A strategy for refining the calculation of contacts in protein-RNA complexes

Luana Luiza Bastos[1*], Rafael P. Lemos[1*], Diego Mariano[1*], Raquel C. de Melo-Minardi [1]

[1]Laboratory of Bioinformatics and Systems (LBS)
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
*These authors have contributed equally

raquelcm@dcc.ufmg.br,luizabastos.luana9@gmail.com

Abstract. *Protein-RNA interactions are essential for several biological processes, including gene expression. However, traditional methods for studying these interactions use superficial criteria to perform this analysis, which can lead to false positives. This study presents a new strategy for modeling protein-RNA contacts. We classify RNA atoms and integrate methods previously used for protein contacts, developing the proposed approach that detects a broader range of interactions. We compare our proposal to an existing benchmark and observe that this method identifies more contacts and provides detailed insights into different types of interactions, such as aromatic stacking and hydrogen bonding. We envision that the strategy could improve the development of protein-RNA interaction databases and deepen our understanding of these complexes.*

1. Introduction

Ribonucleic acid (RNA) is a fundamental macromolecule for the production of proteins. Furthermore, the different types of this molecule play numerous vital functions in cellular physiology, such as regulating gene expression [Corley et al. 2020]. Protein-RNA interactions are essential for regulating several other biological processes in living beings, such as RNA splicing and protein translation. For this reason, the scientific community has made a great effort to study their functions and mechanisms of action. For example, many of these functions performed by RNA in organisms depend on interactions with specific proteins known as RBPs (RNA-binding proteins) [Steinmetz et al. 2023]. Works such as those by [Kang et al. 2020] and [Gebauer et al. 2021] have discussed how a detailed understanding of the mechanisms of function of RBPs is fundamental to understanding genetic diseases and how understanding molecular mechanisms can facilitate the development of therapeutic interventions [Lodde et al. 2023]. However, understanding the functioning of protein-RNA biological processes through experimental methods is expensive and time-consuming. Therefore, computational experiments, such as structural bioinformatics techniques, are being increasingly adopted.

Structural bioinformatics is a field of research that aims to understand the interaction of macromolecules [Pimentel et al. 2021], such as interactions in protein-RNA complexes. Also, it provides tools for analyzing atomic contacts *in silico*. Contacts are representations of biochemical interactions between molecules calculated using computational algorithms. Contacts are weak, stabilizing interactions in the structure of macromolecules

and complexes, such as protein-RNA [Silva et al. 2019, Fassio et al. 2022]. Therefore, they can be used to better understand protein-RNA interactions.

In the literature, we can find examples of previous works of contact analysis for protein-RNA complexes [Jones et al. 2001, Han and Nepal 2007, Treger and Westhof 2001, Puton et al. 2012]. Understanding the contact points between protein-RNA complexes can reveal how proteins and RNAs interact mechanistically. Although there is interest in these complexes, the study of these complexes has yet to be thoroughly explored for protein-protein, protein-peptide, and protein-ligand. In this sense, refining the analysis of protein-RNA contacts is an essential tool for better understanding the interaction of these macromolecules [Zuo et al. 2024, Medina-Munoz et al. 2024].

This paper presents a strategy to refine the calculation of protein-RNA contacts. Initially, we use the LUNA [Fassio et al. 2022] library to predict the types of nucleotide atoms. Then, we use the definition of amino acid atomic types used by the nAPOLI tool [Fassio et al. 2019] and combine it with the contact cutoffs used by the VTR tool [Pimentel et al. 2021]. Our methodology can obtain more refined results for calculating interactions between proteins and RNA molecules. To evaluate our proposal, we collected a sub-dataset of protein-RNA complexes from the Protein Data Bank and compared it with the study by [Puton et al. 2012]. Figure 1 summarizes the workflow adopted in this study.

