Bioinformatics, 2019, 1–8
doi: 10.1093/bioinformatics/btz478
Advance Access Publication Date: 7 June 2019
Original Paper

Structural bioinformatics

P3DOCK: a protein–RNA docking webserver based on template-based and template-free docking

Jinfang Zheng[†], Xu Hong[†], Juan Xie, Xiaoxue Tong and Shiyong Liu *

School of Physics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jan Gorodkin

Received on January 2, 2019; revised on May 24, 2019; editorial decision on June 1, 2019; accepted on June 4, 2019

Abstract

Motivation: The main function of protein–RNA interaction is to regulate the expression of genes. Therefore, studying protein–RNA interactions is of great significance. The information of three-dimensional (3D) structures reveals that atomic interactions are particularly important. The calculation method for modeling a 3D structure of a complex mainly includes two strategies: free docking and template-based docking. These two methods are complementary in protein–protein docking. Therefore, integrating these two methods may improve the prediction accuracy.

Results: In this article, we compare the difference between the free docking and the template-based algorithm. Then we show the complementarity of these two methods. Based on the analysis of the calculation results, the transition point is confirmed and used to integrate two docking algorithms to develop P3DOCK. P3DOCK holds the advantages of both algorithms. The results of the three docking benchmarks show that P3DOCK is better than those two non-hybrid docking algorithms. The success rate of P3DOCK is also higher (3–20%) than state-of-the-art hybrid and non-hybrid methods. Finally, the hierarchical clustering algorithm is utilized to cluster the P3DOCK's decoys. The clustering algorithm improves the success rate of P3DOCK. For ease of use, we provide a P3DOCK webserver, which can be accessed at www.rnabinding.com/P3DOCK/P3DOCK.html. An integrated protein–RNA docking benchmark can be downloaded from http://rnabinding.com/P3DOCK/benchmark.html.

Availability and implementation: www.rnabinding.com/P3DOCK/P3DOCK.html.

Contact: liushiyong@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA-binding proteins (RBPs) bind specifically to their target RNAs in the cell. RBP–RNA interactions play an important role in post-transcriptional gene regulation to fine-tune gene expression (Gerstberger *et al.*, 2014). Therefore, studying RBP–RNA interactions can help us to understand the expression of genes. A variety of experimental techniques have been developed to study RBP–RNA interactions (Konig *et al.*, 2010; Lapointe *et al.*, 2015; Ramanathan *et al.*, 2018; Tome *et al.*, 2014; Van Nostrand *et al.*, 2016; Zarnegar *et al.*, 2016; Zhao *et al.*, 2010). Though lots of RBP–RNA interactions have been determined, atomic interaction details of them remain missing, which are the key to understand the molecular mechanisms underlying the protein–RNA recognition. The number of protein–RNA complex structure in PDB is far less than the number of protein–RNA interactions identified by RBP–RNA interactome capture techniques. To complement experimental methods and fill the gap between them, computational approaches of modeling protein–RNA complex structures are urgently needed for elucidating atomic details of RBP–RNA interactome.

The main strategy of computational modeling of protein–RNA complex structure is protein–RNA docking (Arnautova *et al.*, 2018; Guilhot-Gaudeffroy *et al.*, 2014; Huang *et al.*, 2013; Iwakiri *et al.*, 2016; Perez-Cano *et al.*, 2010; Setny and Zacharias, 2011; Tuszynska *et al.*, 2015; Yan *et al.*, 2017). To date, several protein–RNA docking methods (Arnautova *et al.*, 2018; Guilhot-Gaudeffroy *et al.*, 2014; Huang *et al.*, 2013; Iwakiri *et al.*, 2016; Perez-Cano *et al.*, 2010; Setny and Zacharias, 2011; Tuszynska *et al.*, 2015; Yan *et al.*, 2017) have been developed for the prediction of the complex structures. Most of these methods are borrowed from existing protein–protein docking methods, such as GRAMM (Katchalski-Katzir *et al.*, 1992), FTDOCK (Jackson *et al.*, 1998), ATTRACT (Zacharias, 2003), ICM (Totrov and Abagyan, 1997), ZDOCK (Mintseris *et al.*, 2007) and RosettaDock (Gray *et al.*, 2003). Previous analysis shows that the features of protein–protein interface differ significantly with the features of protein–RNA interface on atomic packing density, positively charged residue propensity, π–π stacking interactions and secondary structure states. Therefore in 2013, we proposed a novel protein–RNA docking approach 3dRPC specially designed for these protein–RNA interface features. The protocol 3dRPC includes a docking approach RPDOCK and a coarse-grained knowledge-based scoring function DECK-RP for reranking decoys. The success rates of RPDOCK are considerably higher (about 20%) than that of FTDock and GRAMM on two docking benchmarks (Huang *et al.*, 2013). In 2014, Huang *et al.* (Huang and Zou, 2014) developed a knowledge-based scoring function ITScore-PR, and reported a success rate of 46.5%, compared with 45.5% for DECK-RP (Huang *et al.*, 2013), 36.4% for DARS-RNP (Tuszynska and Bujnicki, 2011) and 27.3% for the Li potential (Li *et al.*, 2012) on a docking benchmark from Perez-Cano *et al.* (2012) if the top 10 predictions were considered. Recently, Yan *et al.* published a protein–RNA docking webserver HDOCK (Yan *et al.*, 2017) based on ITScore-PR. The success rate increased when it compared with Hdocklite (Huang and Zou, 2010) because HDOCK integrates a template-based docking method but Hdocklite is a method based only on free docking. HDOCK identified a homologous protein–RNA complex structure by HHsearch (Soding *et al.*, 2005) (for protein) and FASTA (Pearson, 1990) (for RNA) against the PDB sequence database. Then the MODELLER (Webb and Sali, 2016) was used to construct the protein model of the input sequence by using a template from a homologous protein–RNA complex. HDOCK allows for the construction of complex structures for proteins without 3D structures, which can expand the application of docking. But the ability to detect templates using HHsearch or FASTA is worse than that based on 3D structural alignment method when 3D structures are available since the 3D structure is more conserved than the sequence (Illergard *et al.*, 2009).

