



A Knowledge-Based Potential for Protein-RNA Docking

R. P. Bahadur, M. Zacharias

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 157-160, 2008.

© 2008 by John von Neumann Institute for Computing
Permission to make digital or hard copies of portions of this work for
personal or classroom use is granted provided that the copies are not
made or distributed for profit or commercial advantage and that copies
bear this notice and the full citation on the first page. To copy otherwise
requires prior specific permission by the publisher mentioned above.

<http://www.fz-juelich.de/nic-series/volume40>

A Knowledge-Based Potential for Protein-RNA Docking

Ranjit P. Bahadur and Martin Zacharias

School of Engineering and Science, Jacobs University Bremen,
Campus Ring 1, D-28759, Bremen, Germany
E-mail: {m.zacharias, r.bahadur}@jacobs-university.de

Protein-RNA interactions play an important role in all cellular processes and it is important to understand the driving forces that govern this interaction. The mechanism by which a protein molecule specifically recognizes a RNA molecule in the cellular environment is not completely known. Here we developed a pair-potential function from the analysis of the 81 non-redundant atomic structures of protein-RNA complexes taken from the Protein Data Bank. This function helps us to understand the specificity of the interactions and could be useful in a protein-RNA docking algorithm where one tries to predict the correct complex structure starting from the individual components.

1 Introduction

Protein and RNA often interact in the cellular environment to perform essential cellular functions such as expression of a gene and its regulation. They can form binary complexes, for example, the aminoacyl-tRNA synthetases bind specific tRNAs for the translation of the genetic code; or multiple RNA and protein molecules can build a complicated cellular machine like a ribosome used for protein synthesis. To understand the functional mechanism of these complexes we have to elucidate the specificity of their interactions.

Several studies have been carried out recently to understand the structural basis of protein-RNA recognition [1-5]. All these methods consider the detailed atomic structures of the biomolecules. Here, we present an alternative approach to represent the protein and RNA chains in a reduced coarse-grained model, where each amino acid is represented by up to four pseudo atoms and each nucleotide by up to five pseudo atoms. We have calculated the pairwise contacts between the pseudo atoms of polypeptide and nucleotide chains and used them to derive a knowledge-based potential from a non-redundant dataset of 81 protein-RNA complexes recently compiled by Bahadur et al. [5]. Furthermore, the potential was included in a protein-RNA docking algorithm which can be used to predict complex structures starting from the individual structures of protein and RNA.

2 Materials and Methods

The dataset consist of 81 non-redundant known protein-RNA complexes taken from the PDB [6]. We have first translated the protein and RNA subunits into a reduced pseudo atom model. In case of the protein the same representation as implemented in the Attract docking program [7] was used. Briefly, each amino acid residues are represented by up to four pseudo atoms, two for main chain (N and O) and two for side chains. The side chains of Ala, Asn, Asp, Cys, Ile, Leu, Pro, Ser, Thr and Val are represented by one pseudo atom located at the center of geometry of all side-chain heavy atoms. Other larger side chains are represented by two pseudo atoms. The main chains pseudo atoms for all residues are rep-

Amino acids	Pseudo atoms
Ala	N, O, SC1
Arg	N, O, SC1, SC2
Asn	N, O, SC1
Asp	N, O, SC1
Cys	N, O, SC1
Gln	N, O, SC1, SC2
Glu	N, O, SC1, SC2
Gly	N, O, CA
His	N, O, SC1, SC2
Ile	N, O, SC1
Leu	N, O, SC1
Lys	N, O, SC1, SC2
Met	N, O, SC1, SC2
Phe	N, O, SC1, SC2
Pro	N, O, SC1
Ser	N, O, SC1
Thr	N, O, SC1
Trp	N, O, SC1, SC2
Tyr	N, O, SC1, SC2
Val	N, O, SC1
Nucleotides	
A	P, S, A1, A2, A3
U	P, S, U1, U2, U3
G	P, S, G1, G2, G3
C	P, S, C1, C2, C3

Table 1: Pseudo atoms for protein-RNA complexes. Each amino acid is represented by two main chain pseudo atoms (N and O) and maximum of two side chain pseudo atoms (SC1 and SC2). Each nucleotide is represented by five pseudo atoms, one each for phosphate and sugar molecules and three for the base. Gly has one extra main chain pseudo atom CA.