2. Material and Methods

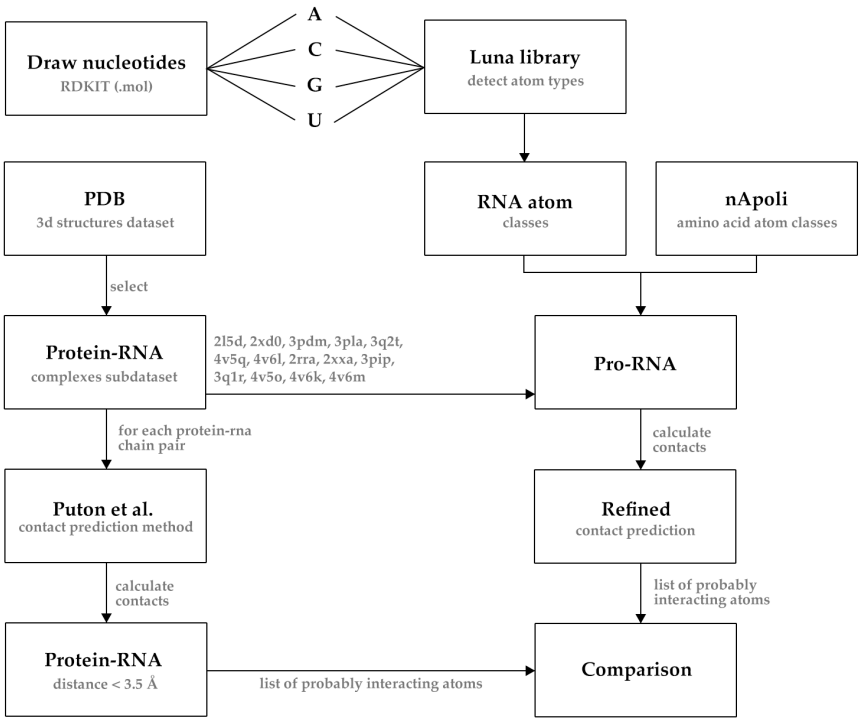

Figure 1. Methodology adopted in this study. RNA atom types were detected using LUNA. Pro-RNA combines this knowledge with protein atom types described by nAPOLI and we use the cutoff distances used by the VTR tool. Finally, we compare our method with the benchmark of Puton *et al.*

2.1. Modeling RNA Contact Types

We modeled the structure of RNA molecules using the Python rdkit library [Landrum et al. 2013]. For each of the four nucleotides (A, C, G, U), we used LUNA to predict the types of each atom. LUNA returns the possible classes: acceptor, donor, hydrophobic, and aromatic. These classes indicate the possible types of interactions each atom can perform. Then, we determined each atom type in amino acids based on the nAPOLI tool definitions [Fassio et al. 2019]. We combined this information to calculate three types of contacts: hydrogen bonds, aromatic stacking, and hydrophobic contacts.

We used the same definitions presented by VTR for distance-based contact calculations [Pimentel et al. 2021]. Contacts were calculated using COCαDA [Lemos et al. 2024][1]. Hydrogen bonds occur when an acceptor atom is within 3.9Å of a donor atom (the bond angle criterion was not adopted). Hydrophobic interactions occur between hydrophobic atoms at 2 to 4.5Å. Aromatic stackings occur when the centroid distance of two aromatic rings is between 2 and 5 Å. The angle θ between the two normal vectors of the rings is then calculated, and the interaction is considered parallel if $160° \leq \theta < 180°$ or $0° \leq \theta < 20°$, or perpendicular if $80° \leq \theta < 100°$ (Table 1).

Table 1. Types of contact and atom pair threshold range (in Å).

Contact Type	Range (Å)
Hydrogen Bond	$0 \leq dist \leq 3.9$
Hydrophobic	$2.0 \leq dist \leq 4.5$
Aromatic Stacking	$2.0 \leq dist \leq 5.0$

2.2. Comparison to Other Approaches

We compared and evaluated our method against the study by [Puton et al. 2012]. In their research, Puton and collaborators collected three-dimensional structures of protein-RNA complexes from the Protein Data Bank (PDB) [Berman 2000]. They then filtered pairs of protein and RNA chains and calculated the contacts. They defined that a polypeptide chain interacts with an RNA molecule if one of the residues is at a distance less than 3.5 Å from any nucleotide. These cutoff patterns are close to the definitions of van der Waals interactions: a set of interactions of a weak nature when compared to covalent interactions or even hydrogen bonds. However, we believe that this simplistic strategy may lead to false positives. To verify this, we collected the same dataset indicated by [Puton et al. 2012] and applied our method for calculating contacts.

