Modelling ssRNA-protein complexes at atomic resolution

Isaure Chauvot de Beauchêne, Sjoerd Jacob de Vries, Martin Zacharias

Modelling ssRNA-protein complexes at atomic resolution

Isaure CHAUVOT DE BEAUCHENE [1], Sjoerd DE VRIES [1], Martin ZACHARIAS

[1] Physics Dep., Technical University Munich, James-Franck Str. 1, 85748 Garching, Germany

Corresponding Author: zacharias@tum.de

Abstract. *RNA-protein specific binding underlies a large variety of fundamental cellular processes. An atomistic description of such binding processes would aid the rational conception of pharmaceutical modulators of those functions. However, while the field of protein-protein docking has achieved considerable improvements in the last decade, protein-RNA docking encounters specific difficulties. This is mainly due to the flexibility and the large conformational space of RNAs compared to proteins, and especially single-stranded RNAs (ssRNAs). Here, we present a novel and highly effective fragment-based approach to tackle this problem, capable of accurate prediction of the structure of a ssRNA bound to a protein, starting from the structure of the protein and the sequence of the RNA. As a proof-of-principle, we focus on the common case of a uniform ssRNA sequence. Without any information on specific contacts or the RNA structure, our method permitted to define accurately the binding site on the protein with 10 Å precision, through the use of a comprehensive fragment library. Moreover, the bound conformation of the ssRNA could be sampled with ~1.5 Å RMSD on heavy atoms, a precision never reached so far. In future research, the method will be extended to dock arbitrary ssRNA sequences to protein structures.*

Keywords: RNA-protein docking, fragment-based docking, *ab initio* modeling, single-stranded RNA, RNA-protein recognition.

Modélisation de complexes ARNsb-protéine à résolution atomique

Résumé : *Les liaisons spécifiques ARN-protéine sont impliquées dans une grande variété de processus cellulaires fondamentaux. Une description atomistique de ces processus de liaison aiderait la conception rationnelle de modulateurs pharmaceutiques de ces fonctions. Toutefois, si le domaine de l'amarrage protéine-protéine a connu des avencées considérables dans la dernière décennie, l'amarrage ARN-protéine se heurte à les difficultés spécifiques. La principale est la flexibilité et le large espace conformationnel des ARNs par rapport aux protéines, surtout pour les regions simple brin des ARN (ARNsb). Nous présentons une approche originale et efficace pour prédire avec précision la structure d'un ARN simple brin lié à une protéine, à partir de la structure de la protéine et la séquence de l'ARN. Nous nous concentrons ici sur le cas courant d'une séquence uniforme d'ARN. Sans aucune information sur des contacts spécifiques, ni sur la structure de l'ARN, notre méthode a permis de définir le site de liaison sur la protéine a 10 Å près. En outre, la conformation liée de l'ARNsb a été approximée a ~ 1,5 Å près, une précision jamais approchée jusqu'à présent. Dans la suite de nos recherches, la méthode sera étendue à l'amarrage de séquences arbitraires d'ARNsb à des structures de protéines.*

Mots-clés : Amarrage ARN-protéine, amarrage par fragment, modélisation ab initio, ARN simple brin, reconaissance ARN-protéine

1. Introduction

RNA-protein specific binding supports a large variety of fundamental cellular processes, from initiation/repression of gene transcription to inter-cellular communication [1 2 3]. However, the number of atomic structures of protein-RNA complexes remains quite low, therefore structural prediction methods are needed. Docking methods aim to assemble a complex from atomic structures of the free, unbound structures of the components (protein or RNA). The first task in docking is to sufficiently sample the space of possible conformations and relative orientations (i.e. *poses*) of the unbound components so as to include near-native structures. In this regard, RNA-protein docking encounters specific limits compared to protein-protein docking, due to the very high flexibility and conformational variety of RNA.

