. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Research. 2021;49(D1):D192-D200. doi: 10.1093/nar/gkaa1047.

43. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol. 2010;403(4):660-70. doi: 10.1016/j.jmb.2010.09.028.
44. Teppa E, Zea DJ, Marino-Buslje C. Protein-protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces. Protein Sci. 2017;26(12):2438-44. doi: 10.1002/pro.3318.
45. Chen Y. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. Nucleic Acids Research. 2004;32(17):5147-62. doi: 10.1093/nar/gkh785.
46. Kagra D, Jangra R, Sharma P. Exploring the Nature of Hydrogen Bonding between RNA and Proteins: A Comprehensive Analysis of RNA : Protein Complexes. ChemPhysChem. 2022;23(2):e202100731. doi: 10.1002/cphc.202100731.
47. Roca-Martínez J, Dhondge H, Sattler M, Vranken WF. Deciphering the RRM-RNA recognition code: A computational analysis. PLOS Computational Biology. 2023;19(1):e1010859. doi: 10.1371/journal.pcbi.1010859.
48. Blanco C, Bayas M, Yan F, Chen IA. Analysis of Evolutionarily Independent Protein-RNA Complexes Yields a Criterion to Evaluate the Relevance of Prebiotic Scenarios. Current Biology. 2018;28(4):526-37.e5. doi: 10.1016/j.cub.2018.01.014.
49. Giacobelli VG, Fujishima K, Lepsik M, Tretyachenko V, Kadava T, Makarov M, et al. In Vitro Evolution Reveals Noncationic Protein-RNA Interaction Mediated by Metal Ions. Mol Biol Evol. 2022;39(3). doi: 10.1093/molbev/msac032. PubMed PMID: 35137196; PubMed Central PMCID: PMCPMC8892947.
50. Wojdyr M. GEMMI: A library for structural biology. Journal of Open Source Software. 2022;7(73):4200. doi: 10.21105/joss.04200.
51. Lu XJ, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 2003;31(17):5108-21. doi: 10.1093/nar/gkg680. PubMed PMID: 12930962; PubMed Central PMCID: PMCPMC212791.
52. Lu XJ, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc. 2008;3(7):1213-27. doi: 10.1038/nprot.2008.104.
53. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026-8. doi: 10.1038/nbt.3988.
54. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-2. doi: 10.1093/bioinformatics/bts565.
55. Bickerton GR, Higueruelo AP, Blundell TL. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. BMC Bioinformatics. 2011;12(1):313. doi: 10.1186/1471-2105-12-313.
56. Yu B, Pettitt BM, Iwahara J. Dynamics of Ionic Interactions at Protein–Nucleic Acid Interfaces. Accounts of Chemical Research. 2020;53(9):1802-10. doi: 10.1021/acs.accounts.0c00212.
57. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637. doi: 10.1002/bip.360221211.
58. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422-3. doi: 10.1093/bioinformatics/btp163.
59. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9(2):173-5. doi: 10.1038/nmeth.1818.
60. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium tU. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926-32. doi: 10.1093/bioinformatics/btu739.
61. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within

their homologues. Bioinformatics. 2002;18 Suppl 1:S71-7. doi: 10.1093/bioinformatics/18.suppl_1.s71.
62. Mayrose I, Graur D, Ben-Tal N, Pupko T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol. 2004;21(9):1781-91. doi: 10.1093/molbev/msh194.
63. Eastman P, Pande V. OpenMM: A Hardware-Independent Framework for Molecular Simulations. Computing in Science & Engineering. 2010;12(4):34-9. doi: 10.1109/MCSE.2010.27.
64. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302-9. doi: 10.1093/nar/gki524.
65. Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. Bioinformatics. 2019;35(21):4459-61. doi: 10.1093/bioinformatics/btz282.
66. Mukherjee S, Zhang Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res. 2009;37(11):e83. doi: 10.1093/nar/gkp318.
67. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences. 1988;85(8):2444-8. doi: 10.1073/pnas.85.8.2444.
68. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. 2008.
69. Cheng H, Liao Y, Schaeffer RD, Grishin NV. Manual classification strategies in the ECOD database: ECOD Manual Classification Strategies. Proteins: Structure, Function, and Bioinformatics. 2015;83(7):1238-51. doi: 10.1002/prot.24818.
70. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX : Structure visualization for researchers, educators, and developers. Protein Science. 2021;30(1):70-82. doi: 10.1002/pro.3943.
71. Django (Version 4.1.7) [Computer Software]. (2023).
72. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW, editors. Web-based molecular graphics for large complexes2016 juillet 22, 2016. New York, NY, USA: Association for Computing Machinery.
73. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlic A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 2018;34(21):3755-8. doi: 10.1093/bioinformatics/bty419. PubMed PMID: 29850778; PubMed Central PMCID: PMCPMC6198858.
74. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 2016;32(2):309-11. doi: 10.1093/bioinformatics/btv557. PubMed PMID: 26415722; PubMed Central PMCID: PMCPMC4708103.