

STRUCTURAL AND ENERGY DETERMINANTS IN PROTEIN-RNA DOCKING

Laura Pérez-Cano,^{a,b#} Miguel Romero-Durana^{a#} and Juan Fernández-
Recio^{a*}

^aJoint BSC-CRG-IRB research program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center (BSC), Jordi Girona 29, Barcelona, 08034, Spain

^bCenter for Neurobehavioral Genetics and Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, California, USA.

*correspondence to: Juan Fernández-Recio, Life Sciences Department, Barcelona Supercomputing Center (BSC), Jordi Girona 29, 08034 Barcelona, Spain. E-mail: juanf@bsc.es

Joint first authors

Abstract

Deciphering the structural and energetic determinants of protein-RNA interactions harbors the potential to understand key cell processes at molecular level, such as gene expression and regulation. With this purpose, computational methods like docking aim to complement current biophysical and structural biology efforts. However, the few reported docking algorithms for protein-RNA interactions show limited predictive success rates, mainly due to incomplete sampling of the conformational space of both the protein and the RNA molecules, as well as to the difficulties of the scoring function in identifying the correct docking models. Here, we have tested the predictive value of a variety of knowledge-based and energetic scoring functions on a recently published protein-RNA docking benchmark and developed a scoring function able to efficiently discriminate docking decoys. We first performed docking calculations with the bound conformation, which allowed us to analyze the problem in optimal conditions. We found that geometry-based terms and electrostatics were the most important scoring terms, while binding propensities and desolvation were much less relevant for the scoring of protein-RNA models. This is in contrast with what we observed for protein-protein docking. The results also showed an interesting dependence of the predictive rates on the flexibility of the protein molecule, which arises from the observed higher positive charge of flexible interfaces and provides hints for future development of more efficient protein-RNA docking methods.

Keywords

Protein-RNA interactions; structural modelling; computational docking; scoring functions; residue-ribonucleotide potentials

1 Introduction

Protein-RNA interactions are involved in all major processes that occur during gene expression. Therefore, the structural determination of the specific complexes formed between proteins and RNA molecules is essential to gain a better understanding of the basis of molecular biology and cellular function. Advances in the structural characterization of protein-RNA interactions have accelerated in the past decade, with major milestones accomplished, such as the structural determination of the ribosome and part of the spliceosome [1–7]. In spite of the advances, there is a clear underrepresentation of protein-RNA structures in the Protein Data Bank [8] given that the experimental structural determination of these complexes at high resolution is still challenging and much more difficult than that of the isolated components. In this line, computational approaches for the structural modelling of protein-RNA complexes are strongly required in order to complement existing experimental data on protein-RNA interactions of interest. One promising tool is computational docking, which has been extensively applied to the structural modelling of protein-protein interactions of biological relevance [9,10]. The inclusion of a protein-RNA target in the Critical Assessment of Prediction of Interactions (CAPRI) [11] showed for the first time, in a systematic way, the potential application of protein-protein docking methods for predicting the structure of a protein-RNA complex [12]. However, most computational efforts have focused on the characterization of protein-RNA interfaces [13]. Given the scarcity of available protein-RNA structures, the initial protein-RNA docking approaches were only evaluated in small data sets, which compromised the reliability of their predictions [14–18]. In this regard, our laboratory

published one of the first two protein-RNA docking benchmarks [19], which provided the opportunity for a more exhaustive evaluation and optimization of existing protein-RNA docking protocols, and constituted a good framework to develop new ones [20–22]. Nevertheless, the number of reported studies on the development of protein-RNA docking and scoring algorithms is still very limited, and they have not always been tested in realistic conditions, that is, on a sufficient large set of cases using only the structures (or models) of the unbound conformations. Therefore, further systematic analyses are required. In this context, we have used here the above mentioned structural data set in order to explore the limits of rigid-body sampling as well as the effectiveness of different energy terms and functions for scoring protein-RNA docking models. For the sake of comparison, we have also evaluated the same parameters on the protein-protein docking benchmark 3.0 [23], which allowed us to identify features that are specifically good for the scoring of each type of interaction. The combination of the best individual parameters provides an efficient scoring function for protein-RNA docking. The detailed analysis performed in this work also gives new insights on the recognition mechanism of protein-RNA complexes, which is strongly dependent on protein flexibility.

2 Material and methods

2.1 Protein-RNA docking protocol

We used the FFT-based docking program FTDock 2.0 [24] for rigid body sampling, with no electrostatics, at a grid resolution of 1.2 Å. During sampling, the protein was always considered as the static receptor, and the RNA molecule as the mobile ligand. The sampling

phase generated 10,000 protein-RNA docking poses for each benchmark case.

2.2 Statistical potentials for scoring

We used the previously calculated pairwise protein-RNA statistical potentials [15] to score the generated protein-RNA rigid-body docking orientations. For every residue-ribonucleotide pq pair at the interface of each docking pose i (considering as pairs those that have at least one atom within 4 Å distance from each other), the corresponding propensity value according to its type was assigned. The propensity-based values of all pairs were summed up to compute the final score of the given docking pose i , as shown in Eq. 1. Finally, all docking solutions were ranked according to these propensity-based scores.

$$\Delta G_i^{stat} = \sum_{pq} \Delta G_{pq}^{stat} \quad (1)$$

For the sake of comparison, success rates of pairwise protein-protein statistical potentials in the scoring of bound protein-protein rigid body docking poses were extracted from the previous work of Pons *et al.*, 2011 [25].

2.3 Energy evaluation of docking poses

Rigid-body docking orientations were also scored by using pyDock binding energy, which is composed of electrostatics, desolvation, and van der Waals terms (Eq. 2-5). The pairwise electrostatics potential (Eq. 2) is based on the Coulombic electrostatics energy with a distance-dependent dielectric constant ($\epsilon = 4.0r$), which was explicitly calculated for all intermolecular ij atom pairs, with q atomic charges from AMBER 94 force field [26]. Pairwise interaction energy values were

truncated to a maximum and minimum of +1.0 and -1.0 kcal/mol, respectively (in order to avoid excessive dependence on incorrect geometries). Desolvation energy (Eq. 3) represents the effective water-to-interface atomic desolvation energy of protein and RNA molecules and it is based on the atomic buried surface areas (BSAs) upon binding, with atomic desolvation parameters (ADPs) originally optimized for rigid-body protein-protein docking [27] and then adapted for protein-RNA complexes [28]. The Lennard-Jones van der Waals energy (Eq. 4) was explicitly calculated for intermolecular atom pairs (i and j atoms from protein and RNA molecules, respectively), with atomic parameters for equilibrium radii (r) and well depth (e) from AMBER 94 force field [26]. Atomic interaction energy values were truncated to a maximum of 1.0 kcal/mol (again to avoid excessive noise from the docking of rigid-body surfaces). The finally used pyDock energy function [29] is defined in Eq. 5, where E_{VDW} is weighted by $W_{VDW} = 0.1$.

$$E_{Ele} = 332.0 \sum_i^{REC} \sum_j^{LIG} \frac{q_i q_j}{4d_{ij}^2} \quad (2)$$

$$E_{Desolv} = \sum_i^{REC} ADP_i BSA_i + \sum_j^{LIG} ADP_j BSA_j \quad (3)$$

$$E_{VDW} = \sum_i^{REC} \sum_j^{LIG} \sqrt{e_i e_j} \left(\left(\frac{r_i + r_j}{d_{ij}} \right)^{12} - 2 \left(\frac{r_i + r_j}{d_{ij}} \right)^6 \right) \quad (4)$$

$$E_{pyDock} = E_{Ele} + E_{Desolv} + W_{VDW} E_{VDW} \quad (5)$$

2.4 Explicit consideration of shape complementarity for scoring

As previously reported, the use of shape complementarity for the scoring of protein-RNA docking poses can improve the predictive rates [15]. Here we scored the rigid-body docking orientations by using the FTDock 2.0 SCscore, which is a correlation score representing the shape complementarity of the components represented as a discrete 3D grid [24]. We then aimed to combine the SCscore with the electrostatics and van der Waals energy terms. Since the variation of the distribution of values for these three terms ranged in the same order, we decided not to use any weighting factor to avoid potential over-fitting problems. Therefore, the three terms were combined in a single scoring function simply by adding the scores, after changing the sign of the SCscore values to be consistent with the other terms (Eq. 6).