2.2.1. Data Collection

Protein-RNA complexes were extracted from the PDB, based on the supplementary material description of [Puton et al. 2012]. Obsolete structures were replaced by the new structures indicated by the PDB. Hence, we collected the PDB IDs: 2L5D, 2XD0, 3PDM, 3PLA, 3Q2T, 4V5Q, 4V6L, 2RRA, 2XXA, 3PIP, 3Q1R, 4V5O, 4V6K, and 4V6M.

3. Results and Discussion

Figure 2 presents a visualization of the atom types predicted by our method (herein called Pro-RNA). Note that all atoms in the aromatic rings were defined as aromatic, although aromatic stackings were calculated between the centroid distance of the atoms of two

[1] Available at https://github.com/LBS-UFMG/COCaDA/.

rings. Only cytosine and uracil can perform hydrophobic interactions (via the C5 atom). The O2' atom can act both as a hydrogen acceptor and donor. On the other hand, O3' was predicted to perform the same interactions when the nucleotides are not connected. When bonded to the P atom of another nucleotide, O3' can act only as an acceptor. The N1 atom of guanine has been classified as both aromatic, acceptor, and donor.

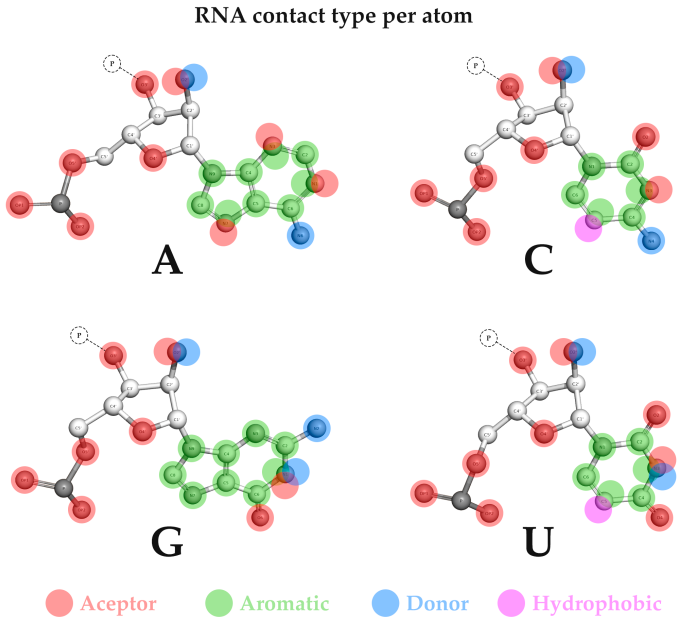

Figure 2. Atom types calculated using LUNA library [Fassio et al. 2022].
A = adenine, C = cytosine, G = guanine, U = uracil. Red = acceptor atoms, Green = aromatic atoms, Blue = donor atoms, Magenta = hydrophobic atoms.

3.1. Method Evaluation

The authors use a simplified approach to calculate contacts in the work by [Puton et al. 2012]. They define an RNA molecule as interacting with a protein if a pair of atoms in each molecule is within a distance of 3.5 Å. This criterion does not consider the atomic types or the types of interactions that certain atoms can perform, so we believe that it does not reflect well the interactions performed between protein and RNA.

After reimplementing the experiments by [Puton et al. 2012] with the original dataset and comparing them with our methodology, Pro-RNA was able to detect more relevant interactions, such as hydrogen bonds (Figure 3, left). Both methods detect several contacts in common (16,240, Figure 3, right), but there are also exclusive contacts (3,703 for Pro-RNA and 6,201 for Puton *et al.*). When we evaluate contacts considering pairs of residues, Puton *et al.*'s method presents more detected contacts (Figure 3, right). However, these contacts tend to be van der Waals (London Dispersal Forces). According to [Neshich et al. 2005], these van der Waals interactions have a contact energy of 0.08 kcal/mol. On the other hand, hydrogen bonds, aromatic stacking, and hydrophobic interactions have contact energies of 2.6 kcal/mol, 1.5 kcal/mol, and 0.6 kcal/mol, respectively. Therefore, despite being considered weak, these interactions stabilize protein-RNA bonds. Thus, our method was the only one capable of detecting these interactions.