Besides free docking, template-based modeling approaches (Zheng *et al.*, 2016), RNP-denovo (Kappel and Das, 2019) and MD simulations (Bahadur *et al.*, 2009; Kim *et al.*, 2014; Zheng *et al.*, 2016) with experimental constrain are also proposed to predict protein–RNA complex structure. The RNP-denovo can *de novo* model the large conformational changes of RNA components with the help of experimental information, herein are limited to several cases. Nithin *et al.* reviewed the current available bioinformatics tools for protein–RNA docking, which may help users to choose appropriate tools to get the 3D structures of protein–RNA complexes (Nithin *et al.*, 2018). On the other hand, template-based modeling approaches are primarily based on an assumption that similar protein sequence may fold into similar 3D structure, which is derived from the study of the sequence–structure relationship on proteins. This assumption has been extended to protein–protein complex, which states that similar protein structure may bind in a similar way. This assumption can also be extended to protein–RNA complex reported by our team (Zheng *et al.*, 2016). A transition point (Illergard *et al.*, 2009) does exist in protein–RNA interaction system similar to proteins (Chothia and Lesk, 1986) and protein–protein interaction system (Aloy *et al.*, 2003; Kundrotas *et al.*, 2012). Based on this principle, we developed a template-based protein–RNA complex structure prediction method PRIME (Zheng *et al.*, 2016) with an accuracy of about 40% for top 1, which is much higher than the accuracy of our previously developed protein–RNA docking algorithm 3dRPC (Huang *et al.*, 2013). PRIME can predict some examples that free docking currently fails. However, the scoring function of RNA alignment algorithm SARA in PRIME is size dependent, which limits its ability to detect good templates in some cases. To enhance the RNA alignment accuracy, we developed a novel RNA 3D structural alignment approach RMalign with a size independent scoring function RMscore (Zheng *et al.*, 2019). The most recent version, PRIME2.0 (Jinfang *et al.*, 2018; Zheng *et al.*, 2019), improves the success rate about 10% than PRIME for top 1. If there is no template structure, the free docking method will be a useful complementary. The combination of these two approaches may improve the prediction accuracy of RNA–protein complexes, which has been demonstrated in protein–protein complex prediction problems (Vreven *et al.*, 2014).

Despite these advances, predicting RNA–protein complex structure remains challenging when two unbound structures are given. There is still no study integrating the template-based and free docking method in protein–RNA docking field. A real combination of free docking and template-based algorithm is needed to be developed. Here, we introduce a novel combined docking protocol P3DOCK, which is a docking webserver based on a template-based approach PRIME (version 2.1) and a template-free docking algorithm 3dRPC, to predict protein–RNA complex structure from unbound protein and RNA structure.

In this article, first, we reveal the transition point from dissimilar and similar binding model as in the previous study (Zheng *et al.*, 2016). Then, we update PRIME 2.0 to version 2.1 so that it can build multi-chains complex structures. Next, we put forward a hybrid docking method, P3DOCK, which combines PRIME2.1 and 3dRPC. P3DOCK holds the advantages of template-based docking and free docking. The success rate of top 10 of P3DOCK in the three docking benchmarks is higher than 3dRPC and PRIME 2.1, respectively. In addition, we also compare the performance of P3DOCK with state-of-the-art methods. The results demonstrate that the top 10 success rate of P3DOCK is much higher than other methods. We also show that the clustering algorithm can improve the success rate of P3DOCK. Finally, we provide a P3DOCK webserver for the interested researchers to use conveniently.

2 Materials and methods

2.1 Dataset

The co-crystal structures of protein–RNA complexes were downloaded from PDB (Berman *et al.*, 2000) (2018-04 released). To consider the functional interaction interface of protein–RNA, we used the biological assemblies in PDB. Structures with resolution greater than 3.0 angstroms (Å) were kept. The minimum length of the protein and RNA monomers were set to 30 and 20, respectively. These conditions are consistent with PRIME (Zheng *et al.*, 2016). RNA redundancy is removed by the CD-hits package (Fu *et al.*, 2012) with sequence identity 0.99 and the coverage 0.99. The parameters to remove redundancy are also consistent with PRIME. From the view of parameters, we just removed the identical RNA and the RNA with little difference (such as a single-point mutation on RNA). After these steps, we obtained 332 protein–RNA complex structures named as PDB332 (Supplementary Table S1). Since 3dRPC (Huang *et al.*, 2013) does not consider the missing/modified amino acids and nucleotides, P3DOCK also ignores them. We will determine the value of the transition point for selecting templates on this dataset. Since we only remove the redundancy with RNA sequence identity 99% and 0.99 coverage. This may create bias in determination of the transition point and testing the method. So, we remove the pairs which have an RNA sequence identity >= 60% in the determination of the transition point using needle (Rice *et al.*, 2000). The relationship between sequence and structure conservation weakens for alignments below this sequence identity (Capriotti and Marti-Renom, 2010). This dataset will be used as a template library for P3DOCK and PRIME 2.1. For benchmarking, we use three published docking benchmarks including different targets (Supplementary Fig. S1) to test our methods. To make the testing result reliable, we exclude the models built on templates with high sequence identity (60%). The relationship between the success rates of P3DOCK and the RNA sequence identity thresholds used to remove redundancy is further investigated.

2.2 Global similarity of protein–RNA complex and iRMSD

In previous researches, TM-score/RMscore refers to the similarity of monomer for that they are fitted on the monomer dataset (Jinfang *et al.*, 2018; Zhang and Skolnick, 2005). However, the assembly of protein–RNA complex usually contains multi-chains in protein and RNA, such as protein–RNA complex 1a34 (PDB ID) contains one protein chain A and two RNA chains CB. Hence, how to describe the similarity of the protein–RNA complex between 1asy (A: R) and 1a34 (A: CB) is a problem. The RNA similarity of 1asy and 1a34 can be described with the one of the RMscore between R-C (1asy chain R and 1a34 chain C), R-B or R-CB. RMscore of R-C and R-B represents the local similarity of RNA and R-CB describes the global similarity of RNA. In this study, the RMscore between 1asy chain R and 1a34 chain CB (because the order of chain C is before chain B in the 3D structure file of PDB) is used to represent the similarity of RNA in protein–RNA complex. In fact, multi-chain RNAs are considered as a single chain. Similarly, protein can be discussed in this way too, if complex contains more than two protein chains. TR-score, which is defined as the minimum of TM-score and RMscore, is used to describe the global similarity of protein–RNA complex following previous studies (Kundrotas *et al.*, 2012; Zheng *et al.*, 2016).