$$E_{Comb} = E_{Ele} + E_{VDW} - SCscore \quad (6)$$

2.5 Benchmarking and evaluation of scoring results

The performance of the different scoring functions analyzed here for bound protein-RNA docking was evaluated on the 106 cases from the protein-RNA docking benchmark v1.0 [19], using the bound structures. The benchmark set is composed of experimental protein-RNA complex structures, in which the unbound protein had available structure (or could be homology-based modeled). It includes 5 unbound-unbound, 4 unbound-pseudo-unbound, 62 unbound-bound, 5 unbound-model, 8 model-unbound, 19 model-bound and 3 model-model cases. The benchmark covers all major functional categories and contains cases with different degrees of difficulty for docking, as far as protein and RNA flexibility is concerned. The benchmark set and more detailed

information are available online (<https://life.bsc.es/pid/protein-rna-benchmark/>). On the other hand, the unbound protein-RNA docking was tested on a subset of cases from the same benchmark for which the unbound or modelled structures are available for both protein and RNA components. In particular, this “unbound” data set comprises a mixture of 5 unbound-unbound, 4 unbound-pseudo-unbound, 5 unbound-model, 8 model-unbound and 3 model-model cases. The total of 25 "unbound" cases were grouped in 6 easy, 13 medium and 6 difficult cases, defined according to the unbound-to-bound conformational deformation of their interfaces as those cases with protein and RNA average interface RMSD below 2.5 Å, between 2.5 and 5.0 Å, and above 5.0 Å, respectively [19]. For comparison, protein-protein docking calculations were evaluated on the protein-protein docking benchmark 3.0 [23].

The assessment of the docking solutions was done by calculating the ligand root-mean-square deviation (ligRMSD) with respect to the reference complex structure, defined as the RMSD between the P atoms of the docking and reference RNA molecules, after superimposing the C_α atoms of the docking and reference protein molecules. In oligomeric proteins, in which additional molecules of RNA could bind in symmetric sites, we calculated the ligRMSD with respect to all the different symmetric positions and chose the smallest value. The docking solutions were classified as excellent-accuracy (ligRMSD ≤ 1 Å), high-accuracy (ligRMSD ≤ 2 Å), medium-accuracy (ligRMSD ≤ 5 Å), acceptable (ligRMSD ≤ 10 Å), or incorrect (ligRMSD > 10 Å). Then, the predictive success rates for a given scoring function were computed as the percentage of cases within the corresponding benchmark set in which an acceptable (or any other accuracy criteria) docking solution was found within the top *N* docking models as sorted

by such scoring function. For comparison, we computed the success rates expected by random based on aleatory permutation of the rank of the generated docking solutions for each case.

3 Results

3.1 Rigid body sampling on bound and unbound protein-RNA docking

The predictive success of a docking scoring function strongly depends on the sampling procedures and their ability to generate near-native solutions. We used here FTDock 2.0 [24] to generate rigid-body docking orientations for protein-RNA complexes under both optimal and realistic conditions. We first calculated the number and the quality of the near-native solutions generated by using the bound coordinates for both protein and RNA molecules, that is, in optimal conditions. Out of the 106 cases that compose the protein-RNA docking benchmark v1.0 [19], we found that FTDock is able to generate acceptable solutions in 95% of the cases, and high-accuracy solutions in 84% of the cases. The difficulties of these cases in finding high-accuracy docking solutions could be related to their smaller interface size. Indeed, 53% of these difficult cases had BSA smaller than 800 \AA^2 (for comparison, only 18% of the cases with high-accuracy solutions had $\text{BSA} < 800 \text{ \AA}^2$). This could be associated to lower binding affinity or less specific binding. In any case, increasing the number of docking poses up to 100,000 by running FTDock with different initial random rotations for the interacting molecules did not help to find any high-accuracy solution in those difficult cases. When tested on the more realistic set of "unbound" cases of the benchmark [19], FTDock was able to generate acceptable solutions in only 44% of the cases, which

corresponded to 3 easy (out of 6), 6 medium (out of 13) and 2 difficult (out of 6) cases. On the other hand, it was unable to find any high-quality solution within the set, underlying the fact that conformational sampling could be the most challenging step in blind protein-RNA docking.

3.2 Decoy discrimination by different energy terms in bound protein-RNA docking

We further explored the effectiveness of the different energy terms and functions for the scoring of protein-RNA poses in those bound docking cases in which FTDock generated at least one high-accuracy solution (Fig. 1A; Supplemental Table S1). Strikingly, the desolvation energy term showed very low predictive success rates. The pairwise propensities also showed a quite low success rate, with only around 16% cases with a predicted high-accuracy solution within the top 10 predictions. In contrast, the terms related to shape or structural complementarity, such as the FTDock scoring function (the SCscore) or the van der Waals energy term, achieved the highest success rates, with 54% and 67% successful cases respectively, i.e. containing at least a high-accuracy docking solution (see Methods), within the top 10 predictions. We noted that electrostatics scoring was surprisingly less efficient than FTDock and van der Waals, with 40% successful cases within the top 10 predictions.

Then we directly combined the most effective scoring terms (FTDock SCscore, van der Waals, and electrostatics) into a new protein-RNA scoring function (Eq. 6) that achieved the highest success rate, with 72% successful cases within the top 10 predictions (Fig. 1A; Supplemental Table S1). This combined function clearly outperformed

the protein-protein scoring function pyDock scoring function, which is composed of electrostatics and desolvation energy terms with a minimal consideration of van der Waals energy [27].

[INSERT FIG. 1 HERE]

3.3 The scoring success highly depends on the quality and quantity of the generated near-native solutions

We evaluated the success rate of our combined scoring function according to the quality of the best near-native solutions generated by FTDock. As illustrated in Fig. 2A, we found that success rate is strongly affected by the presence of high-quality models within the pool of 10.000 protein-RNA docking solutions generated by FTDock (Supplemental Table S1). The combined scoring function is able to identify a near-native solution within the top 10 predictions in around 80% of the cases containing only excellent-accuracy docking solutions within the docking pool. The performance is also good when there are only high-accuracy solutions, but its effectiveness dramatically drops when the docking pools contained only medium or acceptable near-native solutions. Indeed, the combined scoring function is able to identify acceptable solutions within the top 10 predictions in only 20% of the cases containing only acceptable solutions within the total number of docking poses, that is, excluding cases of better quality.

We also evaluated the success rates of the combined score according to the number of high-accuracy solutions generated by FTDock (Fig. 2B). We found that success rates dramatically dropped in

those cases in which only one high-accuracy near-native solution was generated (Fig. 2B, blue bars).

[INSERT FIG. 2 HERE]

3.4 Decoy discrimination according to protein flexibility

Since the structure (or a model) of the unbound protein was available for all the 106 benchmark cases, we analyzed whether the success rates of the combined scoring function depended on the flexibility of the RNA-binding proteins, despite using the bound coordinates of these proteins. For this, we grouped the 89 benchmark cases with at least one high-accuracy docking solution generated by FTDock, according to the conformational differences between the unbound and bound protein structures, as rigid (i.e. cases in which RMSD between the unbound and bound protein conformations is less than 2.5 Å), medium (unbound-bound protein RMSD between 2.5 and 5 Å) and highly flexible cases (unbound-bound protein RMSD > 5 Å). Interestingly, medium flexible cases yielded better predictive values than rigid or highly flexible cases (Fig. 3A). To study in more detail these results, we analyzed the predictive performance of each individual scoring term according to protein flexibility. Interestingly, when docking poses were scored only by electrostatics, we found clearly better predictive success rates for the highly flexible cases (Fig. 3B), while when docking poses were scored only by FTDock default function, predictions for highly flexible cases were clearly worse than for rigid or medium flexible cases (Fig. 3C). The predictive results when scoring only by van der

Waals energy term showed the same trends as the FTDock SCScore, but they did not depend as much on unbound-bound protein flexibility (Fig. 3D). We could not perform a similar analysis on the RNA flexibility given that only 25 of the 106 benchmark cases had an available structure (or a good model) for the unbound RNA.