For example, for PDB 2RRA, Pro-RNA was able to detect an interaction between M381 (chain A) and A505 (chain B) (Figure 4A). This hydrogen bond occurs near the

	Pro-RNA (Atom)	Pro-RNA (Residue)	Puton *et al.* (Residue)
Total	31702	19943	22441
HY	915	-	-
HB	30616	-	-
AS	170	-	-

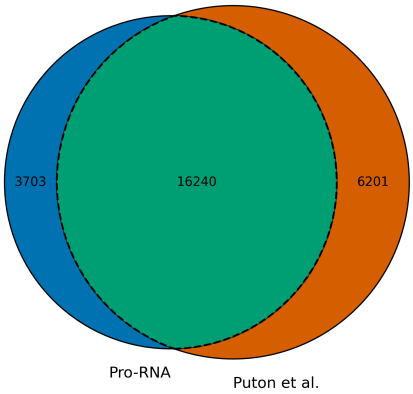

Figure 3. Contacts calculated using Pro-RNA vs. [Puton et al. 2012] approach. Atom = analysis at the atom level, with contact type classification. Residue = analysis at the residue level. HY = hydrophobic, HB = hydrogen bond, AS = aromatic stacking. The Venn diagram shows common and exclusive contacts on the residue level.

cutoff distance limit (3.8 Å), so it is not captured by the method of Puton *et al.* (which limits the cutoff distance to 3.5 Å). Furthermore, note that Pro-RNA was able to find aromatic stacking interactions, such as the one between A:F123 and B:G4 (PDB: 2RRA) (Figure 4B), and even hydrophobic interactions, such as A:U4 and J:Q96 (PDB: 4V6M) (Figure 4C). The complete list of contacts can be found in the supplementary material.

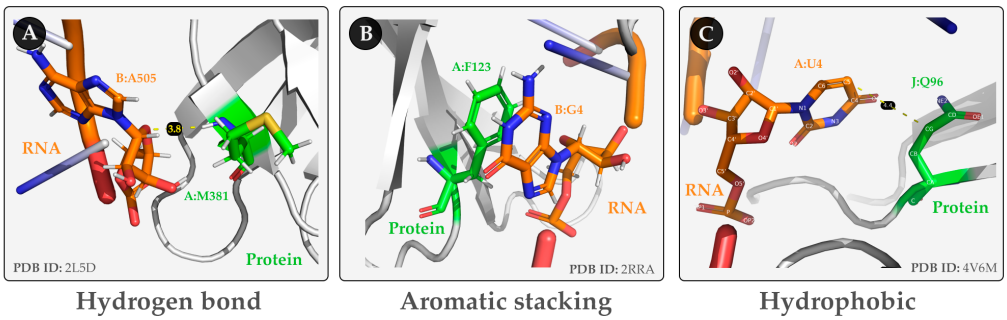

Figure 4. Examples of contacts calculated by Pro-RNA. (A) Hydrogen bond; (B) Aromatic stacking; (C) Hydrophobic interaction.

4. Conclusion

This work presents a strategy to refine the calculation of contacts between proteins and RNA. Our approach can be used to better understand the interactions between these different molecules. However, this work had limitations, such as the size of the dataset used in the case study. Observing how the developed method performs using a more extensive and comprehensive dataset would be possible. This dataset was chosen to compare with the work of Puton *et al.* [Puton et al. 2012] . We considered that a larger dataset would be necessary to obtain better insights. In a larger dataset, it would be possible to evaluate the behavior of the method for different sizes of protein-RNA complexes. With the resolution of increasingly larger complexes, such as ribosomal structures, and the need to understand the interactions performed by protein complexes, we observed the need to optimize the method computationally. Therefore, we have prospects for future work in establishing a database of protein-RNA complexes.