Interaction root-mean-square deviation (iRMSD) (Aloy *et al.*, 2003) was first introduced by Aloy *et al.* to measure the geometric difference between domain orientation in protein–protein interaction. Then we used iRMSD to measure the binding mode in protein–RNA binary complex (Zheng *et al.*, 2016). Given two binary complex P1-R1 and P2-R2, we will get 14 coordinates after superimposing P1 to P2. These 14 coordinates include the mass center of P1 and P2, and other 12 points defined as the mass center is added or subtracted 5 Å to each of the x, y and z coordinates. For example, we assume that the coordinate of mass center of P1 is (x, y, z). So, for another six points of P1, their coordinates are ($x \pm 5$, y, z), (x, $y \pm 5$, z), (x, y, $z \pm 5$). We also get other 14 coordinates after superimposing R1 to R2. Then, iRMSD is defined as the RMSD between 14 coordinates of P1-R1 and P2-R2. In this study, all the protein/RNA chains within one complex are regarded as one chain P/R. The protein is superimposed by TM-align (Zhang and Skolnick, 2005), and RNA is superimposed by RMalign (Zheng *et al.*, 2019).

2.3 Process of P3DOCK

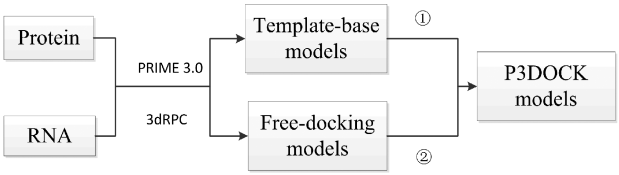
We update PRIME 2.0 to version 2.1, enabling PRIME to build the multi-chain complex. We also expand the template library. Then we integrate RPIME 2.1 and 3dRPC into P3DOCK. In Figure 1, it shows the flowchart of P3DOCK. The monomeric structures of the protein and RNA are docked by two methods: PRIME 2.1 and 3dRPC. The decoys generated by PRIME 2.1 are ranked by TR-score (minimum of RMscore and TM-score) or TM-score. The decoys generated by 3dRPC are reranked by DECK-RP (Huang *et al.*, 2013). In order to combine these two different types of decoys, we use the value of the transition point. These decoys constructing from templates are ranked at the top of P3DOCK's prediction, of whom TR-score is greater than the transition point. The top 1000 decoys generated by 3dRPC are ranked behind those generated by the template-based method.

2.4 Clustering of P3DOCK decoys

Clustering algorithm can gather similar decoys to one cluster, so that it can improve the success rate of top N by keeping a representative decoy within one cluster. In this article, we use clustering algorithm to improve the success rate of docking. For each target, P3DOCK generates 1000 decoys. The similarity matrix of each target is generated with ligand RMSD of all-to-all comparison of 1000 decoys. The average linkage hierarchical clustering algorithm is used to cluster all decoys, which minimizes the average ligand RMSD between all observations of paired clusters. For each cluster, the top-ranking decoy is chosen as the representative decoy.

2.5 Model evaluation

The quality of the model is measured by ligand RMSD, which is consistent with previous studies (Yan *et al.*, 2017; Zhao *et al.*, 2010; Zheng *et al.*, 2016). Models with ligand RMSD less than 10 Å are defined as 'acceptable' (Huang and Zou, 2014). The docking success rate of top N is defined as the number of targets on the top N models containing at least one acceptable model and then divided by the number of all targets. The success rate is used to compare the performance of different docking algorithms.



①: Template-based models with similarity score >= transposition point.

②: Top 1000 free-docking models ranked by DECK-RP.

Fig. 1. Flowchart of P3DOCK. Protein and RNA structures were docked by PRIME 2.1 and 3dRPC. The models generated by PRIME 2.1 are ranked by similarity scores. The model built on the template whose value is greater than the transition point is remained. The models generated by 3dRPC are sorted by DECK-RP. The P3DOCK's models are generated from PRIME 2.1 and 3dRPC. The models of 3dRPC are ranked behind PRIME's

3 Results

3.1 Determination of the transition point

Previous studies about protein–RNA and protein–protein docking revealed the relationship between binding patterns and monomer similarity (Kundrotas *et al.*, 2012; Zheng *et al.*, 2016). The binding modes above the transition point are similar, while the binding patterns below the transition point are random. This feature of the transition point can be used to select templates. The corrected models built by the template-based method are almost above the transition point (Zheng *et al.*, 2016). However, the transition point, which is determined based on binary complexes in the previous study (Zheng *et al.*, 2016), is not applicable any more in multi-chain complexes (this study). So, we need to redetermine the value of the transition point. We perform all-to-all alignment of 332 protein–RNA structures. The protein structure is aligned by TM-align (Zhang and Skolnick, 2005) and the RNA structure is aligned by RMalign (Zheng *et al.*, 2019). Similarity is characterized by TM-score (Zhang and Skolnick, 2004) or TR-score.

In Figure 2, it shows the results of all-to-all alignment of 332 complexes. When TM-score is used to represent the monomer similarity, the transition point occurs at around 0.35. However, the result of previous study shows that the transition point is around 0.5 (Zheng *et al.*, 2016). This difference may be explained by the different composition of the complex. A complete complex contains a larger or same size protein/RNA structure than a binary complex. This makes the value of TM-score smaller (the definition of TM-score). To illustrate whether RNA can eliminate the alternative binding mode in protein–RNA interactions, we use TR-score to represent the similarity of complex. As shown in Figure 2B, the ratio of noise (iRMSD > 5 Å, 1—accumulative fraction) is relatively lower than that in Figure 2A. This suggests that RNA similarity can reduce multiple binding patterns of high similarity proteins, which indicates that protein and RNA similarity are both required in the selection of templates. Overall, in the multi-chain complexes, there is also a transition point between the binding mode and the structural similarity. The similarity between structures which is greater than transition point indicates that they have similar binding mode.

