Feature extraction for the clustering of small 3D structures: application to RNA fragments

Alix Delannoy, Antoine Moniot, Yann Guermeur, Isaure Chauvot de Beauchêne

Docking of RNA Hairpin on Protein Using a Fragment-Based Method

Antoine Moniot[1], Rohit Roy[1,2], Yann Guermeur[1] and Isaure Chauvot de Beauchene[1]
[1] LORIA, Campus Scientifique, BP 239, 54506, Vandœuvre-lès-Nancy Cedex, France
[2] Current: Duke Center for Genomic and Computational Biology, Duke University, NC 27708, Durham, USA

Corresponding author: antoine.moniot@loria.fr

Abstract *We introduce an extension of our fragment-based method for ssRNA-protein docking as it is still a challenging difficulty in docking. It is dedicated to hairpins and makes use of geometrical features of this secondary structure. An initial evaluation establishes that it is promising and could make it possible to overcome the limitations of the state-of-the-art fragment-based methods.*

Keywords RNA hairpin, protein, fragment-based docking

1 Introduction

Protein-RNA interactions are involved in many biological processes, including cell regulation [1] and diseases [2,3]. In that context, the structures of the complexes are major knowledge sources. However, their experimental inference is difficult, when possible [4]. As usual, the approach of choice to overcome this limitation is modeling. A difficulty arises when the interaction involves a single-stranded secondary structure of the RNA: single-stranded RNA is highly flexible and consequently difficult to model. This observation led to the introduction of two methods, one based on molecular dynamics [5] and ours, based on the assembling of structural fragments of RNA docked on the protein surface [6]. The current implementation of these methods, requiring the knowledge either of the exact coordinates of 2 nucleotides [5] or only of anchoring points on the protein surface [6], limits their applicability to few protein families. In this article, we introduce an extension of our method relaxing this requirement in the specific case when the single-stranded RNA is the loop of an hairpin. The additional pieces of information exploited to obtain this improvement are intervals on the distances between the nucleotides at the endpoints of the loop (intervals governed by the distances between the two nucleotides closing the loop). Thus, biological information is not needed except the identification of the nucleotides of the closure of the hairpin, which can be obtained by secondary structure prediction.

The organization of the paper is as follows. Section 2 introduces our method. Section 3 is devoted to its experimental evaluation. At last, we draw conclusions and outline our ongoing research in Section 4.

2 Methods

In this section, we first give a brief description of our fragment-based method, so as to highlight afterwards the specificities of the original contribution: the dedication to hairpins.

2.1 Fragment-Based Docking

Our fragment-based method consists of four main steps. To make the paper self-content, they are now briefly summarized (details are available in [7]). First, for each of the 64 possible trinucleotides, hereafter referred to as *motifs*, a library containing all experimentally observed 3D structures is built, by browsing the Protein Data Bank (PDB [8]). This initial set is then refined by means of a clustering method, to retain a subset, ideally of minimal cardinality, "covering" it with a Root Mean Square Deviation (RMSD) below 1Å(1Å-net). With these libraries at hand, the sequence of interest is first cut into trinucleotides with a step of one (so that two consecutive trinucleotides overlap by two nucleotides). For each of the corresponding motifs, the whole refined library of 3D structures is docked on the protein. This *rigid body docking*, using ATTRACT [9], generates for each trinucleotide a set of *poses*. Then, the assembly consists of searching possible paths in a directed graph. Its vertices are the poses and two successive poses are connected by an edge provided that the RMSD between

their two shared nucleotides is below a threshold. The output at this level is a list of *chains* of poses scored by ATTRACT, that cover the full RNA sequence. The last step consists in an ordering of the list. Two options are implemented. The first one is based on the geometric mean of the ranks of the poses [6]. The second one is a heuristic considering that the best chains are probably made of poses that participate in many chains, for probabilistic and/or entropic reasons [7]. It uses a forward-backward algorithm to count the number of chains in with each pose participates, then selects for every fragment the most connected poses and assemble only those.

In this previous framework, the anchorage takes the form of the knowledge of the location of the interaction between the nucleotides at both ends of the sequence and residues of the protein.