[INSERT FIG. 3 HERE]

3.5 Decoy discrimination in the unbound docking set

In a more realistic test, when using the unbound cases of the benchmark, FTDock generated acceptable or medium-accuracy docking solutions in 11 and 2 cases (out of 25), respectively. However, it was unable to generate high-accuracy solutions in any of the cases. We tested our combined scoring function in these challenging conditions (Table 1). The scoring function was able to identify an acceptable solution within the top 10 predictions for only one case. This case (PDB 1EFW) was probably the easiest case, according to the unbound-to-bound conformational flexibility of the interacting molecules, since it was the only one in which both the protein and the RNA molecules have unbound-bound RMSD $< 2 \text{ \AA}$. Consistently, this was one of the only two cases in which the total docking pool contained at least a medium-accuracy docking solution (Table 1). This confirms the strong dependency of the scoring success on the quality of the generated near-native solutions (as previously described in bound docking conditions). These results are comparable to those obtained in a recent study [30], in which different protein-RNA docking methods were run on the same set of 5 unbound-unbound cases contained within

our "unbound" set (1ASY, 1DFU, 1OB2, 1R3E, 2FMT). They could find near-native solutions in up to 3 of the 5 unbound-unbound cases, with a global best rank of 116 (in our case, we found near-native solutions in 3 of the 5 unbound-unbound cases, with a best rank of 35). In addition, we ran all the 25 cases of our unbound docking set with the NPDock server [22], but in none of the cases it found a near-native solution within the final docking solutions provided by the server (up to 3). This confirms the challenging nature of the unbound protein-RNA docking problem.

[INSERT TABLE 1 HERE]

4 Discussion

4.1 Comparison of the scoring determinants for protein-RNA and protein-protein docking

In order to elucidate the predictive capabilities of the different energy terms and scoring functions in optimal structural conditions, we first used rigid-body docking solutions generated from the bound components of a benchmark set of protein-RNA complexes. One of our main objectives was to evaluate the discriminative power of our previously described residue-ribonucleotide pairwise statistical potentials [15], with the hope that they could be more tolerant to conformational changes in the more realistic unbound protein-RNA docking. Unfortunately, the results indicated that they have a very low predictive value even in bound docking conditions (Fig. 1A). One possible reason is the fact that for each amino acid type, the binding propensities were not significantly different among the four different

ribonucleotides [15]. Therefore, while pairwise propensities can be useful to correctly identify the RNA binding sites on proteins, as previously shown [31], different orientations of RNA molecules that are bound to the correct protein surface are similarly scored. Thus, the use of these pairwise propensities seems too noisy for protein-RNA docking scoring. Contrarily, in protein-protein docking, residue-residue pairwise propensities showed a statistically significantly better predictive value, indicating a higher specificity in the residue-residue contacts than in the residue-ribonucleotide ones (Fig. 1B).

On the other hand, our results indicate that the FTDock score and van der Waals energy, when considered individually, had significantly better predictive rates for protein-RNA bound docking than for protein-protein docking (Fig. 1). Indeed, the FTDock score capabilities for protein-RNA docking were previously suggested on a much more limited set of cases [15]. Therefore, these results indicate an essential and distinctive role of shape and structural complementarity in protein-RNA association, where the lack of clear preferences of amino acids to bind different ribonucleotides could be compensated by geometric constraints that could define for instance H-bonding patterns for specific recognition. Interestingly, electrostatics had slightly better predictive rates for the scoring of protein-RNA docking poses, while the desolvation energy term showed better success rate for protein-protein docking (Fig. 1). Taking all of the above into consideration, the behaviour of all these energy terms for the scoring of protein-RNA and protein-protein docking poses explains why our combined scoring function (composed of the FTDock score, electrostatics and van der Waals energy terms) and the pyDock scoring function (basically composed of the electrostatics and desolvation energy terms) have

different success rates for the scoring of either protein-RNA or protein-protein complexes (Fig. 1). The predictive capabilities of the combined scoring function show that the simple integration of these complementary terms is beneficial for the scoring of protein-RNA docking poses. Future work will focus on the optimization of this combined function to further improve the predictive success rates for realistic protein-RNA docking.

4.2 The role of electrostatics in protein-RNA binding depends on protein flexibility

We found here that the role of electrostatics in protein-RNA bound docking is strongly dependent on the unbound-to-bound conformational flexibility of the RNA-binding proteins. Indeed, electrostatics was especially successful for the highly flexible cases (Fig. 3B). Based on that, one would expect that in highly flexible cases the electrostatics contribution to binding affinity is much more important than in rigid or medium flexible cases. This would be consistent with a situation in which highly flexible proteins would have a larger enthalpic contribution to compensate for the larger conformational entropy penalization. Electrostatics could provide this general enthalpic gain, since it is more tolerant to conformational flexibility as compared to other terms like van der Waals or hydrogen bonding. To confirm this, we studied the number and type of charged residues at the RNA-binding sites in proteins, according to their flexibility. We defined the interface net charge as the difference between the number of positively-charged Arg/Lys and negatively-charged Asp/Glu protein residues that are found within 5 Å from the RNA molecule (Supplemental Table S2). Fig. 4 shows the distribution of the interface net charge values for the different groups of proteins according to unbound-to-bound flexibility.

As much as 81% of the highly flexible proteins have RNA-binding interfaces with larger positive net charge (i.e. $> +5$), as compared to 52% of the medium flexible or rigid proteins. Interestingly, although rigid proteins have less positive interfaces in average, the range of the distribution is broader, with a few extreme cases, from the most positively charged (1C9S and 2GIC, both involving large protein oligomers) to the more negatively charged ones (1E1Y, a tRNA-synthetase from *Thermus Thermophilus*). The higher percentage of positively charged interfaces in the flexible proteins were not caused by a larger interface size, since flexible proteins had in general smaller interfaces than the rigid ones (see section 4.4; Supplemental Table S2). All these findings are consistent with the above mentioned hypothesis that protein-RNA interfaces involving flexible proteins are more electrostatic.

[INSERT FIG. 4 HERE]

Fig. 5 shows two examples of protein-RNA complexes involving rigid and high flexible proteins, respectively. The more flexible case has a higher proportion of Arg/Lys residues, which yields a more positive interface and more favourable electrostatics.

[INSERT FIG. 5 HERE]

4.3 OPRA-based predictions on proteins shows more favourable RNA-binding patches on highly flexible proteins

We applied OPRA predictions [31] to all available unbound coordinates of the proteins in the protein-RNA docking benchmark [19]. As shown in Fig. 6, we found a clear better success rate for OPRA predictions in the highly flexible proteins, despite the fact that we used the unbound coordinates of these proteins, and therefore their conformations were different from those in the native interface. As OPRA predictions are based on residue statistical propensities largely related to the electrostatic character of each residue [31], the results are consistent with the hypothesis that highly flexible proteins could have evolved to have optimal electrostatic patches for RNA recognition to compensate conformational entropy penalization.

[INCLUDE FIG. 6 HERE]

4.4 RNA-binding sites of flexible proteins are smaller and have less surface complementarity

We also found that the role of structural complementarity (as defined by SCscore) in protein-RNA bound docking is strongly dependent on the unbound-to-bound conformational flexibility of the RNA-binding proteins. Indeed, the SCscore yielded much worse scoring for the highly flexible cases than for the rigid or medium flexible cases, despite using the bound coordinates of the protein and RNA components (Fig. 3C). Perhaps highly flexible proteins, which rely more on long-range electrostatics interactions as above described, do

not need to form highly complementary interfaces, which would involve a significant entropy penalization. Indeed, the SCScore values obtained for the native interfaces of protein-RNA complexes involving highly flexible proteins were less favourable in average (246.4 ± 66.2 a.u.) than those of the complexes involving medium flexible (270.4 ± 86.1 a.u.) or rigid proteins (285.6 ± 163.2 a.u.). This could indicate that protein-RNA interfaces involving highly flexible proteins were in average less packed than those of rigid or medium flexible proteins. Consistently, we found that none of the highly flexible proteins had a large interface (i.e. $BSA > 1500 \text{ \AA}^2$). On the contrary, 12% of the medium flexible proteins, and 28% of the rigid ones had such large interfaces, with a few rigid proteins showing the largest interface size among all cases (e.g. 1C9S and 2GIC, both involving large protein oligomers). Accordingly, scoring by van der Waals energy alone was also worse for cases involving highly flexible proteins (Fig. 3D). However, the difference in performance according to unbound-to-bound conformational flexibility was not as large as that yielded by SCScore scoring (Fig. 3C), perhaps due to the noisier character of the van der Waals scoring.