The conformational changes occurring upon RNA association with protein can involve global rearrangements, changes of secondary structure elements and/or flipping-out of bases from intra- to extra-helical position. Therefore, unbound RNA structures are a reasonable starting point only in a limited amount of cases. They are especially unreliable when some single-stranded loops participate in the binding. Due to the high flexibility of *single-stranded RNA (ssRNA)* regions, an experimental structure of the unbound form is very unlikely to exist, and even so would not provide enough data to infer the bound form. Even RNA molecules that are otherwise well-structured contain such single-stranded regions that carry the specificity in most specific RNA-protein recognition processes. Still, essentially all studies in the field have concentrated so far in docking unbound structures of RNAs with limited conformational sampling. They classically sample possible RNA conformations based on the unbound structure by use of coarse-grained models [4], normal modes analysis, elastic network models [5 6] or local conformational perturbations [P. Setny and M. Zacharias, *in prep.*] , then perform rigid [7 8] or nearly-rigid docking [9]. Such methods can perform well when only moderate conformational changes occur. The lack of methodology for modeling larger changes, in particular bound single-stranded loops, limits the accuracy of all current protein-RNA docking methods [5].

Here we present a novel strategy for *ab initio* modeling of ssRNA bound to a protein that does not require any structural information on the ssRNA, assembling it from sequence alone. Our approach consists of cutting the RNA into small overlapping fragments, docking them separately, and assembling the compatible docking poses into a new realistic conformation (Fig. 1). Our strategy uses a large, home-made RNA fragment library to sample the conformational diversity of RNA, and relies on a coarse-grained force field to select effectively the most probable poses.

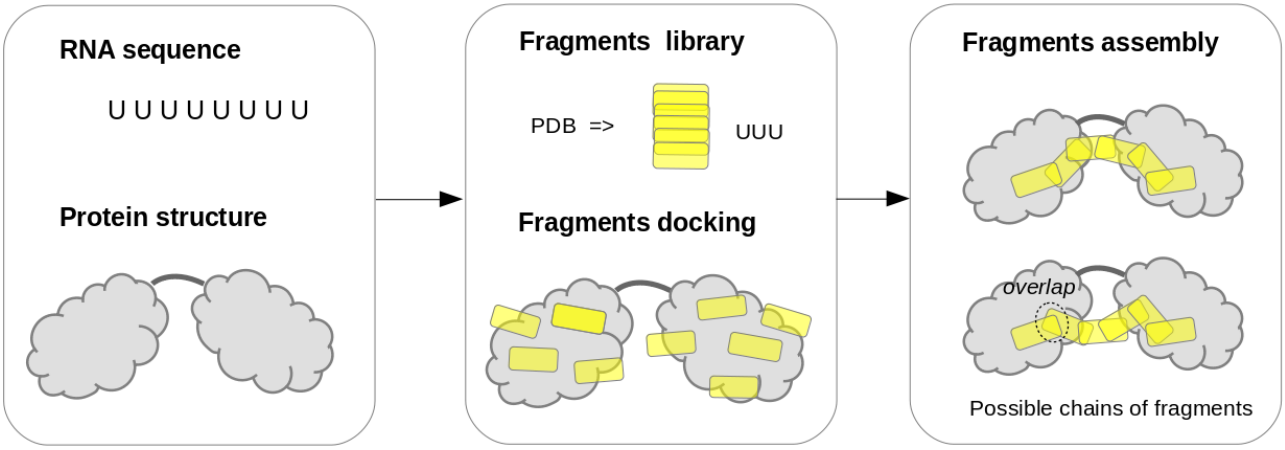

Figure 1. Strategy for homology-driven ssRNA-protein docking with structural fragments.

With this strategy, a structure of poly-U ssRNA in complex with a protein was generated in close agreement with the published crystal structure (< 5 Å RMSD). To the best of our knowledge, this is the first time that a ssRNA has been successfully predicted at such high precision without use of any prior knowledge on the RNA structure.