3 Results and Discussion

The pairwise contact potential between two pseudo atoms of protein and RNA is shown in figure 1. Aromatic residues show no preference to interact with the sugar (S) or phosphate (P) groups in the nucleotide but the interaction with the nucleobases is very favorable with few exceptions. This is due to possible stacking interactions between the aromatic ring and bases that may help to stabilize a protein-RNA complex. However, no interaction between the side chain pseudo atoms of Trp and the pseudo atoms of Uracyl base were found. Similar to aromatic residues, aliphatic (hydrophobic) residues (Ile, Leu and Val) show a preferential interaction with the nucleotide bases. Positively charged residues Arg and Lys prefer to interact with the negatively charged phosphate groups but do not interact favorably

represented by the N and O atoms except in Gly where an additional main chain CA atom is used. In case of RNA chain only one pseudo atom used for phosphate and sugar molecules and three for each bases (Table 1).

A pairwise interaction is counted if the distance between pseudo atoms of protein and RNA is within 4.5 Å. We computed the frequency of all pairwise interactions for the whole 81 complexes and converted them into a contact potential using the following equation:

$$V(P_i N_j) = -RT \ln \frac{\sum_{81} (P_i N_j)}{\sum_{81} P_i * \sum_{81} N_j}$$

Where $P_i N_j$ is the observed frequency of a particular atom pair of protein and RNA that are within the cut-off distance given above, and P_i is the frequency of the i^{th} protein atom interacting with RNA atoms and N_j is the frequency of the j^{th} RNA atom interacting with protein atoms (in the data set). The contact potential for each pair represents the energy minimum or saddle point of a Lennard-Jones (LJ) type potential (as implemented in Attract, [7]). The minimum pairwise distance between pseudo atom pairs represents the effective contact radius in the LJ-potential.

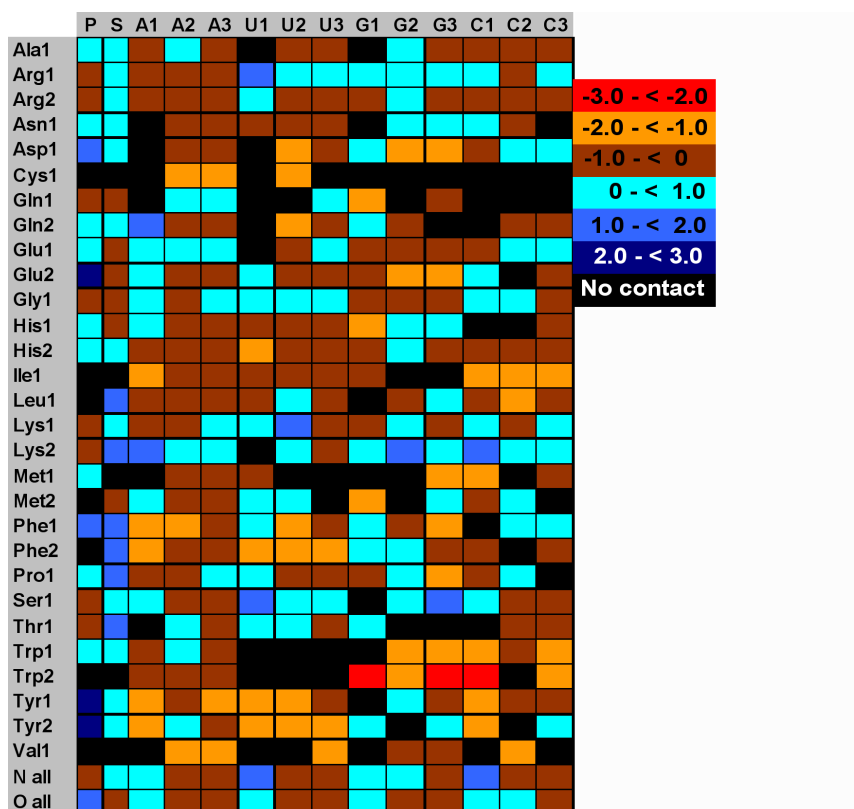


Figure 1. Pair-potential at the protein-RNA interfaces. Detail of the atom types is given in Table 1. If there is no interaction between an atom pair then that cell is colored black. Values should be multiplied by 10^{-3} in real scale.