4.5 Present and future challenges in protein-RNA docking

We found here that a combination of structural complementarity and electrostatic parameters is successful for the scoring of easy protein-RNA docking cases. However, given that most of the known protein-RNA complexes involve significant local and global conformational changes [32–34], further improvement in the sampling and scoring methodology is much needed. One possibility is that scoring functions could integrate more coarse-grained parameters to deal with

inaccuracies derived from sub-optimal conformational sampling in unbound protein-RNA docking. Although we have found here that the use of low-resolution (or residue-level) pairwise propensities is noisy, other previously reported medium resolution propensities could be more effective [16,17] and it remains to be seen whether they could complement the energy terms described in this work. We also found that scoring (either atomic or coarse-grained) becomes extremely challenging for evaluating solutions that differ more than 5 Å RMSD from the bound structures. Thus, despite some scoring schemes have been previously reported to successfully discriminate docking decoys in the presence of multiple high-quality models derived from perturbing the original complex structures, their efficiency might decrease in more realistic conditions [17,18,20–22]. In this context, one of the future major challenges for protein-RNA docking will be modelling the flexibility of the protein or RNA molecules, which should involve a combination of techniques, such as molecular dynamics or normal mode analysis. While major efforts have been made to study protein dynamics during the last decades, biophysical and computational approaches have just begun to shed light on the complexity of RNA dynamics. Modeling the flexibility of RNA molecules requires different parameterization than that used for proteins [35] and involves two main difficulties: the non-ergodic behaviour of RNA and the strong influence of ion molecules on the dynamic processes [36]. Despite this big challenge, all the accumulated experience on the modelling of the structure and dynamics of protein-protein complexes is expected to be useful as a framework that will accelerate the development of efficient docking algorithms for protein-RNA interactions.

5 Conclusions

We have developed and tested a new protein-RNA docking protocol using FTDock to generate docking poses, which are subsequently scored by a combined function, composed of the FTDock score and the pyDock electrostatics and van der Waals energy terms. This combined scoring function showed good discrimination of near-native docking poses in bound docking, but for the more realistic unbound docking, the scoring success strongly depended on the quantity and quality of near-native solutions generated. Thus, one of the major challenges in protein-RNA docking is to improve the description of the conformational flexibility of the interacting molecules. Compared to protein-protein interactions, protein-RNA interactions show lack of specificity for residue-ribonucleotide contacts, which seems to be compensated by a higher shape complementarity that could define specific H-bonding patterns. Interestingly, the RNA recognition mechanism seems to depend on the protein flexibility. Protein-RNA complexes involving rigid proteins show larger and more packed interfaces, while those involving flexible proteins show more positively charged interfaces but not as large in size as those of the rigid proteins. Thus, electrostatics is found to play a major role in highly flexible RNA-binding proteins, likely to compensate the strong conformational entropy penalization, while structural complementarity is more determinant in rigid RNA-binding proteins, in which conformational entropy penalization is much smaller.

Acknowledgements

This work is supported by grant BIO2013-48213-R from Plan Nacional I+D+i (Spanish Ministry of Economy and Competitiveness). LP-C was recipient of an FPU fellowship from the Spanish Ministry of Science.

References

- [1] N. Ban, P. Nissen, J. Hansen, P.B. Moore, T.A. Steitz, The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution, *Science*. 289 (2000) 905–920.
- [2] W.P. Galej, M.E. Wilkinson, S.M. Fica, C. Oubridge, A.J. Newman, K. Nagai, Cryo-EM structure of the spliceosome immediately after branching, *Nature*. 537 (2016) 197–201.
- [3] A.K.W. Leung, K. Nagai, J. Li, Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis, *Nature*. 473 (2011) 536–539.
- [4] D.A. Pomeranz Krummel, C. Oubridge, A.K.W. Leung, J. Li, K. Nagai, Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution, *Nature*. 458 (2009) 475–480.
- [5] R. Wan, C. Yan, R. Bai, G. Huang, Y. Shi, Structure of a yeast catalytic step I spliceosome at 3.4 Å resolution, *Science*. 353 (2016) 895–904.
- [6] C.L. Will, R. Lührmann, Spliceosome structure and function, *Cold Spring Harb. Perspect. Biol.* 3 (2011).
- [7] B.T. Wimberly, D.E. Brodersen, W.M. Clemons Jr, R.J. Morgan-Warren, A.P. Carter, C. Vornrhein, T. Hartsch, V. Ramakrishnan, Structure of the 30S ribosomal subunit, *Nature*. 407 (2000) 327–339.
- [8] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, The protein data bank: A computer-based archival file for macromolecular structures, *Archives of Biochemistry and Biophysics*. 185 (1978) 584–591.
- [9] V.N. Bavro, Z. Pietras, N. Furnham, L. Pérez-Cano, J. Fernández-Recio, X.Y. Pei, R. Misra, B. Luisi, Assembly and channel opening in a bacterial drug efflux machine, *Mol. Cell*. 30 (2008) 114–121.
- [10] C. Pallara, B. Jiménez-García, L. Pérez-Cano, M. Romero-Durana, A. Solernou, S. Grosdidier, C. Pons, I.H. Moal, J. Fernandez-Recio, Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges, *Proteins*. 81 (2013) 2192–2200.
- [11] J. Janin, Protein-protein docking tested in blind predictions: the CAPRI experiment, *Mol. Biosyst.* 6 (2010) 2351–2362.
- [12] J. Janin, The targets of CAPRI Rounds 13–19, *Proteins*. 78 (2010) 3067–3072.

- [13] R.R. Walia, C. Caragea, B.A. Lewis, F. Towfic, M. Terribilini, Y. El-Manzalawy, D. Dobbs, V. Honavar, Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art, *BMC Bioinformatics*. 13 (2012) 89.
- [14] Y. Chen, T. Kortemme, T. Robertson, D. Baker, G. Varani, A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys, *Nucleic Acids Res.* 32 (2004) 5147–5162.
- [15] L. Perez-Cano, A. Solernou, C. Pons, J. Fernández-Recio, Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials, *Pacific Symposium on Biocomputing* (2010) 269–280.
- [16] P. Setny, M. Zacharias, A coarse-grained force field for Protein-RNA docking, *Nucleic Acids Res.* 39 (2011) 9118–9129.
- [17] I. Tuszynska, J.M. Bujnicki, DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking, *BMC Bioinformatics*. 12 (2011) 348.
- [18] S. Zheng, T.A. Robertson, G. Varani, A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins, *FEBS J.* 274 (2007) 6378–6391.
- [19] L. Pérez-Cano, B. Jiménez-García, J. Fernández-Recio, A protein-RNA docking benchmark (II): Extended set from experimental and homology modeling data, *Proteins*. 80 (2012) 1872–1882.
- [20] A. Guilhot-Gaudeffroy, C. Froidevaux, J. Aze, J. Bernauer, Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities., *PLoS One*. 9 (2014) e108928.
- [21] S.-Y. Huang, X. Zou, A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method., *Nucleic Acids Res.* 42 (2014) e55.
- [22] I. Tuszynska, M. Magnus, K. Jonak, W. Dawson, J.M. Bujnicki, NPdock: a web server for protein–nucleic acid docking, *Nucl. Acids Res.* (2015) W425-30.
- [23] H. Hwang, B. Pierce, J. Mintseris, J. Janin, Z. Weng, Protein–protein docking benchmark version 3.0, *Proteins: Structure, Function, and Bioinformatics*. 73 (2008) 705–709.
- [24] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, I.A. Vakser, Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques., *Proc Natl Acad Sci U S A.* 89 (1992) 2195–2199.
- [25] C. Pons, D. Talavera, X. de la Cruz, M. Orozco, J. Fernandez-Recio, Scoring by Intermolecular Pairwise Propensities of