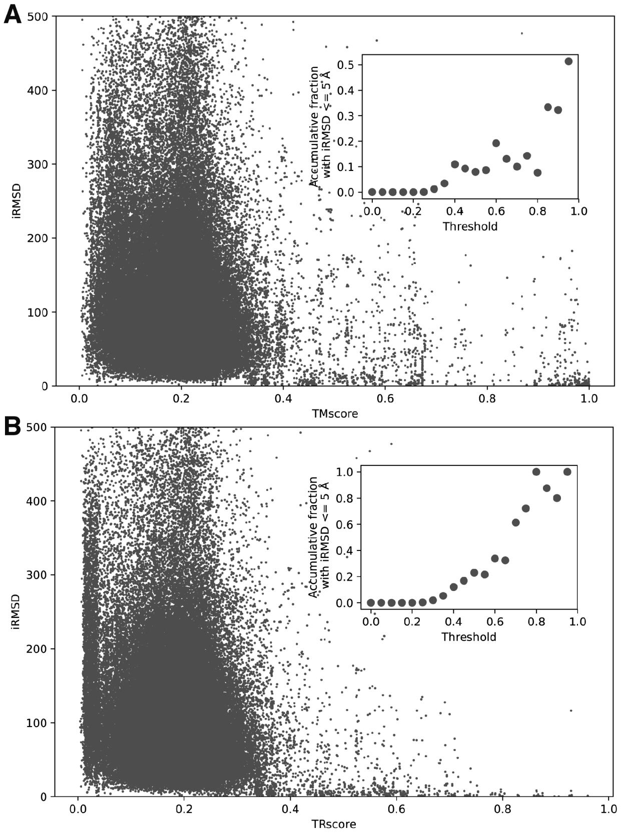


Fig. 2. The binding mode versus similarity score. The iRMSD is plotted against TM-score (**A**) and TR-score (**B**) in all-to-all alignment of 322 protein–RNA complexes. The inset shows that the value of the transition point is 0.35. For the inset, the similarity score is divided into 20 bins with the width of 0.05. Within each bin, we calculate the accumulative fraction of iRMSD <= 5 Å. The accumulative fraction of iRMSD <= 5 Å is defined that the number of pairs with the iRMSD <= 5 Å is divided by all the pairs within one bin. The transition point is defined that the similarity score threshold with which the accumulative fraction of the iRMSD <= 5 Å begins changing from 0 to non-zero value. The value of ratio of noise is equal to 1-accumulative fraction

3.2 The comparison of PRIME 2.1 and free docking

PRIME 2.0 is an algorithm for constructing protein–RNA binary complexes, which cannot be used to build complexes with multi-chains (greater than 2). So, we update PRIME 2.0 to version 2.1 and template library so that it can build the structures with multi-chains. The performance of PRIME 2.1 and the free docking algorithm is then compared on RNAbenchmark (Huang and Zou, 2013), protein–RNA docking benchmark v1.1 (Perez-Cano *et al.*, 2012) and PRDB v2 (Nithin *et al.*, 2017). The models built by PRIME 2.1 are sorted by TR-score or TM-score, while the models constructed by 3dRPC are sorted by DECK-RP (Huang *et al.*, 2013).

In Figure 3, it shows the docking results of PRIME 2.1 and 3dRPC. For top 10 predictions on three docking benchmarks, the success rates of template-based docking are all higher than that of the free docking algorithm. With more predictions, the success rates of the free docking algorithm are higher than PRIME 2.1. The curves of success rate of PRIME 2.1 are different when the models sorted by TM-score or TR-score. In RNAbenchmark (Huang and Zou, 2013) and protein–RNA docking benchmark v1.1 (Perez-Cano *et al.*, 2012), TM-score has a higher success rate than that of TR-score. In PRDB v2 (Nithin *et al.*, 2017), the success rate of TR-score is higher than the success rate of TM-score before top 7 predictions, and then the success rate of TM-score is higher than that of TR-score (a difference of 0.02 in top 10 predictions). In the previous section, we conclude that the TR-score, which takes the similarity of both protein and RNA into account, is better than considering protein alone. However, the success rate of PRIME 2.1 sorted by TM-score is higher than (0.02) that sorted by TR-score in benchmarking. This may be due to the differences in datasets used in benchmarking and determination of transition point. So, both TM-score and TR-score are used to sort the models built by PRIME 2.1. Table 1 shows the number of targets which are correctly predicted by 3dRPC and PRIME 2.1 in top 10. PRIME 2.1 (TM-score) correctly predicts 33/36/40 targets while 3dRPC correctly predicts 31/20/30 targets in PRDB v2/protein–RNA docking v1.1/ RNAbenchmark. However, the overlapping number is only 6/5/21. This shows that the free docking algorithm and the template-based docking algorithm are complementary. Supplementary Table S2 shows the success rates of top 10 of PRIME 2.1 and 3dRPC, which indicates the different performance between template-based and template-free methods. This conclusion is consistent with protein–protein complex docking (Vreven *et al.*, 2014). In Supplementary Figure S2, it shows the distribution of interface RMSD between bound and unbound structures of targets that are correctly modeled by docking methods, which indicates that template-based method performs better in medium (1.5 Å < interface RMSD <= 4 Å) and difficult targets (interface RMSD > 4 Å), and free docking method is better in easy targets (0 Å <= interface RMSD <= 1.5 Å). The *P*-value of Fisher test between the PRIME 2.1 and 3dRPC for easy targets is 0.01.

3.3 Benchmarking of P3DOCK

In the previous section, we compared the performance of PRIME 2.1 and 3dRPC. The performance of PRIME 2.1 is higher than that of 3dRPC for top 10 predictions. However, they have different characteristics. The targets which are successfully predicted by PRIME 2.1 and 3dRPC are not all overlapping. So, we develop P3DOCK to combine these two algorithms. The P3DOCK is tested on three docking benchmarks. We also compare the performance of P3DOCK with other state-of-the-art methods.