2.2 Dedication to Hairpins

One useful feature of hairpins regarding modeling is that the distance between the nucleotides at the endpoints of the loop is known [10]. Our dedicated method is based on the conjecture that an appropriate exploitation of this feature can prove enough to constrain the assembly so as to relax the initial need for anchoring points. The implementation rests on an enrichment of the graph described in Section 2.1. A new type of edge is introduced, to connect poses of the last and first fragment. This connection is added when the Euclidean distance between the phosphate of first nucleotide of the first fragment and the phosphate of the last nucleotide of the last fragment belongs to the interval 11.8-24.7Å(interval observed in the benchmark). The new graph is depicted in Figure 1.

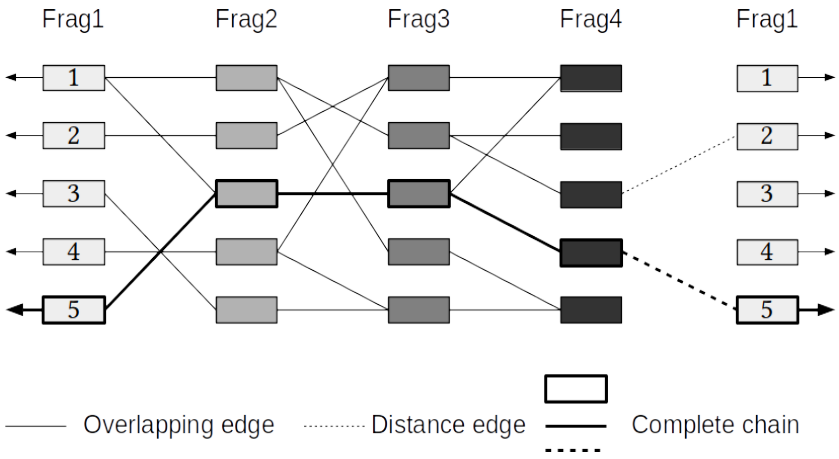

Fig. 1. Graph dedicated to hairpins.

Compared to the general case, the set of chains considered here is smaller, since it retains only those included in a cycle of the new graph.

3 Assessment of the New Method

To assess the method, a data set of hairpin-protein complexes was produced.

3.1 Selection of a Benchmark

The algorithm for this derivation is made up of two main steps. It takes in input the set of all available experimental structures of hairpin-protein complexes, with the redundancies removed. For every hairpin, the docking of all the conformers of all the motifs present in its loop is performed on the corresponding protein using ATTRACT. This leads to retaining only the complexes for which all fragments have at least one *near native* pose. A near native pose is a pose whose RMSD with the native (experimental) position is below 3Å. The second step is a refinement that consists in doing the docking again with a subset of conformers structurally close enough to the native position. This corresponds to eliminating conformers that are structurally too different from the native position to be a near native pose. At this level, the criterion to retain a complex is the following one: for every fragment, the rank of the ATTRACT score of the first-ranked near native pose must be below a threshold. Obviously,

this step, involving pieces of information on principle unknown, turns our experimental approach into a proof of concept.

The initial set of hairpin-protein complexes is obtained by application of NAfragDB [11]. It contains 19 complexes. At the first step of the algorithm, one obtains roughly $10 \cdot 10^3$ poses per conformer, i.e., $30 \cdot 10^6$ poses per fragment, given the fact that the libraries for the motifs contain on average 3000 conformers. At this level, only 2 complexes are selected: 5UDZ [12] and 1RKJ [13]. For both of them, at least one fragment has the top-ranked near native pose at a rank higher than 10^6. This explains, at least for our data set, the need for the second step of the algorithm, to keep a chance to obtain a relevant assembly. This second step is parameterized as follows. For every motif, a subset of conformers is created that contains only the ten closest (according to the RMSD) to the bound form after optimal fitting. On the contrary, the number of poses per conformer is increased to $50 \cdot 10^3$, so that the new number of poses per fragment is $500 \cdot 10^3$. Table 1 provides the set of hairpins selected at the first step, with the corresponding docking results.

		frag1	frag2	frag3	frag4	frag5
5UDZ	first docking	5743309	423163	15403	27578	2138595
	second docking	**90424**	2114	285	1334	34291
1RKJ	first docking	1382237	1110963	13187	599859	
	second docking	7056	**14949**	191	1275	

Tab. 1. First rank for a near native pose for the fragments of the two hairpins.

The figures in Table 1 establish that performing the assembly for 5UDZ and 1RKJ requires to consider a maximum of $6 \cdot 10^{24}$ and $5 \cdot 10^{16}$ possible chains respectively. Those numbers are based on the following computation. For each fragment of each sequence, the number of poses considered is the rounded largest value among the ranks of the first-ranked near native poses (here $90 \cdot 10^3$ for 5UDZ and $15 \cdot 10^3$ for 1RKJ). This arbitrary choice corresponds to a reasonable assumption on what information could be inferred from data. With the processing for 5UDZ being the most time consuming, and currently underway, in the sequel, the results are provided for 1RKJ only.

