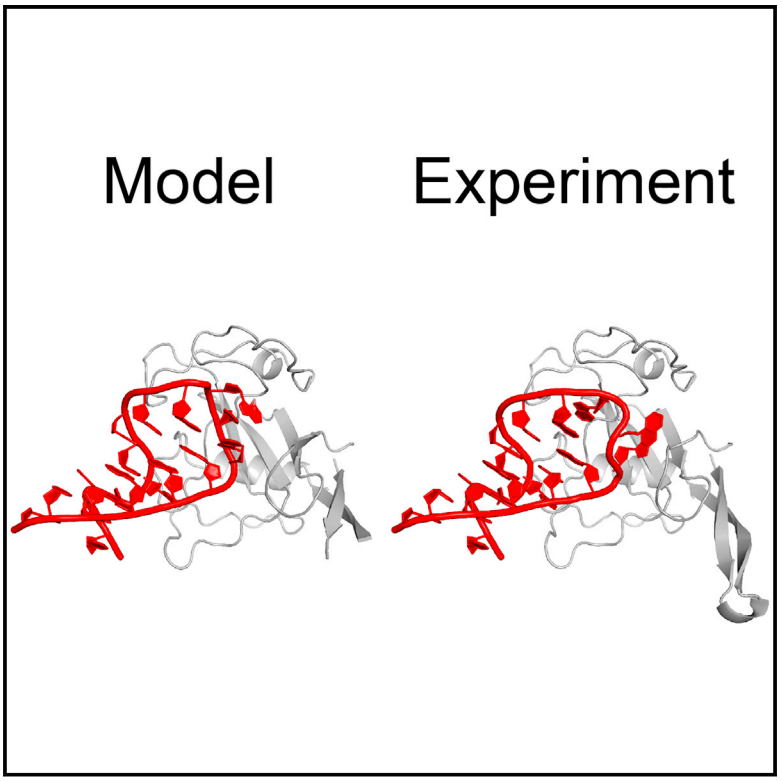
Resource

Structure

Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking

Graphical Abstract



Authors

Kalli Kappel, Rhiju Das

Correspondence

rhiju@stanford.edu

In Brief

RNA and protein molecules interact intimately during ribosome translation, pre-mRNA splicing, and other fundamental cellular processes. Structures of the relevant RNA-protein complexes often elude high-resolution determination. Kappel and Das report a computational Rosetta method *RNP-denovo* that can incorporate sparse experimental data and sample native-like structures of RNA-protein complexes.

Highlights

- A Rosetta method, *RNP-denovo*, folds and docks RNA to protein surfaces
- *RNP-denovo* consistently samples native-like structures for a diverse benchmark set
- Applying *RNP-denovo* to past modeling problems gives models with improved accuracy

Kappel & Das, 2019, Structure *27*, 1–12
January 2, 2019
https://doi.org/10.1016/j.str.2018.10.001

CellPress

Please cite this article in press as: Kappel and Das, Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking, Structure (2018), https://doi.org/10.1016/j.str.2018.10.001

Structure
Resource

Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking

Kalli Kappel[1] and Rhiju Das[1,2,3,4,*]
[1]Biophysics Program, Stanford University, Stanford, CA 94305, USA
[2]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Department of Physics, Stanford University, Stanford, CA 94305, USA
[4]Lead Contact
*Correspondence: rhiju@stanford.edu
https://doi.org/10.1016/j.str.2018.10.001

SUMMARY

RNA-protein complexes underlie numerous cellular processes including translation, splicing, and post-transcriptional regulation of gene expression. The structures of these complexes are crucial to their functions but often elude high-resolution structure determination. Computational methods are needed that can integrate low-resolution data for RNA-protein complexes while modeling *de novo* the large conformational changes of RNA components upon complex formation. To address this challenge, we describe *RNP-denovo*, a Rosetta method to simultaneously fold-and-dock RNA to a protein surface. On a benchmark set of diverse RNA-protein complexes not solvable with prior strategies, *RNP-denovo* consistently sampled native-like structures with better than nucleotide resolution. We revisited three past blind modeling challenges involving the spliceosome, telomerase, and a methyltransferase-ribosomal RNA complex in which previous methods gave poor results. When coupled with the same sparse FRET, crosslinking, and functional data used previously, *RNP-denovo* gave models with significantly improved accuracy. These results open a route to modeling global folds of RNA-protein complexes from low-resolution data.

INTRODUCTION

RNA-protein interactions underlie many critical cellular processes from translation, splicing, and telomere extension to regulation of mRNA stability, alternative splicing, and subcellular localization (Gerstberger et al., 2014; Mitchell and Parker, 2014). Many of these roles require the formation of intricate three-dimensional structures, but the structural heterogeneity and transient nature of many RNP states render them invisible to all but the lowest-resolution methods, such as fluorescence resonance energy transfer (FRET), crosslinking, and mutagenesis tests. For such states, computational techniques will be needed for "hybrid" structure determination integrating low-resolution data into structural models (Schlundt et al., 2017; Ward et al., 2013). Such strategies have proved useful for RNA and proteins separately (Miao et al., 2017; Moult et al., 2018; Ward et al., 2013; Weinreb et al., 2016), but they are not yet in widespread use for RNA-protein complexes because the necessary computational structure prediction methods have not yet been developed.

The majority of existing structure prediction tools for RNA-protein complexes focus on rigid-body docking of RNA and protein partners (Tuszynska et al., 2014). These methods have achieved impressive success when predicting structures of complexes from the corresponding bound RNA and protein structures (Huang and Zou, 2014; Huang et al., 2013; Li et al., 2012; Setny and Zacharias, 2011). However, they typically perform poorly in the more realistic case of starting from the unbound RNA and protein structures (Lensink and Wodak, 2010). This limitation is largely due to the flexibility and conformational variability of RNA; protein-bound RNA structures often differ considerably from the corresponding unbound RNA structures (Rau et al., 2012).

To help address this challenge, a fragment-based method for predicting structures of single-stranded RNA bound to proteins was recently developed (Chauvot de Beauchene et al., 2016a, 2016b). Starting from a protein structure and the positions of a few anchor RNA nucleotides, this method was able to predict structures of RNA recognition motif and Puf proteins bound to single-stranded RNA with high accuracy. However, this method neglects intramolecular RNA interactions and assumes that every RNA nucleotide interacts with the protein, making it applicable to only a small subset of RNA-protein complexes. Currently, there are no computational tools that can model arbitrary protein-bound RNA structures *de novo*, although in principle this could be accomplished by combining RNA structure prediction ("folding") and RNA-protein docking methods sequentially in a "prefold-then-dock" strategy.