As shown in Figure 4, the top 10 success rates of P3DOCK are almost 10% higher than that of PRIME 2.1 in protein–RNA docking v1.1 and PRDB v2. In RNAbenchmark, the success rate of P3DOCK is reduced slightly. This is because P3DOCK only kept models built by templates with similarity greater than the transition point. But the fact is that the template with similarity score less than the transition point may be employed to build the correct model. Such as the target 3lrr, it can be successfully built on template 4bpb. But the TR-score between them is 0.31, which is less than the transition point. The near native decoy is excluded in this situation. So, if the value of the transition point (0.35) is used to combine PRIME 2.1 and 3dRPC, the success rate will be decreased. Regardless, the overall performance of P3DOCK is better than PRIME 2.1 or 3dRPC alone for top 10 predictions. This conclusion is consistent with previous protein–protein docking (Vreven *et al.*, 2014). The details on docking benchmark are listed in Supplementary Table S3. Since three published docking benchmarks can be accessed, which make it difficult to evaluate different algorithms. Therefore, we integrate these three docking benchmarks into one docking benchmark (Supplementary Table S4). The success rates of P3DOCK for top 10 predictions are higher than 3dRPC and PRIME2.1 by 19% and 8% in one docking benchmark (an integrated protein–RNA docking benchmark with 207 cases), respectively (see Supplementary Table S6). HDOCK can only be used to model single-chain proteins, so it cannot be compared here. The compressed file can be downloaded from http://rnabinding.com/P3DOCK/benchmark.html. In Supplementary Figure S3, it shows the distribution of interface RMSD of targets which are correctly predicted by P3DOCK on the three docking benchmarks. We also show the success rates of P3DOCK with different template–target RNA sequence identity (Supplementary Fig. S4).

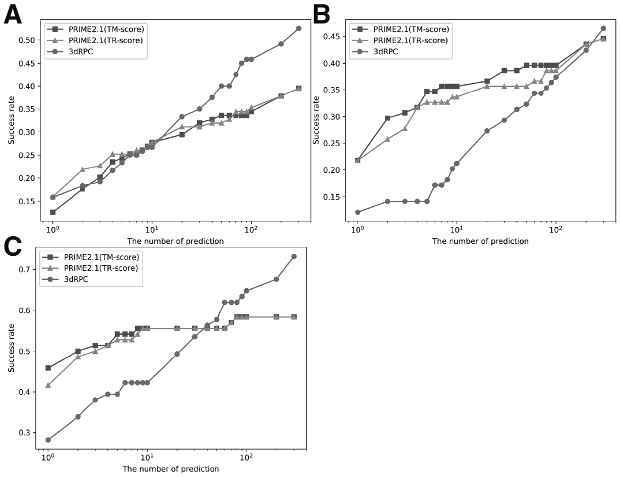


Fig. 3. Success rate of PRIME 2.1 and 3dRPC in PRDB v2 (**A**), protein–RNA docking benchmark v1.1 (**B**) and RNAbenchmark (**C**). The figure shows that PRIME 2.1 performs better than 3dRPC for top 10 predictions

We also compare the performance of P3DOCK and HDOCK on the HDOCK dataset with the same condition, in which the protein high sequence identity (>= 0.3) is removed. The HDOCK dataset only includes 33/33/25 testing cases of 126/104/72 targets from PRDBv2, protein–RNA docking benchmark v1.1 and RNAbenchmark, respectively. The success rate of P3DOCK for top 1 prediction is 6% lower than HDOCK in 33 testing cases from protein–RNA docking benchmark v 1.1. However, the success rates of P3DOCK for top 1 prediction are higher than HDOCK by 15% and 20% in 33 cases from PRDBv2.0 and 25 cases from RNAbenchmark, respectively (see Table 2). The results shown in Table 2 also indicate that P3DOCK performs 15% (Fisher test: *P*-value = 0.20), 3% (Fisher test: *P*-value = 1.0) and 20% (Fisher test: *P*-value = 0.2) better than HDOCK in terms of success rate of top 10 predictions in partial cases from PRDBv2, protein–RNA docking benchmark v1.1 and RNAbenchmark, respectively. At the same time, we also listed the results of removing the homology structures based on RNA side or on both protein and RNA sides in

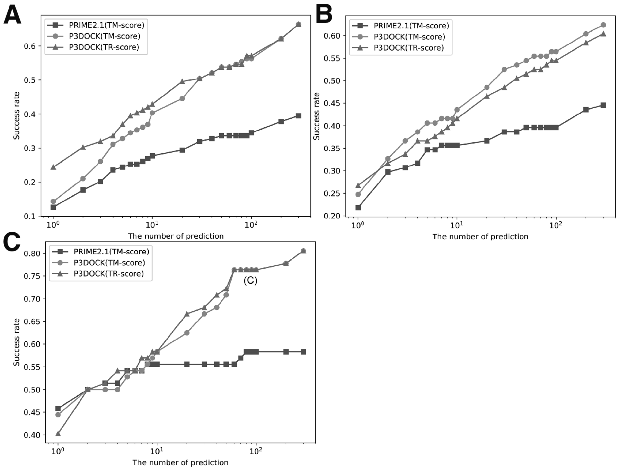


Fig. 4. The results of P3DOCK and non-hybrid docking method. The success rates of P3DOCK in PRDB v2 (**A**), protein–RNA docking benchmark v1.1 (**B**) and RNAbenchmark (**C**). This figure shows that P3DOCK has a higher success rate than free docking or template-based docking algorithm for top 10 predictions

Table 1. The number of targets which are correctly predicted by PRIME 2.1 and 3dRPC in the top 10 predictions

PRDB v2			Protein–RNA docking v1.1			RNA benchmark		
Both	3dRPC	PRIME 2.1 (TM-score)	Both	3dRPC	PRIME 2.1 (TM-score)	Both	3dRPC	PRIME 2.1 (TM-score)
6	31	33	5	20	36	21	30	40

Table 2. Comparison between P3DOCK and HDOCK in partial targets of three docking benchmarks

	PRDB v2[a]		Protein–RNA docking benchmark v 1.1[a]		RNAbenchmark[a]	
	Top 1	Top 10	Top 1	Top 10	Top 1	Top 10
P3DOCK	**0.48**	**0.70**	0.27	**0.58**	**0.72**	**0.84**
HDOCK	0.33	0.52	**0.33**	0.55	0.52	0.64

[a]Success rate of top 1 and top 10 of HDOCK testing in 33/33/25 of 126/104/72 targets, which is derived from HDOCK (Yan *et al.*, 2017). For P3DOCK, we remove the model if the protein sequence identity between the template and the target is greater than 0.3. Bold values indicate top values.