3.2 Assessment for the Hairpin of 1RKJ

The sequence of the hairpin is UCCCGA (thus four fragments). The three parameters to be set to derive the chains are the interval for the distance between the two nucleotides at the endpoints of the loop, the number of poses considered, and the threshold on the RMSD between the common nucleotides of overlapping fragments (see Section 2.1). The two first values have been given above. As for the third one, the value of 2.6Å was retained since it is the smallest value ensuring to generate a chain connecting near native poses only. For this parameterization, the number of chains is 47617288. In that set we obtain 732 acceptable solutions (with an RMSD below 5Å) and 122 good solutions (RMSD below 3Å), with a best model at 1.9Å. The best one and the experimental one are represented in Figure 2.

Thus, the method appears sensitive, but not specific enough, which calls for an investigation of the set of chains. Section 2.1 has introduced the two methods implemented to sort it. We now discuss their effectiveness. When using as criterion the geometric mean of the ranks of the poses, the smallest rank of a solution among the 732 satisfactory ones is 4783648. This is close to the top 10%, but leaves too many false positives above.

	frag1	frag2	frag3	frag4
All poses	206742	502600	654520	1105597
Poses in a solution	5701	16849	366	1672
Average for all poses	11985	15830	24865	20151

Tab. 2. Highest number of chains in which a pose of the 47617288 chains (all poses) and a pose of the 732 solutions (poses in a solution) are involved. The last line is the average of chains for all poses.

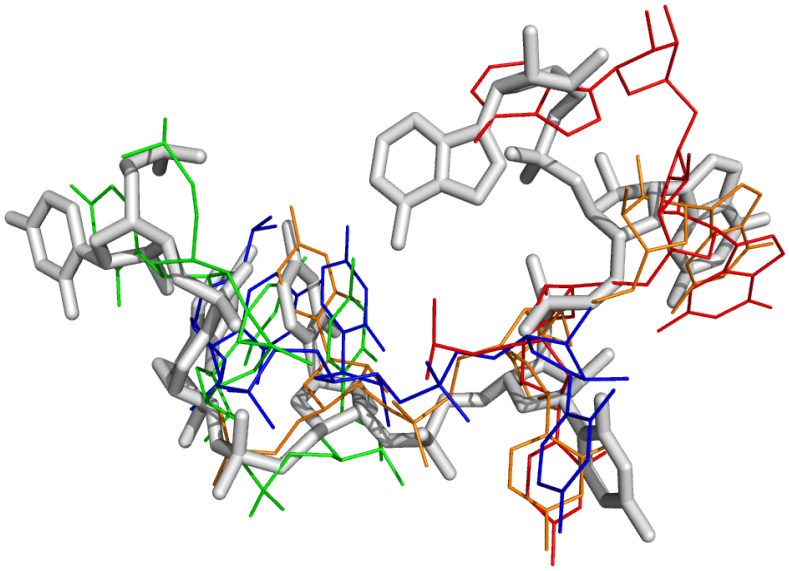

Fig. 2. Poses composing the best chain considering RMSD with the native chain (in white). Frag1 in green, frag2 in blue, frag3 in orange, frag4 in red.

The second criterion is evaluated by regarding the number of chains in which poses are involved (see Table 2). We found that the poses which are involved in good chains are not involved in more chains than the rest of the poses.

In order to decrease the number of retained chains, we tested to lower to 10-15Å the interval for the distance between the phosphate of the first nucleotide of the first fragment and the phosphate of the last nucleotide of the last fragment (which was initially set to 11.8-24.7Å), since this distance is 13.2Å in 1RJK. This resulted in a decrease of the cardinality of more than 30%, obtaining 32765494 chains, but also decreasing the number of good solutions from 732 to 517. It did not increase the percentage of correct solutions in the chains.

4 Conclusions and Ongoing Work

Given the difficulty of the task tackled, our initial results appear promising. The knowledge of the secondary structure seems to be relevant enough to replace the knowledge of the anchoring points. The results were better with anchoring points, with a higher percentage of correct chains and a more accurate best chain in most cases [6]. This was to be expected since distance and position represent a stronger constraint than distance only. Indeed, this knowledge allowed to assemble and evaluate all the possible chains, without our previous heuristic pre-filtering of the most-connected poses [7], and to obtain a more precise model (1.9Å, instead of 3.6-5.7Å). On the other hand, the hypothesis of a loop closure being weaker than that of the exact position of the chain extremities on the protein, the current approach retains more false positives than the anchored docking did (more than 1% correct models in the assembled chains).