Here, we present tests of this prefold-then-dock strategy on a benchmark set of ten diverse RNA-protein complexes and find that it does not lead to accurate models, suggesting that a different modeling approach is needed. We then describe a method, *RNP-denovo*, to model RNA-protein complexes by simultaneously folding and docking RNA to a protein surface. This fold-and-dock approach is implemented in Rosetta and combines the FARNA method for RNA folding (Cheng et al.,

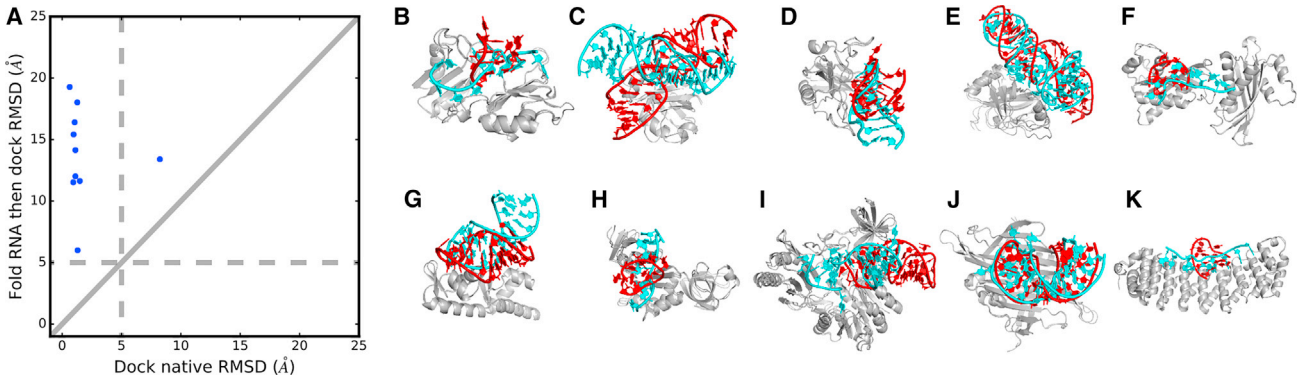


Figure 1. Tests of the Best Accuracies Achievable by a Prefold-then-Dock Strategy

(A) The best RMSDs achieved for ten systems for the prefold-then-dock strategy versus the (unrealistic) best case of rigid-body docking of the native RNA to the native protein structures.

(B–K) Native structures (RNA in cyan) overlaid with the best models from the prefold-then-dock method (RNA in red, protein in gray) for (B) PDB: 1B7F, (C) PDB: 1DFU, (D) PDB: 1JBS, (E) PDB: 1P6V, (F) PDB: 1WPU, (G) PDB: 1WSU, (H) PDB: 2ASB, (I) PDB: 2BH2, (J) PDB: 2QUX, and (K) PDB: 3BX2

See also Figure S1.

2015) with RNA-protein binding and a statistical RNA-protein score function. For the benchmark set of ten RNA-protein complexes, starting from the unbound protein structure and a few RNA residues fixed relative to the protein (to simulate sparse experimental data), *RNP-denovo* recovered native-like models with an average root-mean-square deviation (RMSD) over the best models of 4.3 Å, which is comparable with what has previously been achieved for low-resolution protein and RNA structure prediction on similar size problems (Das and Baker, 2007; Simons et al., 1997). Additional tests demonstrate the importance of including both RNA-protein and intramolecular RNA interactions when modeling protein-bound RNA structures. Finally, we apply our *RNP-denovo* fold-and-dock method to previous structure modeling challenges based on limited experimental data for the human telomerase core RNP, an RNA methyltransferase, and the human spliceosomal C complex active site. We find improvements over previous blind models in all cases, and achieve the correct global folds in the telomerase and RNA methyltransferase cases. These results demonstrate the applicability of this method to real modeling challenges and highlight areas for future improvement. Accurate fully *de novo* prediction of protein-bound RNA structures is not yet feasible, but we expect the method described here to be immediately useful for modeling arbitrary RNA-protein complexes in cases where sparse experimental data are available.

RESULTS

Testing a Prefold-then-Dock Approach

To evaluate different protocols for modeling RNA-protein complexes, we built models for a set of ten RNA-protein complexes for which crystal structures are available. The benchmark set contains relatively small RNA-protein complexes with between 94 and 417 protein residues and between 7 and 45 RNA nucleotides per system. This size range is typical for initial structure prediction tests for RNA alone, proteins alone, and protein-protein docking (Das and Baker, 2007; Gray et al., 2003; Simons et al., 1997); tests on larger RNA-protein complexes are described below. Given previous results for low-resolution modeling of RNA and proteins separately, the target modeling accuracy for systems in this size range is around 2–7 Å RMSD (Das and Baker, 2007; Simons et al., 1997). Because the methods considered here do not include final full-atom refinement, we focused on evaluating whether native-like conformations are sampled. This is important because current high-resolution refinement methods typically do not modify structures dramatically. We therefore report the best RMSD accuracy of the top 100 scoring models (out of thousands) in all cases. This procedure is also consistent with typical evaluation criteria in structure prediction challenges, where multiple models are often considered and the number of models is increased to assess progress on more difficult problems with large search spaces (Lensink et al., 2017; Miao et al., 2017; Miao and Westhof, 2017; Moult et al., 2018).

To first address whether a combination of existing tools might be sufficient to predict structures of protein-bound RNA, we tested a prefold-then-dock strategy on a set of ten RNA-protein complexes using the FARFAR method (Cheng et al., 2015) to fold the RNA and then RPDock to dock the resulting RNA structures (Huang et al., 2013). We first tested whether RPDock could accurately predict structures of RNA-protein complexes if starting from the bound protein structure and bound RNA structure, i.e., if the RNA structure could somehow be predicted perfectly. Indeed, RPDock recovered near-native models with RMSDs $\leq$1.5 Å for the best of the top 100 scoring models in nine out of ten cases, with a mean RMSD of 1.8 Å over all ten cases. However, if we use the unbound protein structures, as is more realistic for macromolecule docking, the mean RMSD overall increases to 7.1 Å. Furthermore, the results became significantly worse if we assume that the bound RNA structures are unknown, as is typically the case in realistic modeling scenarios. For these tests, we folded the RNA with FARFAR and clustered the resulting RNA structures, retaining the centers of the ten most populated clusters for subsequent docking in the hopes of capturing some of the conformational heterogeneity of the unbound RNA and including conformations similar to the bound structures. After docking these structures with RPDock, the mean RMSD increased to 13.8 Å, with RMSDs worse than 11 Å in nine out of ten cases (Figures 1A–1K and S1A, see the