Table 3. Performance of P3DOCK in partial targets on three docking benchmarks with different conditions

	PRDB v2[a]		Protein–RNA docking benchmark v 1.1[a]		RNAbenchmark[a]		Conditions (sequence identity cutoff)	
	Top 1	Top 10	Top 1	Top 10	Top 1	Top 10	Protein	RNA
P3DOCK	0.39	0.64	0.24	0.52	0.68	0.84	0.30	0.60
P3DOCK	0.52	0.70	0.27	0.55	0.72	0.84	0.99	0.60
P3DOCK	0.48	0.70	0.27	0.58	0.72	0.84	0.30	0.99
P3DOCK	0.58	0.76	0.36	0.64	0.76	0.84	0.99	0.99

[a]Success rate of P3DOCK on the HDOCK dataset including 33/33/25 targets, which is derived from HDOCK (Yan *et al.*, 2017). Conditions represent removing homology structure based on protein sequence identity only, RNA sequence identity only or both protein and RNA sequence identity. The sequence identity cut-offs of protein and RNA are 0.30, 0.60 and 0.99, respectively.

Table 3. For real practical applications, we can use all templates except for the target itself. So, the result by using sequence identity cut-off 0.99 for both protein and RNA are also reported in Table 3. Table 4 shows a more comprehensive comparison with other docking methods in RNAbenchmark. P3DOCK/PRIME obtains the highest success rate 0.58/0.56 for top 10 predictions, which is 0.06/0.11 higher than ZDOCK-ITscore-PR/RPDock-ITscore-PR. For top 1 prediction, PRIME 2.0 and P3DOCK still achieve the best success rate with the advantage of template-based method. In addition, we compare the performance of P3DOCK and PRIME 2.0 on the unbound set in order to test the performance of building binary complexes. The result indicates that P3DOCK outperforms PRIME 2.0 (Supplementary Fig. S5). At last, we also compare the performance of P3DOCK with RNP-denovo on 10 cases provided by this method (Kappel and Das, 2019). The result is presented in Table 5. The result shows that P3DOCK can build more accurate models in some cases, but RNP-denovo can build some models successfully while P3DOCK fails.

3.4 Comparison of before and after clustering

In the previous section, we benchmark the performance of P3DOCK and compare it to the other 11 methods. P3DOCK achieves the highest success rate of top 1 and top 10 predictions in RNAbenchmark. However, some near-native decoys generated by 3dRPC or PRIME 2.1 cannot be picked out (Fig. 4). The clustering algorithm can rank the near-native decoys to top in protein–protein docking (Kozakov *et al.*, 2005). Therefore, in this section, we cluster the decoys generated by P3DOCK (both PRIME's and 3dRPC's) to further improve the success rate of top 10 of P3DOCK.

Different ligand RMSD cut-offs are used to cluster decoys. But different cut-offs have little effect on the final success rates (Supplementary Fig. S5). Therefore, 5 Å is selected as the cut-off finally. The results of the clustering are shown in Figure 5 and Supplementary Table S5. In Figure 5, it shows that the success rates after clustering is higher than that without clustering in three protein–RNA docking benchmarks. After clustering, the success rates of top 10 are increased by 4%, 3% and 3%, respectively (Supplementary Table S5). An example shows that the near-native decoy of target 3ol9

Table 4. Performance of docking methods on RNAbenchmark (all targets)

Docking protocol	Top 10	Top 1
NPDock[a]	0.29 (top 3)	
P3DOCK	**0.58**	**0.40**
PRIME2.1	**0.56**	**0.41**
RPDock-DECK-RP[b]	0.38	0.28
RPDock-3dRPC-score[b]	0.42	0.32
RPDock-ITscore-PR[b]	0.45	0.32
ZDOCK-ITscore-PR[b]	0.52	0.36
ZDOCK-3dRPC-score[b]	0.44	0.30
ZDOCK-DECK-RP[b]	0.40	0.25
ICM[c]	0.48	–

Note: – stands for not implemented. Bold values indicate top values.

[a]Data are derived from Tuszynska *et al.* (2015) which only provides success rate of top 3.

[b]Data are extracted from Li *et al.* (2017) and the success rate is calculated on partial targets. So, we recalculate it with all targets.

[c]Data are derived from Arnautova *et al.* (2018).

Table 5. The comparison of P3DOCK and RNP-denovo in 10 cases

PDB ID	Best RMSD of top 100 scoring models (Å), RMSD of best scoring model (Å)			
	RNP-denovo[a]		P3DOCK	
1B7F	4.2	8.9	7.3	12.3
1JBS	3	3.2	1.2	1.2
1WPU	3.9	10.1	0.2	0.2
1WSU	2.4	2.8	1.6	6.3
2ASB	3.1	4	1.8	1.8
2BH2	5.8	6.8	1.8	49.4
2QUX	4.7	5.5	3.5	23.1
3BX2	3.8	4.3	5.3	19
1P6V	6.3	8.9	25.4	36
1DFU	5.5	9.1	28.6	56.2

[a]Data are derived from RNP-denovo (Kappel and Das, 2019).