An improvement of our method should result from a change of target for the distance. The distance between the phosphate of the first nucleotide and the sugar of the last nucleotide seems to be a stronger constraint. The interval of distance is 13.7-21.6Å, which is tighter than the phosphate-phosphate interval (see Section 2.2). The variance for the phosphate-sugar distance, 2.4, is smaller than for the phosphate-phosphate distance (5.9). These values are observed on our bigger benchmark of 191 structures. Currently, the limiting factor of the new method is still the docking of the trinucleotides by ATTRACT, as in most complexes, not all fragments have at least one near-native pose. A major reason for this problem is inherent to the fragment-based approach: minimizing the interaction energy for such small fragments is not equivalent to minimizing this energy for the whole sequence. Another one, but directly related, is the inadequation of ATTRACT scoring function for ssRNA, developed on protein- double-stranded RNA complexes. Thus, our current research consists in developing a new

scoring function specific for ssRNA fragments. In parallel, we consider an addition to the distance constraint, to check if the loop-closing nucleotides have a geometry compatible with the base-pairing of the neighbor nucleotides. This will consist in evaluating if at least one conformer of the library can be fitted on the two terminal nucleotides on one side so as to etablish a base-pairing with at least one conformer fitted on the other side of the loop.

Finally, a natural extension of this work consists in applying its principle to different RNA secondary structures (known or predicted), with the aim of the global docking of the complex. Eventually, the constraint should not necessarily involve the knowledge of the secondary structure, but could benefit from any knowledge of distance between two nucleotides.

References

[1] Y. Huang, JL. Zhang, XL. Yu, TS. Xu, ZB. Wang, and XC. Cheng. Molecular functions of small regulatory non-coding RNA. *Biochemistry Moscow*, 78(3):221–230, 2013.

[2] T. Vanderweyde, K. Youmans, L. Liu-Yesucevitz, and B. Wolozin. Role of stress granules and RNA-bindingproteins in neurodegeneration: a mini-review. *Gerontology*, 59(6):524–533, 2013.

[3] M. Derrigo, A. Cestelli, G. Savettieri, and I. Di Liegro. RNA-protein interactions in the control of stability andlocalization of messenger RNA (review). *International Journal of Molecular Medicine*, 5(2):111–134, 2000.

[4] S. Jones. Protein–RNA interactions: structural biology and computational modeling techniques. *Biophysical Reviews*, 8(4):359–367, 2016.

[5] K. Kappel and R. Das. Sampling native-like structures of RNA-protein complexes through rosetta folding and docking. *Structure*, 27(1):140–151, 2019.

[6] Chauvot de Beauchene I, S. de Vries, and M. Zacharias. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Research*, 44(10):4565–4580, 2016.

[7] Chauvot de Beauchene I, S. de Vries, and M. Zacharias. Binding site identification and flexible docking of single stranded RNA to proteins using a fragment-based approach. *PLoS Computational Biology*, 12(1), 2016.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 18(1):235–242, 2000.

[9] M. Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science*, 12(6):1271–1282, 2003.

[10] P. Clote, Y. Ponty, and JM. Steyaert. Expected distance between terminal nucleotides of RNA secondary structures. *Journal of Mathematical Biology*, 65(3):581–599, 2011.

[11] A. Moniot, S. de Vries, D. Ritchie, and I. Chauvot de Beauchene. NAfragDB: a multi-purpose structural database of nucleic-acid–protein complexes for advanced users. In *GGMM*, page 21, 2019.

[12] L. Wang, Y. Nam, A.K. Lee, C. Yu, K. Roth, C. Chen, E.M. Ransey, and P. Sliz. LIN28 zinc knuckle domain is required and sufficient to induce let-7 oligouridylation. *Cell Reports*, 18(11):2664–2675, 2017.

[13] C. Johansson, L.D. Finger, L. Trantirek, T.D. Mueller, S. Kim, I.A. Laird-Offringa, and J. Feigon. Solution structure of the complex formed by the two N-terminal RNA-binding domains of nucleolin and a pre-rRNA target. *Journal of Molecular Biology*, 337(4):799–816, 2004.