Table 1. Accuracy of Prefold-then-Dock versus Rosetta *RNP-denovo* Fold-and-Dock

System	Residue Numbers of Fixed Nucleotides	Residue Numbers of All Nucleotides Modeled	Best RMSD of Top 100 Scoring Models (Å), RMSD of Best Scoring Model (Å)	
			Prefold-then-Dock[a]	Rosetta *RNP-denovo* Fold-and-Dock
Sex-lethal RRM (PDB: 1B7F)	3, 12	3–12	6.7, 10.6	4.2, 8.9
Ribotoxin restrictocin – SRL analog (PDB: 1JBS)	7, 23	7–23	3.1, 5.4	3.0, 3.2
HutP antitermination complex (PDB: 1WPU)	1, 7	1–7	7.3, 8.1	3.9, 10.1
mRNA binding domain of SelB elongation factor (PDB: 1WSU)	13–15, 33–35	13–35	2.4, 7.8	2.4, 2.8
NusA transcriptional regulator (PDB: 2ASB)	1, 10	1–10	5.6, 9.1	3.1, 4.0
Methyltransferase RumA in complex with rRNA (PDB: 2BH2)	834, 847–849, 861–863	834–863	6.5, 8.4	5.8, 6.8
PP7 coat protein and viral RNA (PDB: 2QUX)	245–247, 267–269	245–269	5.2, 7.5	4.7, 5.5
Puf4 bound to 3′ UTR of target transcript (PDB: 3BX2)	652, 660	652–660	6.3, 11.2	3.8, 4.3
SmpB-tmRNA complex (PDB: 1P6V)	19–20, 37–38, 41–42, 56–57	13–57	5.2, 6.7	6.3, 8.9
E. coli L25-5S rRNA (PDB: 1DFU)	83–85, 91–93	70–86, 91–106	5.2, 7.4	5.5, 9.1
Human telomerase[b]	46–56, 301–305, 311–315, DNA: 1–11	33–145, 170–191, 301–305, 311–315, DNA: 1–11	51.4 (25.5), 147.0[c]	15.0, 88.6
RNA methyltransferase (CAPRI T33)	N/A	1–74	18.2 (31.0), 20.7[d]	13.6, 35.9
Human spliceosome	N/A	U5: 30–52, U2: 20–49, intron: BP-5 – BP+1, intron: +1 – +10, U6: 30–78, 5′ exon: -3 to -1	13.8 (34.5), 20.1[e]	8.0, 19.5
Average	–	–	10.5, 20.8	6.1, 16.0

[a]For the ten small RNA-protein systems, the RNA was folded keeping the same residues fixed relative to the protein as in the *RNP-denovo* fold-and-dock runs (therefore, docking was not performed; see Figures 1 and S1 for full prefold-then-dock results).
[b]Results for models built without FRET data. RMSDs calculated over pseudoknot ends (see the STAR Methods).
[c]Because the position of the template RNA relative to the protein is known, prefold-then-dock was not attempted. Instead, the results here are for the *RNP-denovo* fold-and-dock runs with just RNA-only score terms and RNA-protein sterics. Accuracy of best of top 100 previous blind models shown in parentheses.
[d]Accuracy of the best of the previously submitted blind models shown in parentheses.
[e]Accuracy of the previously published blind model shown in parentheses. RMSDs were calculated over active site residues (see the STAR Methods).

STAR Methods for complete details). Indeed, even assuming the best possible docking—by aligning predicted RNA structures to the native RNA coordinates—the mean RMSD remained at 6.5 Å, with RMSDs >7 Å in five out of ten cases (Figure S1A). We emphasize that the RMSD values here and below are computed for the best of 100 models and therefore represent "best-case" assessments. The poor results suggest that sequentially folding and docking RNA structures does not generally lead to accurate models of RNA-protein complexes.

However, in many realistic cases, the structure prediction problem may be simplified by the availability of limited experimental data. In favorable cases, these data can elucidate a few specific RNA-protein contacts and/or relative orientations of the RNA and protein partners. To test the prefold-then-dock strategy in such scenarios, we simulated the availability of limited experimental data by assuming the bound conformations of the 3′ and 5′ nucleotides for the single-stranded RNA-protein complexes (analogous to previous work; Chauvot de Beauchene et al., 2016a, 2016b), or the positions of RNA helices relative to the protein for the remaining complexes (see the STAR Methods for complete details). When the RNA was folded with these constraints, the mean RMSD was 5.4 Å over the ten RNA-protein complexes, representing an improvement over the naive prefold-then-dock tests, but still far from the 1.8 Å mean RMSD achieved when docking the bound RNA conformations (Table 1; Figure S1B). Together these results suggest that a prefold-then-dock strategy alone is insufficient to recover near-native conformations of RNA-protein complexes, and that a strategy that allows simultaneous optimization of the RNA fold and the docked conformation might improve predictions.

Developing a Fold-and-Dock Method

Motivated by these results, we hypothesize that folding the RNA in the context of the protein rather than pre-folding and docking would further improve the accuracy of computational models. We developed a fold-and-dock algorithm *RNP-denovo*