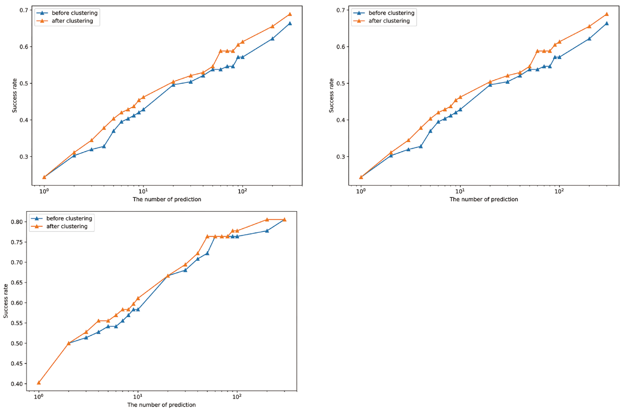
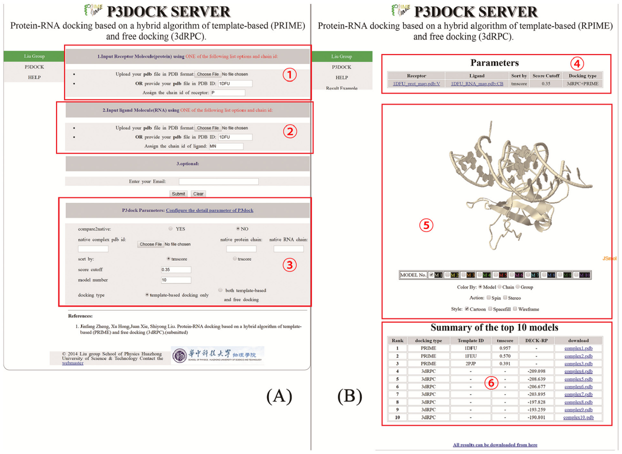


Fig. 5. The results of P3DOCK before and after clustering in three docking benchmarks. The success rates of P3DOCK in PRDB v2 (**A**), protein–RNA docking benchmark v1.1 (**B**) and RNAbenchmark (**C**). Curve labeled with 'before clustering' stands for success rate of before clustering. Curve labeled with 'after clustering' stands for success rate of after clustering with ligand RMSD = 5 Å as cut-off. This figure shows that the success rates of top 10 were improved 3–4% in three docking benchmarks



P3DOCK SERVER

Protein-RNA docking based on a hybrid algorithm of template-based (PRIME) and free docking (3dRPC).

1.Input Receptor Molecule(protein) using ONE of the following list options and chain id:

- Upload your **pdb** file in PDB format: No file chosen
- **OR** provide your **pdb** file in PDB ID: 1DFU

Assign the chain id of receptor: P

2.Input ligand Molecule(RNA) using ONE of the following list options and chain id:

- Upload your **pdb** file in PDB format: No file chosen
- **OR** provide your **pdb** file in PDB ID: 1DFU

Assign the chain id of ligand: MN

3.optional:

Enter your Email:

P3dock Parameters: Configure the detail parameter of P3dock

compare2native: ○ YES ● NO

native complex pdb id: No file chosen native protein chain: native RNA chain:

sort by: ● tmscore ○ trscore

score cutoff 0.35

model number 10

docking type ● template-based docking only ○ both template-based and free docking

References:

1. Jinfang Zheng, Xu Hong,Juan Xie, Shiyong Liu. Protein-RNA docking based on a hybrid algorithm of template-based (PRIME) and free docking (3dRPC).(submitted)

© 2014 Liu group School of Physics Huazhong University of Science & Technology Contact the webmaster

華中科技大學 物理學院 SCHOOL OF PHYSICS . HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

P3DOCK SERVER

Protein-RNA docking based on a hybrid algorithm of template-based (RPIME) and free docking (3dRPC).

Parameters

Receptor	Ligand	Sort by	Score Cutoff	Docking type
1DFU_prot_map.pdb:V	1DFU_RNA_map.pdb:CB	tmscore	0.35	3dRPC+PRIME

MODEL No. ☑ M 1 □ M 2 □ M 3 □ M 4 □ M 5 □ M 6 □ M 7 □ M 8 □ M 9 □ M 10

Color By: ● Model ○ Chain ○ Group

Action: □ Spin □ Stereo

Style: ☑ Cartoon □ Spacefill □ Wireframe

Summary of the top 10 models

Rank	docking type	Template ID	tmscore	DECK-RP	download
1	PRIME	1DFU	0.957	-	complex1.pdb
2	PRIME	1FEU	0.570	-	complex2.pdb
3	PRIME	2PJP	0.391	-	complex3.pdb
4	3dRPC	-	-	-209.098	complex4.pdb
5	3dRPC	-	-	-208.639	complex5.pdb
6	3dRPC	-	-	-206.677	complex6.pdb
7	3dRPC	-	-	-203.895	complex7.pdb
8	3dRPC	-	-	-197.828	complex8.pdb
9	3dRPC	-	-	-193.259	complex9.pdb
10	3dRPC	-	-	-190.801	complex10.pdb

All results can be downloaded from here

Fig. 6. Schematic diagram of the P3DOCK Webserver. (**A**) shows the page for parameters setting, and (**B**) shows the page of the docking results. A more detailed introduction can be found in the main text

(RNAbenchmark) is reranked to top 9 after clustering, while it is ranked at top 17 before clustering. Overall, the clustering algorithm improves the success rate of P3DOCK.

3.5 Webserver of P3DOCK

For the convenience of other researchers, we provide P3DOCK webserver. As shown in Figure 6, it is divided into two parts:

1. Input and parameters configuration
   ①: Upload the PDB file of the protein or provide the PDB ID. At the same time, the chain ID must be specified.
   ②: Upload the PDB file of the RNA or provide the PDB ID. The chain ID of RNA must be assigned.
   ③: The parameters of P3DOCK can be set in detail. Of course, users can also use the default parameters.
2. Results section
   ④: Parameters that user set.
   ⑤: Ten best models of P3DOCK showed by JSmol (Hanson and Lu, 2017).
   ⑥: Summary tables with top 10 models. If the model is built by PRIME 2.1, the PDB ID of the template and the similarity scores of the target and template are displayed. If the model is built by 3dRPC, the table will give the result of DECK-RP.