2. Results

a. Bound docking on a test-case

In order to get a first idea of the validity of our fragment-based approach and an upper-limit of its performance, we performed docking tests on one case using not the fragment library but the bound form of the RNA. We chose the PDB structure 1B7F representing the sex-lethal protein bound to a 5'-UUUUUUUU single-stranded RNA, and which constitute a canonical case of a RRM-containing RNA-protein complex [10 11]. The RNA is bound to a deep cleft delimited by RNA-binding domains. Most nucleotides bind by their base and/or sugar, and establish 1 to 5 hydrogen-bonds with the protein (Fig. 2).

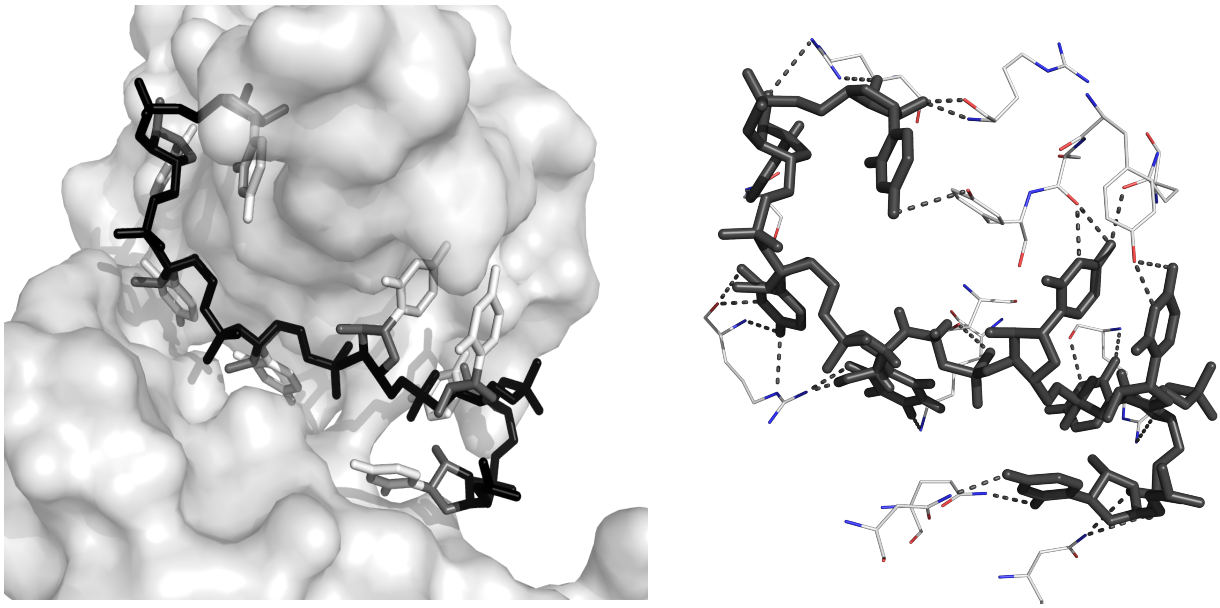

Figure 2. Crystallographic structure of the human sex-lethal protein (surface) bound to a 5'-U8 ssRNA (sticks) (PDB ID 1B7F). Left: The RNA backbone and bases are distinguished in black and white respectively. Right: The nucleotides and amino-acids establishing hydrogen bonds (dashes) are represented as black sticks and gray lines respectively, with the same orientation as on the left picture.

We cut the RNA into trinucleotides that overlap by two nucleotides. Each trinucleotide fragment was docked onto the bound structure of the protein, both partners in *coarse grained (CG)* representation, using our docking program ATTRACT [9 10] , and the poses were assembled using distance restraints between overlapping nucleotides. . This test resulted in a *quasi-native* RNA chain (1.2 Å RMSD toward the bound form) (Table 1). This result was comparable to rigid bound-docking with the entire RNA in its bound form (0.7 Å RMSD , result not shown).