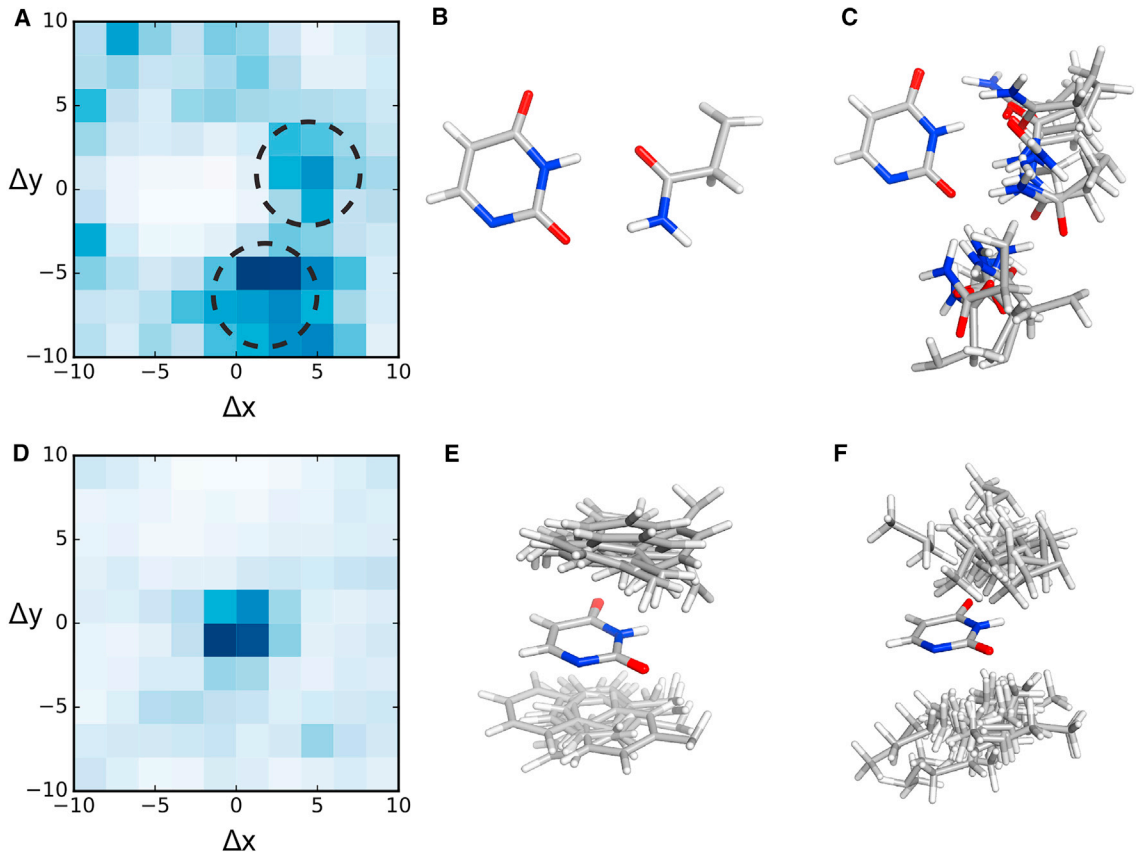


Figure 2. Statistical RNA-Protein Potential in Rosetta

(A) The distribution of glutamine side-chain centroids around uracil in the plane of the base ($0 < |z| < 3$ Å), from the non-redundant set of RNA-protein crystal structures from the PDB (darker blue represents higher frequency). Two major hotspots are circled. Distributions of all protein side chains around all four RNA bases are shown in Figures S2A–S2D.

(B) A pseudo-pair between glutamine and uracil.

(C) Conformations from the two major hotspots circled in (A) show that the interactions between glutamine and uracil are not highly stereotyped.

(D) The distribution of phenylalanine side-chain centroids around uracil above and below the plane of the base ($3 < |z| < 6.5$ Å; darker blue represents higher frequency). Distributions of all protein side chains around all four RNA bases are shown in Figures S2E–S2H.

(E) Representative conformations from the hotspot in (D) show stereotyped stacking interactions.

(F) Conformations of valine around uracil also reveal frequent stacking interactions.

See also Figures S2–S4.

in Rosetta by modifying the FARNA algorithm for RNA folding to include RNA-protein docking moves and to take into account RNA-protein interactions. Currently, *RNP-denovo* does not include full-atom refinement; as described above, we wish to cleanly test here whether native-like conformations are sampled with a low-resolution protocol. We also developed an RNA-protein statistical potential to score conformations of RNA as it makes contact with a protein surface. While prior studies have developed statistical potentials for RNA-protein docking, we sought a scoring function that could be rapidly computed at the same time as Rosetta RNA score terms and have a similar level of coarse graining. Overall, our scoring function includes all previously published score terms describing RNA structure (Das and Baker, 2007), and additional terms describing interactions between RNA and proteins. As with the Rosetta RNA statistical potential, we took a coarse-grained knowledge-based approach. Score terms were based on the frequencies of interactions observed in a non-redundant set of 154 crystal structures of RNA-protein complexes with resolution better than 2.5 Å, curated from the PDB. Structures of DNA-protein complexes were not included in this set because of the differences between DNA and RNA backbone conformational preferences.

The score terms describe three major features of RNA-protein interactions. First, pseudo pairs between nucleotides and protein side chains have been observed and are thought to contribute to both the specificity and affinity of RNA-protein interactions (Kondo and Westhof, 2011). To capture these effects, we developed a potential based on the distributions of protein side-chain centroids in the plane of each of the four RNA bases. As described previously for the RNA score terms (Das and Baker, 2007), a coordinate system was set up on each base with the origin at the centroid of the base heavy atoms, the x axis going through the N1 atom for purines or the N3 atom for pyrimidines, and the z axis perpendicular to the plane of the base. Analysis of the protein side-chain distributions indeed revealed positional preferences within the plane of the base (Figures 2A and S2A),

although the interactions were not as highly stereotyped as, for example, RNA base pairs (Figures 2B and 2C). Statistical potentials were derived by taking the negative logarithm of these frequencies (see the STAR Methods). These terms include RNA-protein pseudo pairs previously identified by expert inspection (Kondo and Westhof, 2011) as well as less stereotyped RNA-protein contacts.

Second, the potential captures the effect of stacking interactions frequently found at RNA-protein interfaces (Rahman et al., 2015). Analysis of the distributions of protein side chains above and below the plane of the base revealed that the aromatic amino acids tryptophan, tyrosine, and phenylalanine, and two additional hydrophobic amino acids, valine and leucine, frequently stack on RNA bases (Figures 2D–2F and S2B). A stacking bonus is encoded in the potential for any of these five side chains with 3.0 Å < |z| < 6.5 Å and $\sqrt{x^2+y^2} < 4.0$ Å, again analogous to the bonus for RNA-RNA stacking.

Third, the potential includes the effect of interactions with the RNA backbone, which often confer both affinity and structural specificity to RNA-protein interactions (Iwakiri et al., 2012). Scores were inferred by taking the negative logarithm of the frequencies of distances between RNA phosphate atoms and protein side-chain centroids (see the STAR Methods). The resulting statistical potentials exhibit several expected features (Figure S3A), most notably that interactions with positively charged amino acids (arginine and lysine) are among those that score most favorably, and interactions with polar amino acids are generally preferred over interactions with nonpolar amino acids.