Supplementary material Supplementary Material is available at https://github.com/LBS-UFMG/Pro-RNA

Acknowledgements The authors would like to thank the research funding agencies CAPES, FAPEMIG, and CNPq. The authors also thank Alexandre Fassio for his support in using the LUNA library and Alessandra Cioletti for her valuable contributions. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Berman, H. M. (2000). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

Corley, M., Burns, M. C., and Yeo, G. W. (2020). How rna-binding proteins interact with rna: molecules and mechanisms. *Molecular cell*, 78(1):9–29.

Fassio, A. V., Santos, L. H., Silveira, S. A., Ferreira, R. S., and de Melo-Minardi, R. C. (2019). napoli: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(4):1317–1328.

Fassio, A. V., Shub, L., Ponzoni, L., McKinley, J., O'Meara, M. J., Ferreira, R. S., Keiser, M. J., and de Melo Minardi, R. C. (2022). Prioritizing virtual screening with interpretable interaction fingerprints. *Journal of Chemical Information and Modeling*, 62(18):4300–4318.

Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M. W. (2021). Rna-binding proteins in human genetic disease. *Nature Reviews Genetics*, 22(3):185–198.

Han, K. and Nepal, C. (2007). Pri-modeler: extracting rna structural elements from pdb files of protein–rna complexes. *FEBS letters*, 581(9):1881–1890.

Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M., and Thornton, J. M. (2001). Protein–rna interactions: a structural analysis. *Nucleic acids research*, 29(4):943–954.

Kang, D., Lee, Y., and Lee, J.-S. (2020). Rna-binding proteins in cancer: functional and therapeutic perspectives. *Cancers*, 12(9):2699.

Landrum, G. et al. (2013). Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281.

Lemos, R. P., Mariano, D., Azevedo, S., and Melo-Minardi, R. (2024). Cocαda - large-scale protein interatomic contact cutoff optimization by cα distance matrices. In *Proceedings of the Brazilian Symposium on Bioinformatics, 2024*. Accepted for publication.

Lodde, V., Floris, M., Zoroddu, E., Zarbo, I. R., and Idda, M. L. (2023). Rna-binding proteins in autoimmunity: From genetics to molecular biology. *Wiley Interdisciplinary Reviews: RNA*, 14(4):e1772.

Medina-Munoz, H. C., Kofman, E., Jagannatha, P., Boyle, E. A., Yu, T., et al. (2024). Expanded palette of rna base editors for comprehensive rbp-rna interactome studies. *Nature Communications*, 15(1):875.

Neshich, G., Mancini, A. L., Yamagishi, M. E., Kuser, P. R., et al. (2005). Sting report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the sting database. *Nucleic acids research*, 33(suppl_1):D269–D274.

Pimentel, V., Mariano, D., Cantão, L. X. S., Bastos, L. L., Fischer, P., de Lima, L. H. F., Fassio, A. V., and Melo-Minardi, R. C. d. (2021). Vtr: a web tool for identifying analogous contacts on protein structures and their complexes. *Frontiers in Bioinformatics*, 1:730350.

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., and Bujnicki, J. M. (2012). Computational methods for prediction of protein–rna interactions. *Journal of structural biology*, 179(3):261–268.

Silva, M. F., Martins, P. M., Mariano, D. C., Santos, L. H., Pastorini, I., Pantuza, N., Nobre, C. N., de Melo-Minardi, R. C., and Players, P. (2019). Proteingo: motivation, user experience, and learning of molecular interactions in biological complexes. *Entertainment Computing*, 29:31–42.

Steinmetz, B., Smok, I., Bikaki, M., and Leitner, A. (2023). Protein–rna interactions: from mass spectrometry to drug discovery. *Essays in Biochemistry*, 67(2):175–186.

Treger, M. and Westhof, E. (2001). Statistical analysis of atomic contacts at rna–protein interfaces. *Journal of Molecular Recognition*, 14(4):199–214.

Zuo, Y., Chen, H., Yang, L., Chen, R., Zhang, X., and Deng, Z. (2024). Research progress on prediction of rna-protein binding sites in the past five years. *Analytical Biochemistry*, page 115535.