- Exposed Residues (SIPPER): A New Efficient Potential for Protein–Protein Docking, *J. Chem. Inf. Model.* 51 (2011) 370–377.
- [26] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, *J. Am. Chem. Soc.* 117 (1995) 5179–5197.
- [27] J. Fernández-Recio, M. Totrov, R. Abagyan, ICM-DISCO docking by global energy optimization with fully flexible side-chains, *Proteins.* 52 (2003) 113–117.
- [28] C. Pons, A. Solernou, L. Perez-Cano, S. Grosdidier, J. Fernandez-Recio, Optimization of pyDock for the new CAPRI challenges: Docking of homology-based models, domain-domain assembly and protein-RNA binding, *Proteins.* 78 (2010) 3182–3188.
- [29] T.M.-K. Cheng, T.L. Blundell, J. Fernandez-Recio, pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking, *Proteins.* 68 (2007) 503–515.
- [30] J. Iwakiri, M. Hamada, K. Asai, T. Kameda, Improved Accuracy in RNA–Protein Rigid Body Docking by Incorporating Force Field for Molecular Dynamics Simulation into the Scoring Function, *J. Chem. Theory Comput.* 12 (2016) 4688–4697.
- [31] L. Pérez-Cano, J. Fernández-Recio, Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins, *Proteins.* 78 (2010) 25–35.
- [32] C. Dominguez, M. Schubert, O. Duss, S. Ravindranathan, F.H.-T. Allain, Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy, *Prog. Nucl. Magn. Reson. Spectrosc.* 58 (2011) 1–61.
- [33] C. Hyeon, R.I. Dima, D. Thirumalai, Size, shape, and flexibility of RNA structures, *J. Chem. Phys.* 125 (2006) 194905.
- [34] A. Ke, J.A. Doudna, Crystallization of RNA and RNA-protein complexes, *Methods.* 34 (2004) 408–414.
- [35] S. Fulle, H. Gohlke, Analyzing the flexibility of RNA structures by constraint counting, *Biophys. J.* 94 (2008) 4202–4219.
- [36] H.M. Al-Hashimi, N.G. Walter, RNA dynamics: it is about time, *Curr. Opin. Struct. Biol.* 18 (2008) 321–329.

Figure Legends

Figure 1. Predictive success rates for different scoring functions on docking poses generated for bound protein-RNA (A) and protein-protein (B) benchmark data sets. For each scoring function, the percentage of cases with high-accuracy docking solutions within the top N number of predictions is shown. The following scoring functions are shown: FTDock-SCscore (purple), residue-ribonucleotide statistical propensities (red), electrostatics (yellow), desolvation (green), van der Waals (brown), combined scoring (cyan), pyDock (grey), random (black). Only those cases in which FTDock generated at least one high-accuracy solution were considered.

Figure 2. Success rates of the combined scoring function for protein-RNA docking poses as a function of the quality of the near-native docking models contained in the docking sets. (A) Percentage of cases in which a near-native solution is found within the top 1, 5 or 10 predictions considering only the cases with excellent-accuracy docking solutions (blue), high-accuracy or worse (red), medium-accuracy (yellow) or acceptable (green). (B) Percentage of cases in which a high-accuracy docking solution is found within the top 1, 5 or 10 predictions among the cases containing a total of one (blue), between one and five (red), or more than five (yellow) high-accuracy docking solutions.

Figure 3. Success rates for different scoring functions on the benchmark cases as grouped according to the unbound-to-bound

conformational flexibility of the protein. (A) Combined score. (B) Electrostatics. (C) FTDock SCScore. (D) Van der Waals. Results for rigid proteins (blue), medium flexible proteins (red), and high flexible proteins (green) are shown. Only those cases in which FTDock generated at least one high-accuracy solution were considered.

Figure 4. Distribution of interface net charge of native protein-RNA interfaces according to the unbound-to-bound conformational flexibility of the protein. The histogram bar values show the normalized population for each range of interface net charge values. A smoothed curve representing the values has been added for a clearer visualization.

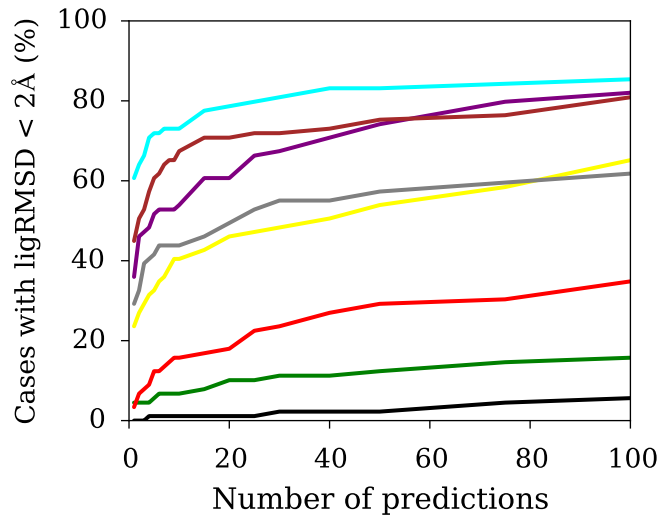
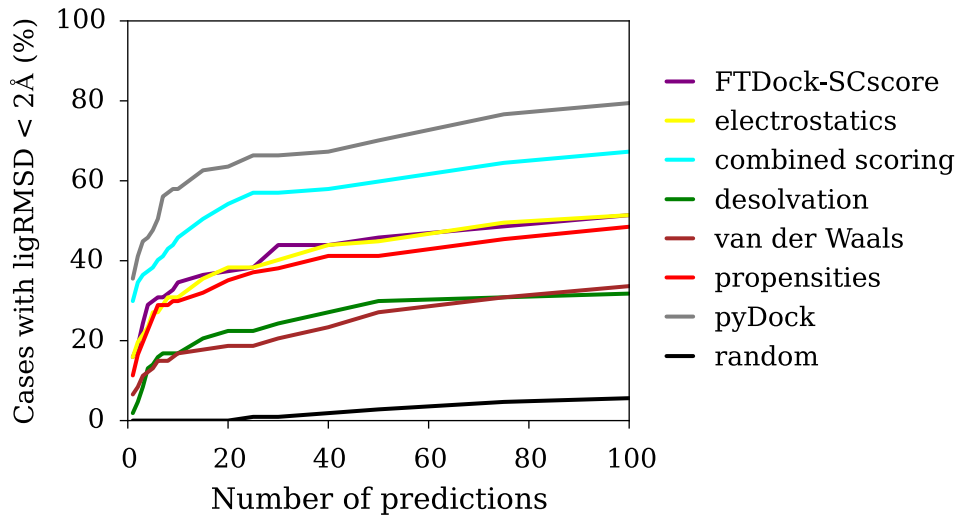
Figure 5. Location of positively (blue) and negatively (red) charged residues at protein-RNA interfaces in two examples. (A) A rigid case: PDB code 1N78 (interface net charge +2). (B) A highly flexible case: PDB code 1OOA (interface net charge +8).