To account for additional interactions that may be missed by the three score terms described above, we also included a general distance-dependent potential based on the observed distributions of distances between representative RNA and protein atoms (Figure S3B), analogous to several prior efforts (Guilhot-Gaudeffroy et al., 2014; Huang and Zou, 2014; Li et al., 2012; Perez-Cano et al., 2010; Setny and Zacharias, 2011; Simons et al., 1997; Tuszynska and Bujnicki, 2011; Zheng et al., 2007). Finally, steric clashes between the RNA and protein are penalized in a manner similar to clashes within the RNA (Das and Baker, 2007). A penalty is applied when representative RNA and protein atoms come within a distance smaller than the minimum distance observed in the set of crystal structures from the PDB (Figure S4).

Benchmarking *RNP-denovo* on Ten RNA-Protein Complexes

We benchmarked the performance of the *RNP-denovo* fold-and-dock method on the same set of ten RNA-protein complexes that was used for the prefold-then-dock tests described above. Again, to simplify the problem and simulate the availability of limited experimental data, we assumed the positions of some of the RNA nucleotides. For single-stranded RNA binding proteins, we assumed the positions of the 5′ and 3′ nucleotides relative to the protein because, for many single-stranded RNA binding proteins, RNA-protein contacts around these positions are highly conserved (Maris et al., 2005), therefore allowing these nucleotides to serve as anchor points during modeling. This is also analogous to what has been done in previous modeling work (Chauvot de Beauchene et al., 2016a, 2016b). For RNA-protein complexes containing RNA helices, we assumed the relative positions of the RNA helices. This is information that could, for example, be obtained from low-resolution cryoelectron microscopy (cryo-EM) maps (Kappel et al., 2018). The exact residues that were kept fixed are listed in Table 1 and in the STAR Methods section. We then built models of the remaining RNA residues in the presence of the protein with *RNP-denovo*. To assess the effect of the RNA-protein score terms, we first built a set of models with a score function that included only the RNA-specific score terms and the RNA-protein steric penalty. The best of the top 100 scoring models (top 1.4%) achieved an average RMSD of 6.0 Å, with RMSDs better than 5 Å in three out of ten cases (Table S1). For models built with all of the RNA-protein score terms included, the average RMSD over the best of the top 100 scoring models improved to 4.3 Å, with RMSDs better than 5 Å in seven out of ten cases (Tables 1 and S1), recovering near-native RNA folds for all systems (Figures 3A–3J). In all cases, the inclusion of the RNA-protein score terms resulted in a shift in the distribution of *RNP-denovo* models toward lower RMSDs (Figures 3A–3J).

The Rosetta RNA-protein score function was derived from a set of complexes that includes some of the ten RNA-protein systems tested here. To test whether this affected the accuracy of our models, we additionally built models using a separate RNA-protein potential derived from a set of structures that did not include any of these ten RNA-protein systems or homologs of these structures. The accuracy of these models was similar to the accuracy of the models built with the original score function, with an average RMSD over the best of the top 100 scoring models of 4.0 Å for models built with the new score function compared with 4.3 Å for models built with the original score function (Table S1). We also tested whether the inclusion of fragments from structures homologous to those in the benchmark set affected our results. Models built using a fragment library that excluded fragments from homologous structures had an average RMSD over the best of the top 100 scoring models of 4.2 Å, again similar to the accuracy of the original models (Table S1).

Testing Alternative Score Functions

The major difference between the Rosetta RNA-protein potential described here and previously developed RNA-protein docking potentials is the inclusion of terms describing RNA-RNA interactions, which can safely be neglected for rigid-body docking problems. To test whether existing docking potentials might produce similar results if integrated into an algorithm for structure prediction of RNA at protein interfaces, such as the *RNP-denovo* fold-and-dock method described here, we rescored our *RNP-denovo* models with the 3dRPC (Huang et al., 2016) and DARS-RNP docking potentials (Tuszynska and Bujnicki, 2011). For the single-stranded RNA binding proteins, the docking potentials picked out models with accuracy similar to the full Rosetta RNP potential (Table S2; Figure 4A). However, for three of the RNA-protein systems containing structured RNA, the docking potentials picked out models that deviated significantly from the native conformations with RMSDs $\geq$ 14.7 Å. Over all ten systems, the average RMSD of the best-scoring models was 11.6 Å for 3dRPC and 10.2 Å for DARS-RNP compared with 6.4 Å for the Rosetta RNA-protein score function (Table S2; Figure 4A). The docking potentials performed worst for 1DFU, with