4 Discussion and conclusion

We updated PRIME 2.0 to version 2.1 and it can build a complete protein–RNA structure instead of a binary complex. The performance of PRIME 2.1 and 3dRPC was compared on three docking benchmarks. The results show that the template-based approach is better than the docking-based approach. Like protein–protein docking, free docking and template-based docking have their own advantages in protein–RNA docking. In other words, they are complementary. Therefore, we developed P3DOCK by combining free docking and template-based algorithm. We systematically benchmarked P3DOCK on three docking benchmarks and found that the success rates of P3DOCK are higher than that of free docking or template-based algorithms for top 10 predictions alone. P3DOCK obtains the highest success rate in comparing to other state-of-the-art methods for top 10 predictions. We also discuss the ability of clustering algorithms to pick up the near-native decoys of P3DOCK. The success rate can be improved by introducing the clustering process to the P3DOCK. Finally, we provide the P3DOCK webserver, which is convenient for researchers who need it.

Funding

This work was supported by National Natural Science Foundation of China [31100522]; National High Technology Research and Development Program of China [2012AA020402]; the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) [U1501501] and the Fundamental Research Funds for the Central Universities [2016YXMS017].

Conflict of Interest: none declared.

References

Aloy,P. *et al.* (2003) The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, **332**, 989–998.

Arnautova,Y.A. *et al.* (2018) Protein-RNA docking using ICM. *J. Chem. Theory Comput.*, **14**, 4971–4984.

Bahadur,R.P. *et al.* (2009) Binding of the bacteriophage P22 N-peptide to the boxB RNA motif studied by molecular dynamics simulations. *Biophys. J.*, **97**, 3139–3149.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Capriotti,E. and Marti-Renom,M.-R. (2010) Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, **11**, 322.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *Embo J.*, **5**, 823–826.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Gerstberger,S. *et al.* (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.

Gray,J.J. *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.

Guilhot-Gaudeffroy,A. *et al.* (2014) Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities. *PLoS One*, **9**, e108928.

Hanson,R.M. and Lu,X.J. (2017) DSSR-enhanced visualization of nucleic acid structures in Jmol. *Nucleic Acids Res.*, **45**, W528–W533.

Huang,S.Y. and Zou,X. (2010) MDockPP: a hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. *Proteins*, **78**, 3096–3103.

Huang,S.-Y. and Zou,X. (2013) A non-redundant structure dataset for benchmarking protein-RNA computational docking. *J. Comput. Chem.*, **34**, 311–318.

Huang,S.Y. and Zou,X. (2014) A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Res.*, **42**, e55.

Huang,Y. *et al.* (2013) A novel protocol for three-dimensional structure prediction of RNA-protein complexes. *Sci. Rep.*, **3**, Article Number 1887.

Illergard,K. *et al.* (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins*, 77, 499–508.

Iwakiri,J. *et al.* (2016) 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**, 4688–4697.

Jackson,R.M. *et al.* (1998) Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.*, **276**, 265–285.

Kappel,K. and Das,R. (2019) Sampling native-like structures of RNA-protein complexes through Rosetta folding and docking. *Structure*, **27**, 140–151.e145.

Katchalski-Katzir,E. *et al.* (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA*, **89**, 2195–2199.

Kim,H. *et al.* (2014) Protein-guided RNA dynamics during early ribosome assembly. *Nature*, **506**, 334–338.

Konig,J. *et al.* (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–U166.

Kozakov,D. *et al.* (2005) Optimal clustering for detecting near-native conformations in protein docking. *Biophys. J.*, **89**, 867–875.

Kundrotas,P.J. *et al.* (2012) Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. USA*, **109**, 9438–9441.

Lapointe,C.P. *et al.* (2015) Protein-RNA networks revealed through covalent RNA marks. *Nat. Methods*, **12**, 1163–1170.

Li,C.H. *et al.* (2012) A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. *Proteins*, **80**, 14–24.

Li,H. *et al.* (2017) A pair-conformation-dependent scoring function for evaluating 3D RNA-protein complex structures. *PLoS One*, **12**, e0174662.

Mintseris,J. *et al.* (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins*, **69**, 511–520.

Nithin,C. *et al.* (2017) A non-redundant protein-RNA docking benchmark version 2.0. *Proteins Struct. Funct. Bioinform.*, **85**, 256–267.

Nithin,C. *et al.* (2018) Bioinformatics tools and benchmarks for computational docking and 3D structure prediction of RNA-protein complexes. *Genes*, **9**, pii: E432.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.

Perez-Cano,L. *et al.* (2010) Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. *Pac. Symp. Biocomput.*, 293–301.

Perez-Cano,L. *et al.* (2012) A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. *Proteins*, **80**, 1872–1882.

Ramanathan,M. *et al.* (2018) RNA-protein interaction detection in living cells. *Nat. Methods*, **15**, 207–212.

Rice,P. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet*, **16**, 276–277.

Setny,P. and Zacharias,M. (2011) A coarse-grained force field for Protein-RNA docking. *Nucleic Acids Res.*, **39**, 9118–9129.

Soding,J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–248.

Tome,J.M. *et al.* (2014) Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nat. Methods*, **11**, 683–688.

Totrov,M. and Abagyan,R. (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins*, **Suppl. 1**, 215–220.

Tuszynska,I. and Bujnicki,J.M. (2011) DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. *BMC Bioinformatics*, **12**, 348.

Tuszynska,I. *et al.* (2015) NPDock: a web server for protein-nucleic acid docking. *Nucleic Acids Res.*, **43**, W425–430.

Van Nostrand,E.L. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.

Vreven,T. *et al.* (2014) Evaluating template-based and template-free protein--protein complex structure prediction. *Brief Bioinform.*, **15**, 169–176.

Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinform.*, **54**, 5 6 1–5 6 37.

Yan,Y. *et al.* (2017) HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.*, **45**, W365–W373.

Zacharias,M. (2003) Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.*, **12**, 1271–1282.

Zarnegar,B.J. *et al.* (2016) irCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods*, **13**, 489.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.

Zhao,J. *et al.* (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–953.

Zheng,J. *et al.* (2016) Template-based modeling of protein-RNA interactions. *PLoS Comput. Biol.*, **12**, e1005120.

Zheng,J. *et al.* (2019) RMalign: an RNA structural alignment tool based on a novel scoring function RMscore. *BMC Genomics*, **20**, 276.