	Bound docking	**Biased docking**	**Unbiased docking**	
Fragment	Best RMSD			Nb poses < 5 Å
n5-n7	1.7 Å	1.8 Å	3.3 Å	5
n6-n8	1.7 Å	1.8 Å	4.4 Å	7
n7-n9	0.5 Å	1.1 Å	4.4 Å	72
n8-n10	0.9 Å	1.4 Å	3.5 Å	255
n9-n11	0.7 Å	0.9 Å	3.5 Å	165
n10-n12	1.4 Å	2.7 Å	6.1 Å	0
Average / total	1.2 Å	1.6 Å	4.3 Å	496 (6%)

Table 1. Comparison to the bound form of the poses obtained by bound, biased or unbiased docking.

b. Biased docking using a fragment library

We built a fragment library sampling the conformational space for each of the 64 possible trinucleotide sequences, based on the available ssRNA-protein structures in the Protein Data Bank (*to be published*). The test-case 1B7F was not included in the library building process. In order to get a first evaluation of the capacity of our library to sample efficiently near-native solutions, with a reduced computational cost, we performed biased docking of a poly-U octo-nucleotide (noted n5-n12) on the sex-lethal protein, corresponding to complex 1B7F in our benchmark. After excluding from our library the fragments issued of this complex, we selected for each bound fragment the best fitting conformer in the UUU sub-library, ending up with six UUU conformers.

Quasi-native solutions (RMSD < 2.1 Å) were sampled for each fragment, among ~15,000 non-redundant

poses per conformer . By selecting the 20 % best-scored solutions, we retained near-native solutions (RMSD < 3 Å) for each fragment (Table 1). The worst-docked fragment corresponded to nucleotides n10-n12, with the best solution in top 20% at 2.7 Å RMSD. The structure of this fragment corresponds to a conformation not closely sampled in the library. The best conformer for this fragment displays 1.8 Å RMSD when fitted to the bound form, *versus* 0.4 Å to 1.1 Å for the other fragments (results not shown). These results indicate that ATTRACT was able to sample and rank solutions very close to the optimal position of each conformer in the top 20 %.

We further tried to assembly the overlapping fragments into chains, by evaluation of the inter-atomic distances of the overlapping nucleotides (*see Methods*). We retained the top 20 % poses for each of the six conformers, for a total of 18,222 poses. Out of the 18222 x 18222 possible pairings, 23,458 were found to possibly represent consecutive overlapping fragments. These pairs were arranged into 5-fragment chains by identifying pairs A-B, B-C, C-D, and D-E. The 75,311 resulting chains n5-n11 were filtered by total overlap energy, leaving 3,174 chains. The best chain had an average RMSD of 1.4 Å (Fig. 3). More importantly, this RMSD was representative for the whole result : 41 % of the chains were under 2.0 Å RMSD, and 97 % under 5.0 Å. In conclusion, the correct structure was essentially the only possible way to build a poly-U hepta-nucleotide onto the protein, when docking the proper conformers.