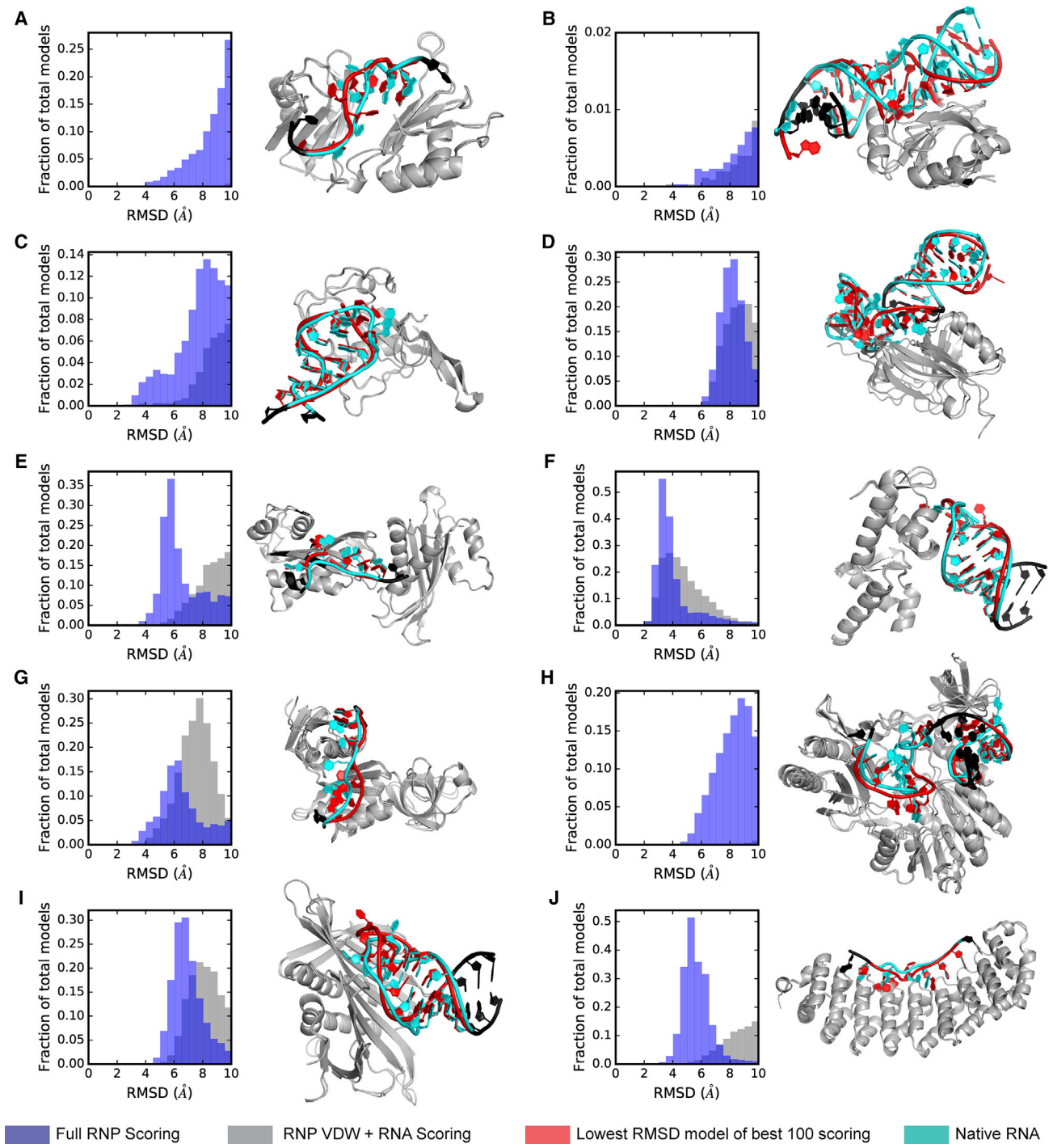
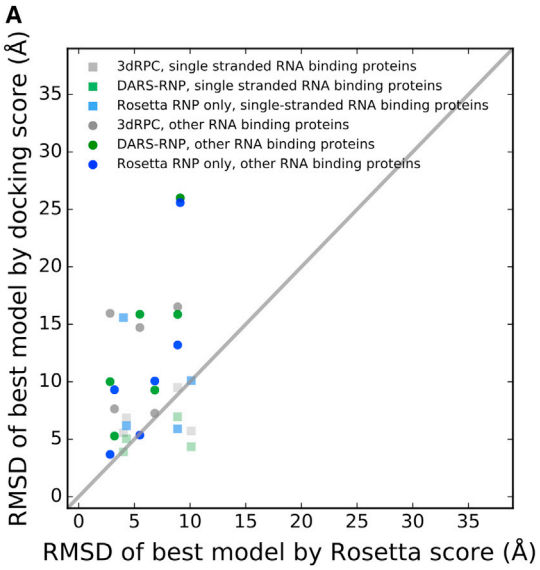


Figure 3. Rosetta *RNP-denovo* Results

Left, histograms of *RNP-denovo* RMSDs relative to native for Rosetta models built with the full RNP score function (blue) and the RNP van der Waals term plus RNA-only score terms (gray); and right, the best models (by RMSD) out of the top 100 scoring (RNA in red) overlaid with the native RNA-protein complexes (RNA in cyan) for (A) PDB: 1B7F, (B) PDB: 1DFU, (C) PDB: 1JBS, (D) PDB: 1P6V, (E) PDB: 1WPU, (F) PDB: 1WSU, (G) PDB: 2ASB, (H) PDB: 2BH2, (I) PDB: 2QUX, and (J) PDB: 3BX2. RNA residues that were kept fixed relative to the protein are colored black. Unbound protein structures (gray) were used for modeling to simulate a realistic prediction scenario.

See also Table S1.



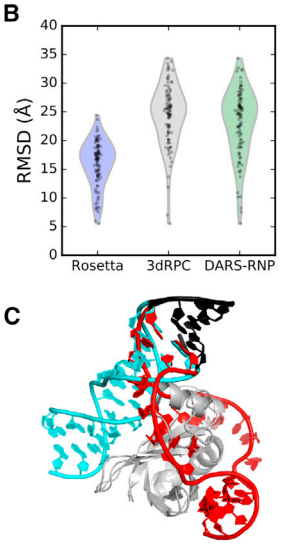


Figure 4. Rescoring *RNP-denovo* Rosetta Models with Docking Potentials

(A) The RMSD of the best-scoring Rosetta models assessed by the DARS-RNP (green), 3dRPC (gray), and Rosetta (RNA-protein score terms only; blue) docking potentials versus by the full Rosetta RNP score function for systems shown in Figure 3. Values for single-stranded RNA binding proteins are shown as squares.

(B) Distributions of RMSDs for the top 100 scoring models assessed by the DARS-RNP (green), 3dRPC (gray), and full Rosetta RNP potentials (blue).

(C) The best-scoring model assessed by 3dRPC for 1DFU (RNA colored red) overlaid with the native structure (RNA colored cyan). Residues that were kept fixed during modeling are colored black.

See also Table S2.

the distribution of model accuracy over the top scoring models shifting dramatically toward poorer RMSDs (Figure 4B). The best-scoring *RNP-denovo* model picked out by both DARS-RNP and 3dRPC adopts a conformation in which the two RNA strands, which interact in the native complex, wrap around opposite sides of the protein to maximize the number of RNA-protein contacts (Figure 4C). As an additional comparison, we rescored the *RNP-denovo* models with just the five Rosetta RNA-protein (intermolecular) score terms. The results were similar to the poor results with prior RNA-protein docking potentials, with an average RMSD over the best-scoring models for each of the ten systems of 10.5 Å (Table S2; Figure 4A). These results highlight the importance of RNA-RNA interactions in RNA-protein complexes. All "decoy" models are being made available at https://purl.stanford.edu/gt072md8147 to allow testing of new RNP scoring functions.