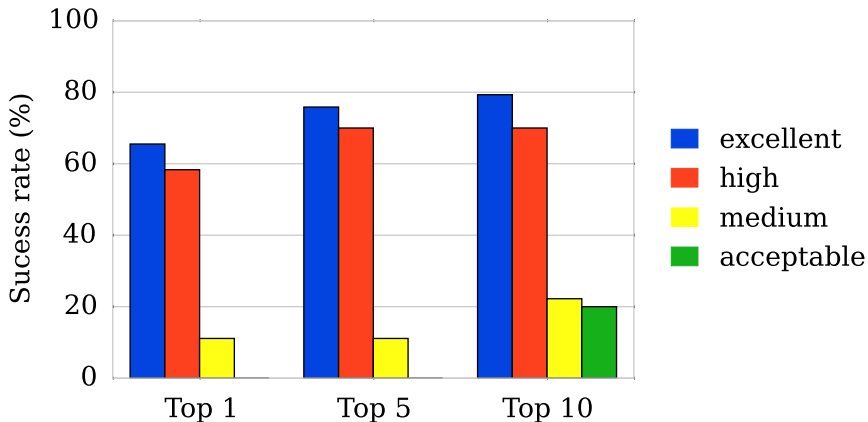
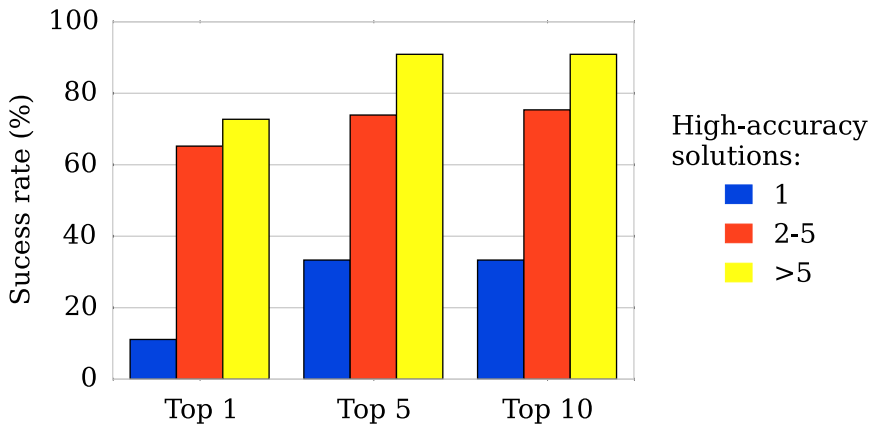
Figure 6. OPRA predictive rates on the unbound proteins of the protein-RNA docking benchmark according to the unbound-to-bound conformational flexibility of the protein. Results for rigid proteins in blue, medium flexible proteins in red, and high flexible proteins in green.

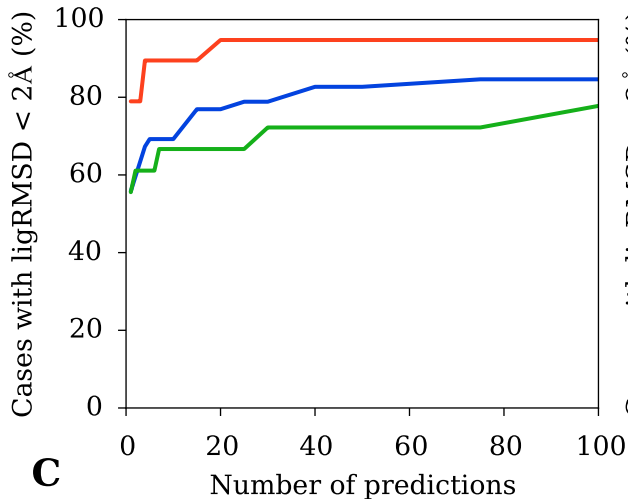
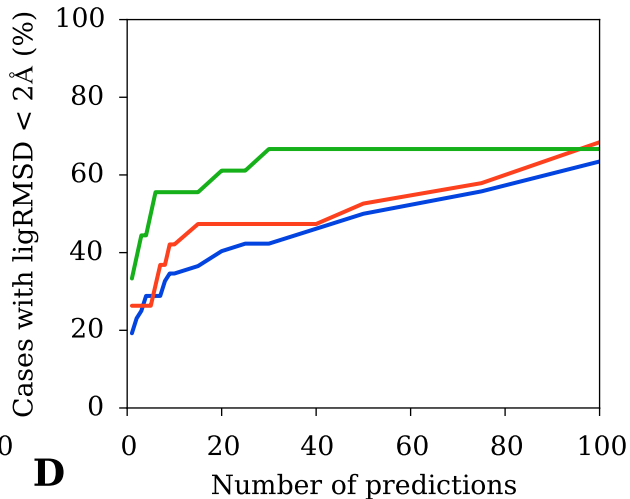
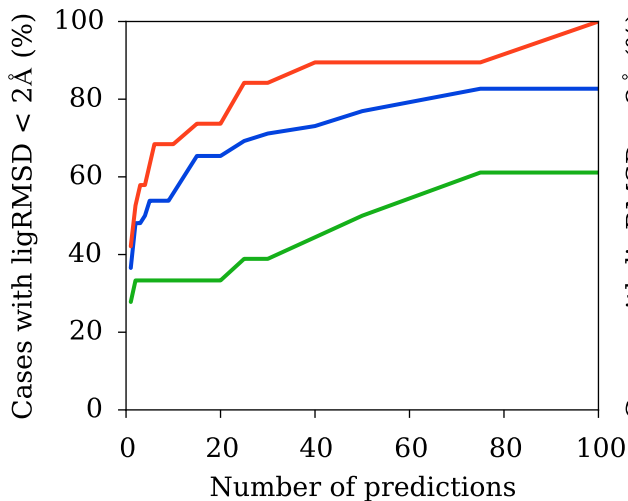
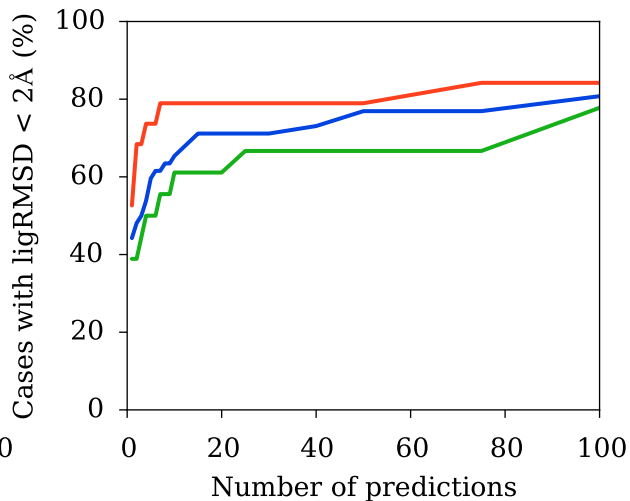
Table 1. Results for scoring with the unbound docking set

Complex ^a	Receptor ^b	Ligand ^b	Rank ^c	RMSD ^d	f _{nat} ^e
1DFU_p:n,m (2)	1B75_a (4.3)	364D_b,c (3.3)	>1000	9.8 (9.8)	0.05
1R3E_a:c,d,e (3)	1ZE1_a (5.2)	1EHZ_a (2.7)	562	9.0 (7.1)	0.36
2FMT_a:c (1)	1FMT_a (1.2)	3CW5_a (2.9)	32	8.1 (7.9)	0.27
1EFW_a:c (1)	1L0W_a (1.3)	1C0A_b (1.6)	7	4.2 (4.2)	0.57
2DRA_a:b (1)	1R89_a (1.9)	1VFG_d (3.0)^f	83	9.9 (4.5)	0.23
1FEU_a:b,c (2)	1NJP_t (4.4) ^f	1A4D_a,b (5.0)	816	9.2 (5.9)	0.28
1HQ1_a:b (2)	1KVV_a (1.2) ^f	1CQL_a (9.5)	>1000	10.0 (10.0)	0.45
2V3C_c:m (3)	3DM5_a (13.0) ^f	1Z43_a (1.7)	>1000	9.4 (9.4)	0.00
2ZKO_a,b:c,d (2)	2Z0A_b (1.0) ^f	2ZI0_c,d (4.4)	>1000	7.9 (7.6)	0.33
1VFG_a:c (2)	3H38_a (4.0) ^f	1E1Y_c (3.1) ^f	>1000	8.6 (8.6)	0.10
2DU3_a,b,c,d:e,f (2)	2DU7_a,b,c,d (4.5) ^f	2ZZN_d (3.1) ^f	>1000	8.0 (8.0)	0.17