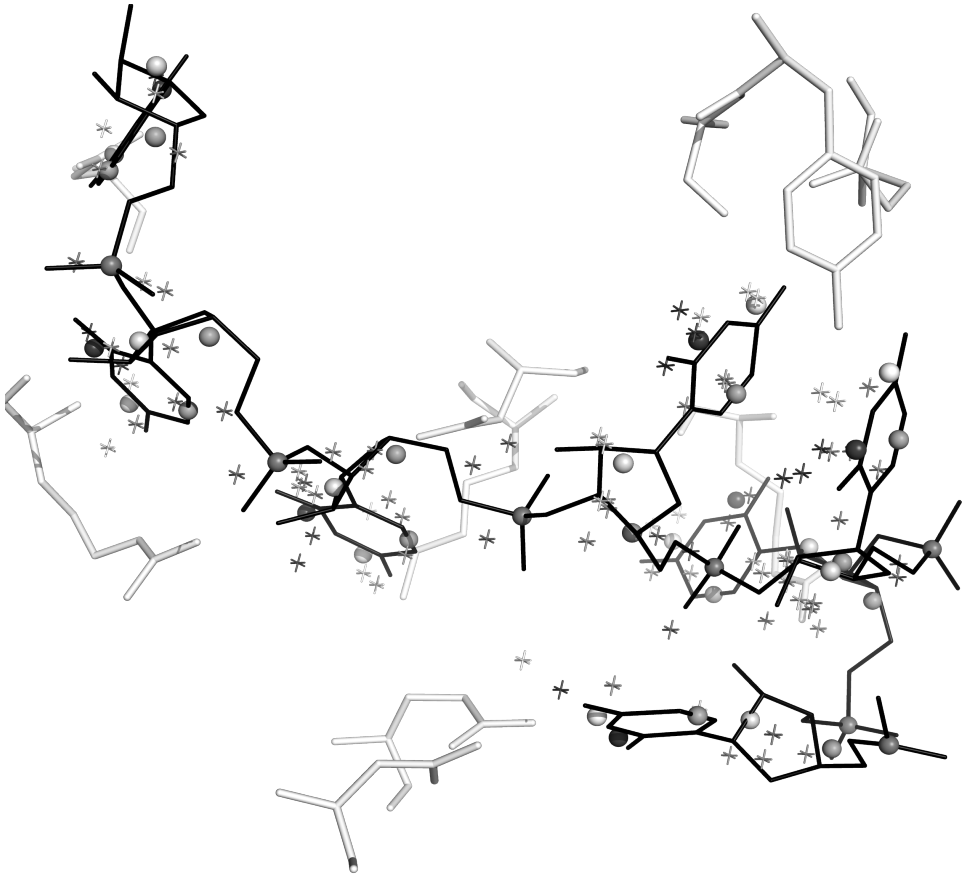

Figure 3. Best prediction obtained by unbound fragments docking. The amino-acids interacting with the RNA in the crystal structure are represented in white sticks. The bound RNA is represented both in black lines and in spheres corresponding to beads of ATTRACT coarse-grain representation, with one color per bead type. Each bound nucleotide n is approximated by up to 3 docked fragments {n, n+1, n+2}, {n-1, n, n+1} and {n-2, n-1, n}. The beads of the docked fragment are represented as asterisks, with same color code as for the bound nucleotides.

d. Unbiased docking with the complete fragment library

A similar procedure was applied considering not only the best conformers but the whole UUU library (1305 conformers) for each fragment in n5-n12. The docking produced more than 22 millions of non-

redundant poses, from which we retained the top 20%. Out of this large pool, the fragments with the highest propensity to form complete hepta-nucleotide chains were selected. This resulted in 8,441 chain-forming fragments, corresponding to 7,798 poses (one pose can correspond to different fragments, by shifting its position in the chain). Some correct poses were found toward all bound fragment except n10-12, similarly to what was found by the biased six-conformer docking. In total, 6 % of the poses were correct (RMSD < 5 Å from at least one bound fragment) (Table 1).

In addition, these poses permitted to define accurately the binding site (Fig. 4). The worse pose was at only 15.4 Å from the closest fragment, and more than 50 % of the poses were under 10 Å. Moreover, the poses outside the binding site scored very poorly : the 8 poses at more than 15 Å from any fragment were ranked in the last 10 %. Therefore, our procedure for fragment assembly proved an efficient filter of wrong solutions.