with the ribose group. In addition, Arg interacts preferentially with the Adenine base but only moderately with the other three bases. Aspartic acid, being negatively charged, has less favorable interactions with P and S but it interacts favorably with the bases. Another negatively charged residue Glu has less preference to interact with P but more to S. Main chain atoms of amino acid residues have a mixed preference to interact with the nucleotide atoms. The knowledge-based potential has been integrated in the flexible docking program ATTRACT which employs energy minimization in translational and rotational degrees of freedom of the interacting partners [7]. Initial tests indicate that the potential can reproduce in many cases near-native protein-RNA complexes in good agreement with experimental complex structures (Figure 2).

4 Concluding Remarks

The approach was already used to predict a protein-RNA complex given in the CAPRI challenge [8] starting from two unbound structures which is under evaluation. We are

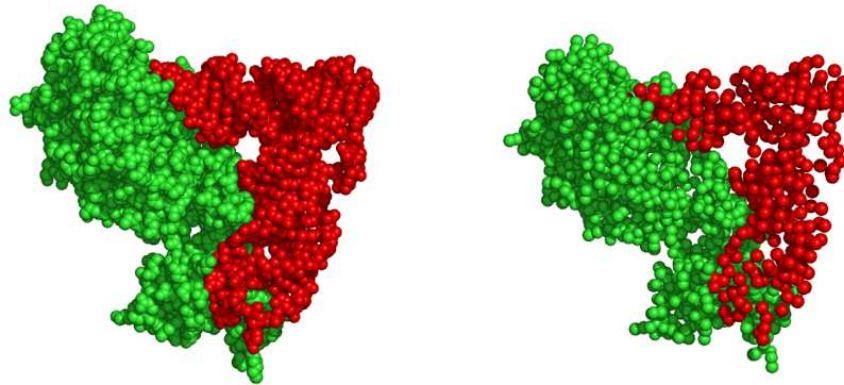


Figure 2. An all atom (left) and reduced model (right) of the Aspartyl tRNA synthetase complexed with tRNA (1asy).

now working on a set of known protein-RNA complexes to test the performance of the potential to predict the native complex structure starting from the individual components in systematic docking searches.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (DFG) for financial support through grant Za153-11 to M.Z.

References

1. K. Nadassy, S. J. Wodak and J. Janin J, *Structural features of protein-nucleic acid recognition sites*, *Biochemistry* **38**, 1999–2017, 1999.
2. S. Jones, D. Daley, N. Luscombe, H. Berman and J. Thornton J, *ProteinRNA interactions: a structural analysis*, *Nuc. Acids Res.* **29**, 943–954, 2001.
3. M. Treger and E. Westhof, *Statistical analysis of atomic contacts at RNA-protein interfaces*, *J. Mol. Recog.* **14**, 199–214, 2001.
4. J. J. Ellis, M. Broom and S. Jones, *ProteinRNA interactions: structural analysis and functional classes*, *Proteins* **66**, 903–911, 2007.
5. R. P. Bahadur, M. Zacharias M and J. Janin, *Dissecting protein-RNA recognition sites*, *Nuc. Acids Res.* **36**, 2705–2716, 2008.
6. H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, et. al, *The Protein Data Bank*, *Acta Crystallog. Sect. D* **58**, 899–907, 2002.
7. M. Zacharias, *Protein-protein docking with a reduced protein model accounting for side-chain flexibility*, *Protein Sci.* **12**, 1271–1282, 2003.
8. J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, S. J. Wodak, *CAPRI: a Critical Assessment of PRedicted Interactions*, *Proteins* **52**, 2–9, 2003.