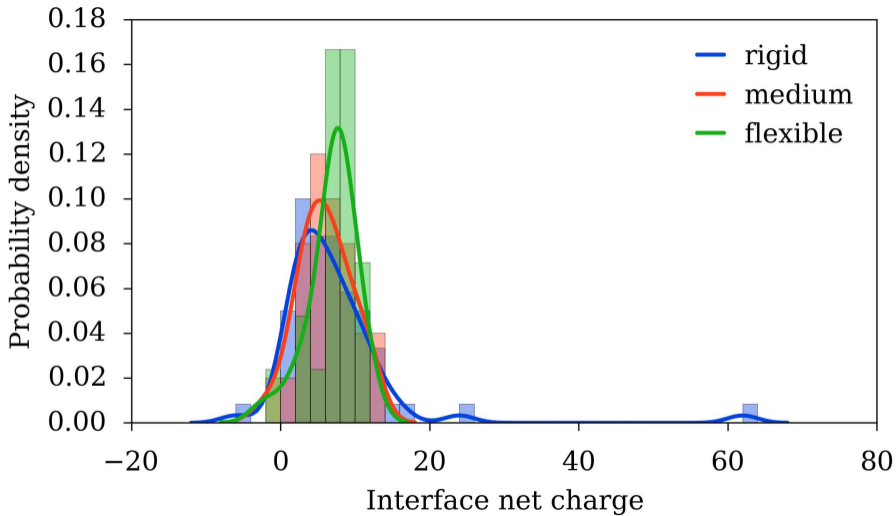
^a PDB code of the protein-RNA complex. Protein and RNA chains are separated by colon. In brackets, each case is classified as easy (1), medium (2) and difficult (3), according to the RMSD of the P and C_α atoms composing the interface [17]. ^b PDB code of the protein and RNA subunits, followed by chain IDs. Numbers in brackets correspond to RMSD (C_α atoms for protein and P atoms for RNA) with respect to the bound reference structure. ^c Best rank of any acceptable solution. ^d RMSD between the P atoms of the RNA molecule in the best-ranked acceptable solution and the corresponding P atoms of the bound structure, after superimposing the C_α atoms of the protein onto the corresponding ones of the native complex (ligRMSD). In brackets are shown the best ligRMSD of any acceptable solution within the total pool of generated solutions (in bold are shown the only two cases with a medium-accuracy solution). ^e Fraction of native contacts, defined as the number of native (correct) residue-ribonucleotide contacts (i.e. if any of their atoms were within 5Å distance) in the docking model divided by the number of contacts in the reference complex (native). ^f PDB code of the templates used to build the models.

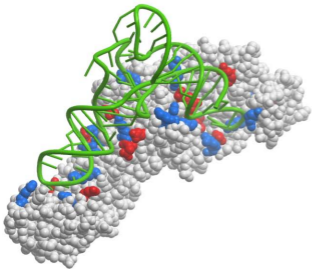
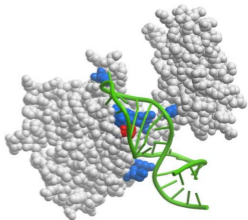
A**B**

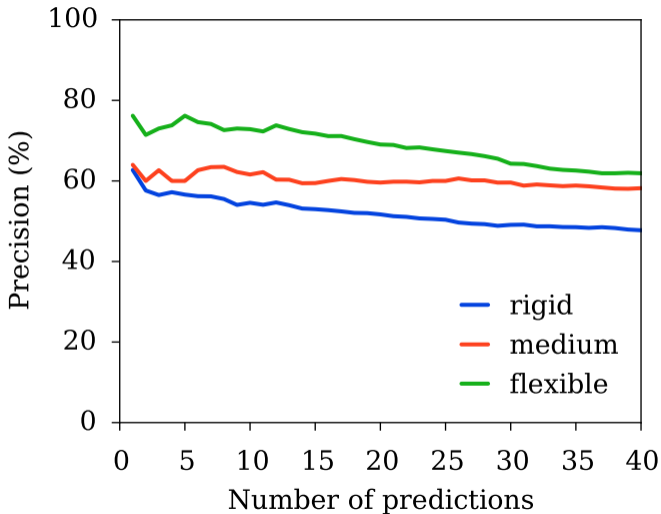
A**B**

A**B****C****D**

— rigid — medium — flexible



A**B**



Supplemental Table S1: Scoring results for the bound docking, only cases with high-accuracy docking solutions.

Complex PDB, number of hits (high-accuracy docking solutions), best ligRMSD (Å) of any docking solution with respect to reference complex, best rank for any high-accuracy docking solution according to the indicated scoring function.

complex	hits	best RMSD	Rank							combined
			FTDock-SCscore	electrostatics	desolvation	van der Waals	propensities	pyDock		
1ASY	3	0.68	32	1667	4901	1	100	1888	1	
1B23	3	0.82	103	646	2403	1	2263	766	7	
1B7F	3	1.09	1	1	1	1	1	1	1	
1C0A	3	0.76	1	238	5556	1	4346	445	1	
1C9S	3	0.80	1	1	50	77	118	1	1	
1DDL	4	1.23	107	526	1870	8	499	398	33	
1DFU	3	1.08	25	78	4493	1	276	77	1	
1DK1	3	1.04	1	6	5572	69	35	13	1	
1E7K	3	0.97	77	73	1128	168	930	55	4	
1EC6	1	1.83	2	59	2296	2	5	20	3	
1EFW	3	0.77	72	78	7179	11	4860	116	13	
1EIY	3	1.08	1059	7018	9824	1	3319	7404	3512	
1EKZ	3	1.35	474	2216	307	4436	1741	1523	1281	
1F7U	3	1.05	1	1	7063	1	4198	1	1	
1FEU	3	1.20	1	325	6919	1	1263	414	1	
1FXL	8	1.04	2	1	176	1	2	1	1	
1H3E	2	1.30	7493	2	8463	8299	1649	3	197	
1H4S	2	1.24	11	380	738	14	348	465	4	
1HC8	3	1.00	1	185	3871	1	1730	183	1	
1HQ1	3	1.02	5	1219	550	1	110	1157	21	
1J1U	7	1.00	1	196	1	166	366	183	1	
1J2B	3	0.87	1	1	9922	1	9998	1	1	
1JBR	3	1.33	116	615	582	110	467	612	124	
1K1G	2	1.88	1	6	27	91	211	1	1	
1K8W	3	1.04	1	1	4195	1	89	1	1	
1KOG	1	1.91	156	219	369	913	2721	264	127	
1KQ2	15	0.97	2	36	1	2	1	1	1	
1LNG	3	1.28	3	15	1267	1294	25	18	1	
1M5O	3	0.82	14	1802	798	1	1333	1739	385	
1M8V	3	1.37	56	670	53	191	572	572	603	
1M8W	17	0.66	1	81	6	4	13	1	1	
1MFQ	3	0.92	82	9002	1180	4	4632	8409	5294	
1MJI	1	1.49	25	110	6992	1	233	138	18	
1MMS	3	0.83	6	7	1548	1	1234	4	1	
1N78	3	0.92	1	8	5723	1	728	42	1	
1OB2	3	1.27	195	817	1950	3	2018	651	129	
1OOA	3	1.11	71	1	4622	10	4	1	1	
1Q2R	2	1.81	2	1	2588	5	2	1	1	
1QF6	3	1.32	1	90	6233	1	5394	194	1	
1QTQ	3	1.13	1	2	9998	1	6947	2	1	
1R3E	4	0.90	75	5	3639	1	42	3	1	
1RLG	2	1.07	182	1112	421	2663	29	1346	662	