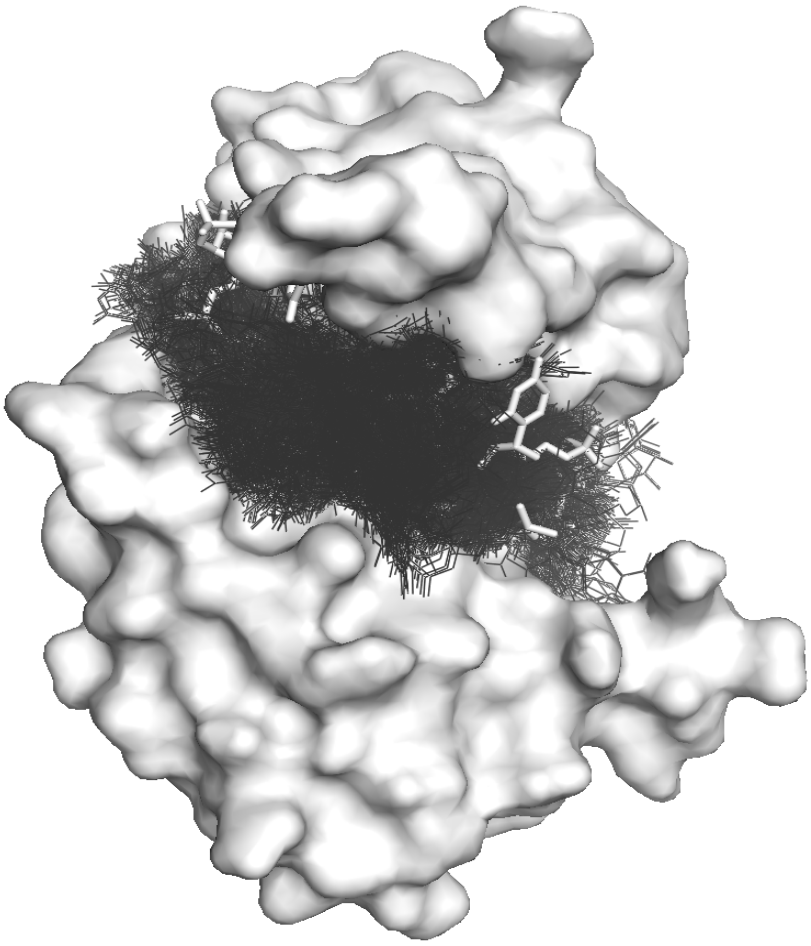

Figure 4. Delineation of the RNA-binding site of sex-lethal protein by unbiased docking. The 8441 chain-forming poses are represented as dark lines, the bound RNA as white sticks and the protein as surface. None of the poses are occluded by the protein.

3. Discussion and Perspectives

Here we present a method to predict the structure of ssRNA-protein complexes, based on the structure of the protein and the sequence of the RNA, using a fragment-based approach. For the first time, \ the structure of such a complex could be predicted at high precision.

First, we validated our fragment-based approach by preliminary tests of bound-bound ssRNA-protein docking. This led to a precision of ~1A RMSD towards to the crystal structure, comparable to the control run where the bound ssRNA was docked in its entirety. Therefore, the process of cutting the RNA into fragments does not lead to any loss in accuracy or precision.

Subsequently, we performed "*ab initio*" RNA–protein docking, using the structure of the protein and the sequence of the RNA. The structures used for docking were obtained by extracting an exhaustive library of trinucleotide fragments out of the 571 available structures of ssRNA-protein complexes (PDB, July 2014). By a biased fragment docking, selecting the library conformers closest to the RNA structure, we generated accurate near-native structures at ~ 1.5 Å RMSD for an hexa-nucleotide, providing a proof-of-principle for *ab initio* fragment-based docking of ssRNA on protein. Using unbiased docking with the full library, we were able to select a pool of fragments that delineate accurately the binding site.

Our results constitute a considerable improvement compared to the limited success that had been achieved so far in docking ssRNA-protein complexes. Almost all current methods are limited to structured RNA. Only the RNA-lim method [14] has attempted to predict ssRNA-protein structures based on fragments, focusing the sampling around a pre-defined binding site. They predicted the position (but not the orientation) of RNA fragments at around 5 Å from the binding site. In contrast, our method achieves this precision for both position and orientation for unbiased docking , without the use of any knowledge on RNA conformation or binding site. For biased docking, our method predicts the RNA structure at high precision. We emphasize that the only difference between biased and unbiased docking is that in biased docking, the fragment library was limited to relevant conformations. All structures generated in biased docking were also generated in unbiased docking, albeit among millions of others.