Applying *RNP-denovo* to Three Past Modeling Challenges of Large RNPs

To test whether *RNP-denovo* would be useful for real modeling problems, we revisited three past modeling challenges for large RNA-protein complexes in which only sparse experimental data were available. These three systems are substantially larger than the complexes in our initial benchmark set; the average number of protein and RNA residues per system is 702 and 117 residues, respectively, compared with 215 and 21 residues, respectively, for the initial benchmark set. Due to the increased size of these problems, we expected the RMSD values to be higher than for the initial benchmark set. Using a previously determined relationship between number of residues and RMSD (Carugo and Pongor, 2001), we calculated that the structural similarity specified by an RMSD of 4.3 Å (average for the initial benchmark set) for complexes with 21 RNA residues on average would correspond approximately to an RMSD of 21 Å for the larger complexes (see the STAR Methods). We therefore targeted RMSD accuracies of better than 21 Å as representative of correct global fold recovery for these three larger systems. First, we built models of the human telomerase core RNP based on FRET measurements, which provided ten distance restraints between specific pairs of RNA residues (Parks et al., 2017). Models of this system were previously built in 2015 using FARNA with these FRET constraints and an additional score term that penalized RNA-protein steric clashes (Parks et al., 2017). Here, we followed the previous modeling procedure, but used the *RNP-denovo* fold-and-dock method and the more advanced Rosetta RNP statistical potential described here (see the STAR Methods). As before, the positions of the template hybrid and CR4/5 RNA were kept fixed relative to the homology model of the telomerase reverse transcriptase (TERT) protein. Although a high-resolution structure of the human telomerase core RNP has still not been determined, the accuracy of the newly and previously built models could be assessed by comparison with the recently solved 7.7 Å resolution cryo-EM structure of the human telomerase RNP (Nguyen et al., 2018) (which was determined after the original and *RNP-denovo* models were built). Specifically, we considered the positioning of the highly conserved RNA pseudoknot motif relative to the TERT protein. Qualitatively, both the previously built and the new *RNP-denovo* fold-and-dock models agree well with the cryo-EM structure, with the pseudoknot positioned on the correct face of TERT. The best *RNP-denovo* model (out of the top 100 scoring models) has an improved RMSD accuracy of 13.2 Å over the ends of the pseudoknot motif (see the STAR Methods), compared with 17.1 Å for the best of the top 100 previously built models (Figures S5A–S5C). Although the accuracy here is worse than for the systems in the initial benchmark set, it is still reasonable given the increased size of this problem. Notably, at 13.2 Å RMSD, the global fold is still correctly recovered (Figure 5A, middle and right panels).

We additionally tested whether the inclusion of the FRET data was necessary to build accurate models of the telomerase core RNP, as was noted for the previously built models (Parks et al., 2017). The top scoring models built without FRET data using *RNP-denovo* were on average more accurate than top scoring models built with the previous method without FRET data (Figures S5D and S5E). The most accurate *RNP-denovo* model of the top 100 scoring models had an RMSD of 15.0 Å, compared with 51.4 Å for models built with the previous method (scoring with RNA score terms and RNA-protein sterics only) and

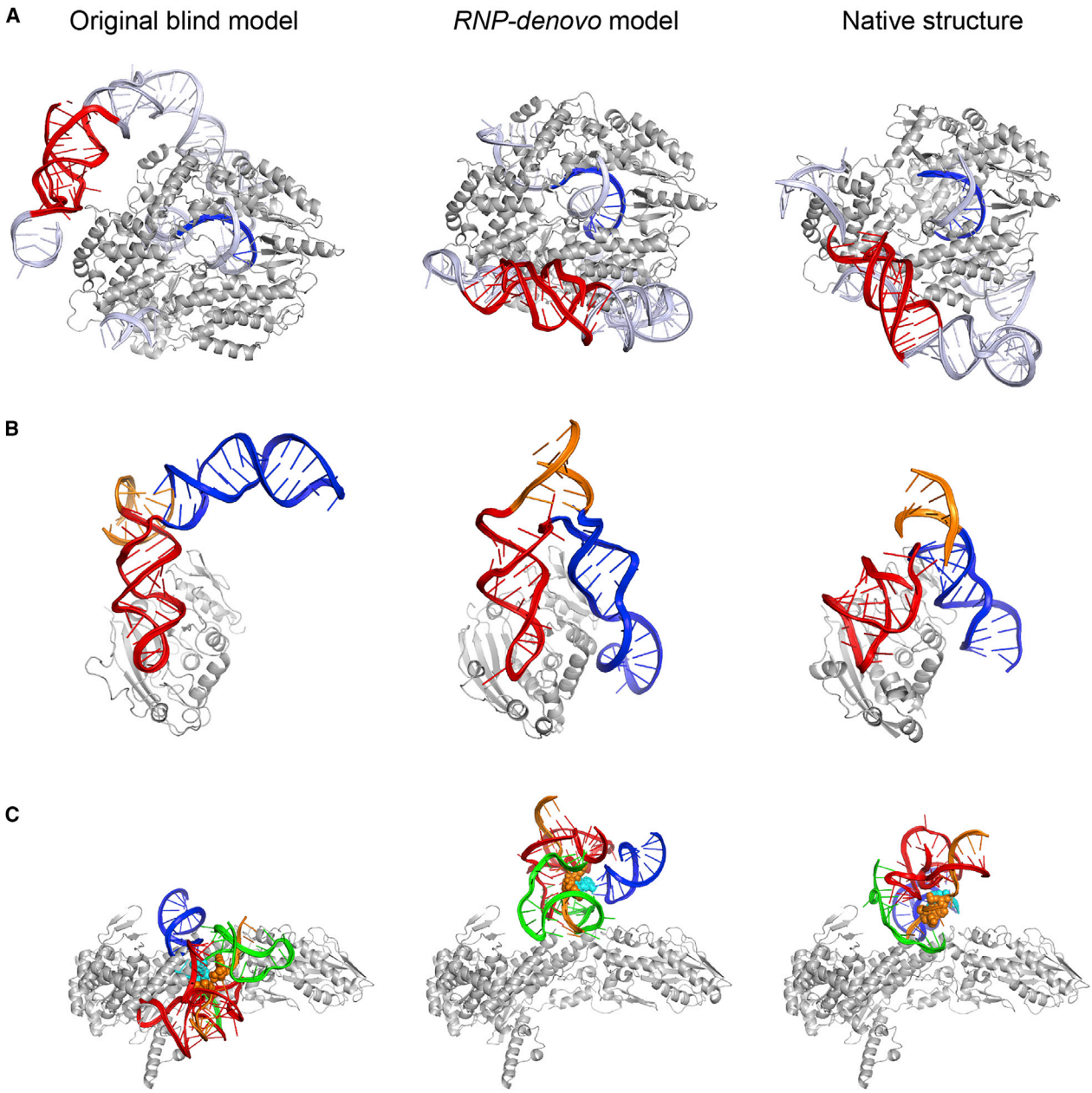


Figure 5. Revisiting Three Past Modeling Challenges with the Rosetta *RNP-denovo* Fold-and-Dock Method

(A) The best of the previously selected blind human telomerase core RNP models built without FRET data (Parks et al., 2017) (left; RMSD over select pseudoknot residues = 78.8 Å; pseudoknot RNA motif colored red, template RNA colored blue, other modeled RNA colored light blue, protein colored gray), the best *RNP-denovo* fold-and-dock model by RMSD of the top 100 scoring models built without FRET (middle; RMSD over select pseudoknot residues = 15.0 Å), and the cryo-EM structure of human telomerase (Nguyen et al., 2018) (right).