complex	hits	best RMSD	Rank							combined
			FTDock-SCscore	electrostatics	desolvation	van der Waals	propensities	pyDock		
1T0K	1	1.26	2949	72	80	47	3780	23	11	
1T4L	3	0.89	235	85	1760	480	588	29	4	
1U0B	3	1.25	43	41	6557	1	1173	169	1	
1U63	2	0.98	105	99	4702	1	632	122	33	
1WNE	3	0.61	2	19	9705	1	818	94	1	
1WPU	7	0.72	10	274	396	40	471	238	130	
1YVP	3	0.80	5	1	3279	1	2083	1	1	
2AD9	10	1.37	33	1	191	2	39	1	4	
2ADB	5	1.21	5	1	677	297	162	3	1	
2ASB	3	1.44	1	4	16	1	1698	1	1	
2AZ0	1	1.41	1	48	8522	10	538	122	1	
2B3J	7	1.03	11	3	712	3	87	1	1	
2BGG	3	1.27	1	9	810	292	333	5	1	
2BH2	3	1.06	1	1	3988	490	46	1	1	
2BTE	3	1.28	13	115	3513	1	4693	107	1	
2C0B	1	1.77	119	30	2150	24	97	30	26	
2CSX	3	0.83	15	575	1685	1	2407	713	12	
2CZJ	2	1.45	25	1	1615	11	158	1	1	
2DB3	5	1.60	129	3	2566	4	226	17	2	
2DER	3	0.87	1	44	6002	1	149	42	1	
2DLC	2	1.98	2533	1443	925	4333	6527	1148	1608	
2DRA	3	1.00	2	175	4489	1	1472	234	1	
2ERR	7	1.05	1	109	1	7	25	6	1	
2F8K	1	1.91	61	8	4427	1940	20	24	11	
2FK6	1	1.81	21	4	4396	5	579	3	2	
2FMT	3	1.32	12	1	7343	1	22	1	1	
2FY1	3	0.96	1	1	69	78	8	1	1	
2GJW	3	0.82	1	21	4049	746	35	12	1	
2HGH	3	1.34	36	1	7684	584	545	1	1	
2HW8	3	1.41	1	19	2515	7	378	21	1	
2I82	3	1.29	2	1	3513	1	23	1	1	
2I91	2	1.23	1	1	9959	1	5	1	1	
2IPY	2	1.30	21	1	2616	1	61	1	1	
2IX1	3	1.20	1	33	398	1	9	3	1	
2JPP	4	0.69	2	1	1449	2	7	1	1	
2NUG	3	1.35	1	407	5583	1	136	540	1	
2QUX	4	1.15	1	17	2733	1	189	6	1	
2R7R	2	1.49	26	1	413	4	5	2	5	
2R8S	3	0.86	4	14	1971	49	1074	1	1	
2V3C	3	1.20	50	467	9494	1	327	813	81	
2ZKO	3	1.21	1	57	5578	1	5746	65	1	
3BO2	3	1.17	47	3229	881	6	4907	3026	1320	
3BSB	8	0.64	1	606	17	5	1	233	3	
3BSO	3	1.45	2	2	9519	1	430	1	2	
3BSX	8	1.23	1	222	14	1	3	2	1	
3BX2	6	1.01	1	9	5	2	2	3	1	
3CIY	1	1.96	162	920	8170	92	1364	1157	67	

Supplemental Table S2: Parameters for the native protein-RNA interfaces.

Complex PDB, number of interface residues, net charge, buried surface area (\AA^2), complex type, flexibility type.

complex	# interface residues	net charge	BSA protein	type	flexibility
1ASY	126	+4	2088	1	rigid
1DFU	23	+3	815	2	medium
1OB2	43	+7	1241	1	flexible
1R3E	38	+11	1349	1	flexible
2FMT	35	+11	1573	1	rigid
1B23	42	+7	1286	1	flexible
1MFQ	15	+1	538	5	medium
1QTQ	72	+9	2360	1	rigid
1U0B	67	+9	2099	1	rigid
1B7F	41	+9	1348	9	flexible
1C9S	138	+24	16466	9	rigid
1DK1	33	+12	1226	2	medium
1E7K	18	+3	610	4	medium
1EC6	24	+6	873	8	rigid
1EIY	142	-6	2143	1	rigid
1EKZ	15	+4	596	6	flexible
1F7U	76	+11	2591	1	medium
1G1X	5	+3	179	2	rigid
1H3E	40	+8	1223	1	flexible
1H4S	32	+1	1123	1	rigid
1HC8	31	+2	990	2	medium
1HVU	26	+3	621	9	flexible
1JBR	24	+8	702	2	rigid
1K8W	39	+6	1407	1	rigid
1KOG	23	+2	851	6	rigid
1KQ2	27	+13	1358	9	rigid
1M5O	27	+4	869	3	rigid
1M8V	11	0	309	9	rigid
1M8W	27	+5	941	6	rigid
1MMS	35	+7	1200	2	medium
1N78	68	+2	2064	1	rigid
1Q2R	32	+6	1232	1	rigid
1SER	34	+12	1095	1	rigid
1T0K	15	+3	466	6	rigid
1T4L	22	+4	924	4	rigid
1U63	24	+8	1135	2	rigid
1WNE	44	+1	1426	6	rigid
1WPU	24	+3	608	6	rigid
1WSU	12	+2	444	6	rigid
1YVP	31	+9	1072	9	rigid
2AD9	18	+6	617	9	medium
2ADB	23	+4	594	9	medium
2ADC	52	+11	1497	9	medium

complex	# interface residues	net charge	BSA protein	type	flexibility
2ASB	36	+6	1076	6	rigid
2AZ0	31	+4	1090	9	rigid
2AZX	58	+10	1958	1	rigid
2BGG	24	+4	1008	6	rigid
2BH2	59	+11	2067	2	rigid
2BTE	54	+5	1628	1	medium
2BU1	16	+4	397	7	rigid
2C0B	31	+8	989	9	flexible
2CZJ	49	+8	1799	6	rigid
2ERR	25	+4	840	9	medium
2F8K	10	+7	429	9	rigid
2FK6	43	+12	1095	1	rigid
2GIC	278	+62	10013	7	rigid
2GJE	26	+8	859	6	medium
2GJW	35	+11	1508	9	rigid
2HGH	27	+11	1314	2	flexible
2HW8	31	+11	1170	6	flexible
2I91	51	+15	1832	2	rigid
2IX1	62	+10	1877	10	rigid
2PY9	16	+4	492	6	medium
2QUX	29	+3	780	7	rigid
2R7R	33	+7	890	9	rigid
3BO2	24	+3	860	3	rigid
3BSB	35	+7	1203	6	rigid
3BSO	52	+3	1513	9	rigid
3BSX	34	+6	1205	6	rigid
3BX2	34	+5	1100	6	medium
3CIY	26	+8	1133	9	rigid
1C0A	64	+7	2140	1	rigid
1EFW	33	+1	1201	1	rigid
1J1U	64	+16	1049	1	rigid
1J2B	94	+13	3254	1	rigid
2DRA	45	+10	1363	1	rigid
1FEU	21	+7	789	2	medium
1HQ1	17	+5	669	5	rigid
1LNG	26	+7	1099	5	medium
1OOA	23	+8	951	8	flexible
1RKJ	35	+7	1056	2	flexible
2R8S	34	0	1181	3	rigid
2V3C	43	+7	1371	5	flexible
2ZKO	29	+2	1197	9	rigid
1DDL	22	+3	572	7	rigid
1E8O	13	+5	519	5	rigid
1FXL	39	+9	1139	6	flexible
1K1G	35	+8	1142	6	flexible
1MJI	35	+9	862	2	medium
1RLG	19	+1	631	10	rigid
1S03	25	+3	833	6	rigid
2B3J	54	+7	1011	1	rigid

complex	# interface residues	net charge	BSA protein	type	flexibility
2CJK	41	+7	1165	9	flexible
2CSX	31	+5	1023	1	rigid
2D6F	45	+3	616	1	medium
2DB3	20	+2	492	9	flexible
2DER	36	+6	1125	1	medium
2DLC	62	-2	999	1	flexible
2FY1	26	+8	1072	8	flexible
2I82	35	+12	1476	1	medium
2IPY	43	+6	1331	6	flexible
2JPP	55	+8	970	6	medium
2NUG	44	+7	1479	9	flexible
1QF6	60	+9	2231	1	medium
1VFG	12	+5	422	1	medium
2DU3	36	-2	631	1	medium