In this study, we have focused on sampling, *i.e.* the generation of correct ssRNA-protein conformations. Although the fragment approach proved an efficient sampling strategy, the ranking of the best structures among decoys (scoring) remains an issue. As our test sequence is only made of UUU fragments, the 8,441 chain-forming fragments selected by our unbiased docking could theoretically form 4×10^{19} chains for n5-n10. To reduce the number of chains to select for further refinement and scoring, we plan to use insights of specific RNA-protein contacts from conserved RNA-binding motifs in proteins.

Our method currently focuses on the common case of a binding ssRNA that is uniform in sequence. For this case, the current study provides in important proof-of-principle for RNA-protein modeling. The method is in principle extendable to arbitrary sequence, and this will be the direction of further research. If this method proves generalizable to a larger number of cases and sequences, it will constitute a major methodological breakthrough.

4. Methods

a. Docking of the RNA fragments.

Both the protein and the RNA fragment were in coarse grain representation. The pyrimidine/purine were represented by 6/7 beads and the amino-acids by 5 to 8 beads [4 15] . Throughout this paper, the results are given in coarse-grain RMSD if not otherwise mentioned. A comparison tests on 21,704 RNA fragments showed a 10 % decrease in RMSD numbers when converting from CG to all-atoms for RMSD < 5 Å , with a regression coefficient of 0.90.

For each docked conformer, the starting positions of the ligand and receptors were produced by the "randsearch" procedure of ATTRACT [13], generating starting positions and respective orientation of the two partners (protein and fragment). The positions of the *center of mass (COM)* of the fragment at each starting configuration are equidistant on a unit sphere of 75 Å radius centered on the protein. A gravity was applied, consisting in a harmonic distance restraint toward the COM of the protein with harmonic constant of 0.0015 kcal/Mol/Å. For each fragment were performed 1000 minimization steps, the pairwise interactions between ligand and receptor being approximated on a pre-calculated receptor grid. A final re-scoring was performed without grid, pairwise interactions being considered until a squared distance of 50 Å. The final poses were sorted by score, and the redundant poses (within 0.05 Å from a better scored pose) were discarded. We first docked each of the 6 best-fitting UUU conformers starting from 30 000 initial positions, producing ~ 15,000 non-redundant poses. Then each conformers of the whole UUU sub-library was docked starting from 200,000 position to account for the increased difficulty.

b. Overlap evaluation.

The overlap of two fragments was evaluated using ATTRACT scoring function with harmonic distance

restraints and no force-field. The restraints were defined between the 2nd and 3rd nucleotides of the 1rst fragment and the 1rst and 2nd nucleotides of the 2nd fragment. The restraints were defined such that the two backbone beads and the most remote base bead must occupy the same position, with some margin. The margin was defined with smaller values for the backbone than for the base (2.3 Å and 2.8 Å respectively), to account for the necessity to further link the backbone atoms in the refinement procedure.

c. Sequential clustering.

For the unbiased docking, the 20% best-scored poses for each conformer were selected and grouped, then clustered by 2 Å. The centers of the 2Å-clusters were clustered into 3Å-clusters, and the centers of the 3Å-clusters into 4Å-clusters. The overlap between the center of mass of the central structure of the 4Å-clusters was evaluated, and all pairs of clusters with low overlap-energy were kept. The same procedure was applied inside each pair of overlapping 4Å-clusters at the 3Å-clusters level, with a lower overlap margin. Each 3Å-cluster belonging to the first 4Å-cluster was paired with each 3Å-cluster within the second 4Å-cluster. The same procedure was followed with each pair of overlapping 3Å-clusters at the 2Å-clusters level, then at the level of individual fragments, with decreasing overlap margins. According to optimization tests, we used distance restraints of 2.23 Å for backbone and 2.83 Å for side chain, with decreasing margins for overlap-energy for the different clustering levels.

References

[1] S. Geisler and J. Coller, “RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts,” *Nat. Rev. Mol. Cell Biol.*, vol. 14, no. 11, pp. 699–712, Nov. 2013.

[2] Y. Huang, J. L. Zhang, X. L. Yu, T. S. Xu, Z. B. Wang, and X. C. Cheng, “Molecular functions of small regulatory noncoding RNA,” *Biochem. Mosc.*, vol. 78, no. 3, pp. 221–230, Mar. 2013.

[3] C. Maris, C. Dominguez, and F. H.-T. Allain, “The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression,” *FEBS J.*, vol. 272, no. 9, pp. 2118–2131, May 2005.

[4] P. Setny and M. Zacharias, “A coarse-grained force field for Protein–RNA docking,” *Nucleic Acids Res.*, vol. 39, no. 21, pp. 9118–9129, Nov. 2011.

[5] S. Fulle and H. Gohlke, “Molecular recognition of RNA: challenges for modelling interactions and plasticity,” *J. Mol. Recognit. JMR*, vol. 23, no. 2, pp. 220–231, Apr. 2010.

[6] P. Setny and M. Zacharias, “Elastic Network Models of Nucleic Acids Flexibility,” *J. Chem. Theory Comput.*, vol. 9, no. 12, pp. 5460–5470, Dec. 2013.

[7] L. Pérez-Cano, A. Solernou, C. Pons, and J. Fernández-Recio, “Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials,” *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, pp. 293–301, 2010.

[8] Y. Huang, S. Liu, D. Guo, L. Li, and Y. Xiao, “A novel protocol for three-dimensional structure prediction of RNA-protein complexes,” *Sci. Rep.*, vol. 3, May 2013.

[9] E. Mashiach, D. Schneidman-Duhovny, A. Peri, Y. Shavit, R. Nussinov, and H. J. Wolfson, “An integrated suite of fast docking algorithms,” *Proteins*, vol. 78, no. 15, pp. 3197–3204, Nov. 2010.

[10] N. Handa, O. Nureki, K. Kurimoto, I. Kim, H. Sakamoto, Y. Shimura, Y. Muto, and S. Yokoyama, “Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein,” *Nature*, vol. 398, no. 6728, pp. 579–585, Apr. 1999.

[11] E. Sakashita and H. Sakamoto, “Protein-RNA and Protein-Protein Interactions of the Drosophila Sex-Lethal Mediated by Its RNA-Binding Domains,” *J. Biochem. (Tokyo)*, vol. 120, no. 5, pp. 1028–1033, Nov. 1996.

[12] A. May and M. Zacharias, “Protein–protein docking in CAPRI using ATTRACT to account for global and local flexibility,” *Proteins Struct. Funct. Bioinforma.*, vol. 69, no. 4, pp. 774–780, Dec. 2007.

[13] S. J. de Vries, C. E. M. Schindler, I. Chauvot de Beauchêne, and M. Zacharias, “A Web Interface for Easy Flexible Protein-Protein Docking with ATTRACT,” *Biophys. J.*, vol. 108, no. 3, pp. 462–465, Feb. 2015.

[14] D. Hall, S. Li, K. Yamashita, R. Azuma, J. A. Carver, and D. M. Standley, “RNA-LIM: a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure,” *Anal. Biochem.*, vol. 472, pp. 52–61, Mar. 2015.

[15] M. Zacharias, “Protein-protein docking with a reduced protein model accounting for side-chain flexibility,” *Protein Sci. Publ. Protein Soc.*, vol. 12, no. 6, pp. 1271–1282, Jun. 2003.