(B) The best of the previously submitted ten RNA methyltransferase CAPRI T33 models (left; RMSD = 31.0 Å; RNA colored blue, red, and orange; protein colored gray), the best of the top 100 scoring *RNP-denovo* fold-and-dock models (middle; RMSD = 13.6 Å), and the best T34 model, which closely resembles the crystal structure (right; interface RMSD to crystal structure = 1.5 Å).

(C) The previously published human spliceosomal C complex model (Anokhina et al., 2013) (left; RMSD over key active site residues [shown as spheres] = 34.5 Å, U2 RNA colored green, U5 RNA colored blue, U6 RNA colored red, intron colored orange, 5′ exon colored cyan, protein colored gray), the best RMSD model of the top 100 scoring *RNP-denovo* fold-and-dock models of the human C complex (middle; RMSD over key active site residues = 8.0 Å), and the cryo-EM structure of the yeast C complex (right).

See also Figure S5.

25.5 Å for the top 100 scoring previously built blind models (Figure 5A). In addition, predicted FRET values calculated from models built using the previous method correlate poorly with the experimental FRET values, with a maximum correlation of 0.37 (Figure S5F). This correlation is improved for many of the models built with the *RNP-denovo* method, with the best

RMSD model achieving the maximum correlation of 0.60 (Figure S5F). These results suggest that, in contrast to the previous approach, *RNP-denovo* does not require FRET data to sample accurate models, although inclusion of the data still enriches for and helps select accurate models.

We then revisited target 33/34, an RNA methyltransferase, from the 2008 CAPRI blind modeling challenge (Janin, 2010). Target 33 challenged modelers to build a structure of the full RNA-protein complex, starting from sequence only. Subsequently, for target 34, modelers were provided with the crystal structure of the bound RNA and asked to predict the structure of the RNA-protein complex. Originally, models for target 33 were built with Rosetta using an *ad hoc* prefold-then-dock approach, with restraints based on RNA chemical structure probing data, highly conserved protein residues likely to interact with the RNA, and a homology model of the methyltransferase (Fleishman et al., 2010) and (STAR Methods). Although the crystal structure of the complex has not yet been released, the accuracy of the models can be assessed by comparison with the best target 34 model of the bound RNA crystal structure docked to the protein, which CAPRI evaluators confirmed to be close to the crystal structure of the complex (interface RMSD [I_rmsdBB] of 1.5 Å; http://www.ebi.ac.uk/msd-srv/capri/round15/round15.html). Using this structure for comparison, the previously submitted target 33 model achieved 31.0 Å RMSD over the RNA after aligning over the protein. To determine whether the *RNP-denovo* fold-and-dock method could build more accurate models, we applied it to this problem and additionally repeated the prefold-then-dock modeling to control for possible changes to the RNA modeling procedure (see the STAR Methods). In each case, all RNA residues were allowed to move relative to the protein. When comparing the accuracy over the full complex (RMSD of the RNA after aligning to the protein), the best of the top 100 scoring *RNP-denovo* models was more accurate than the best of the top 100 scoring prefold-then-dock models, with RMSDs of 13.6 and 18.2 Å, respectively (Figure S5G, Figure 5B).

Finally, we applied the *RNP-denovo* fold-and-dock method to build a model of the human spliceosomal C complex active site. Prior to the determination of high-resolution cryo-EM structures, Anokhina et al. (2013) built a model of the core RNA elements bound to the Prp8 protein based on RNA chemical structure probing data, crosslinking data, homology to the group II intron, and a crystal structure of the Prp8 protein alone. While many features of this model, particularly the relative arrangement of the intron and U2 and U6 RNA, agree well with the later-solved cryo-EM structures, the positioning of these RNA elements relative to the Prp8 protein was not highly accurate, with a 34.5 Å RMSD over key active site RNA residues (the 3′ residue of the 5′ exon, the 5′ residue of the intron, and the residue immediately 3′ of the branch point adenosine) after alignment over Prp8 (Figure 5C). This inaccuracy can potentially be explained by the fact that the RNA model was docked rigidly to Prp8 despite explicitly noted uncertainty in the U5 RNA positioning (Figures S5H and S5I). Applying the Rosetta prefold-then-dock approach to this problem, without keeping any RNA residues fixed relative to the protein, gave improved RMSD accuracy over active site residues of 13.8 Å for the best of the top 100 scoring models for which the crosslinking constraints were satisfied, although the global fold over the entire RNA-protein complex was not recovered (Figure S5I). The complete *RNP-denovo* fold-and-dock method has the potential to further improve the accuracy of the model by explicitly accounting for the uncertainty in U5 RNA positioning and allowing it to move relative to the rest of the RNA during docking to Prp8. Starting from the published RNA model and using the same crosslinking constraints, but allowing U5 to move, and again without keeping any RNA residues fixed relative to the protein, the *RNP-denovo* Rosetta fold-and-dock method resulted in models of the Prp8-RNA complex that were more accurate than the previously published model and the prefold-then-dock models, although the global fold was still not completely accurate (Figure 5C). Over just active site residues, the best of the top 100 scoring models for which the crosslinking constraints were satisfied achieved 8.0 Å RMSD (Figures S5I and 5C orange and cyan space-filled residues).

Overall, for three large RNA-protein systems with sparse experimental restraints, the *RNP-denovo* method resulted in models with improved or similar accuracy compared with previously published models. These tests suggest that this method is useful for real modeling challenges, sampling an ensemble of models with biologically correct global folds or placement of functional residues.

DISCUSSION

Structure prediction of RNA-protein complexes has remained a relatively unexplored area of research, with efforts predominantly focused on RNA structure prediction without consideration of RNA-protein binding or separately on RNA-protein rigid-body docking. A critical bottleneck is the computational difficulty of sampling *de novo* the new conformations that RNAs form when interacting with protein surfaces. Tests presented here show that combining existing tools in a prefold-then-dock strategy does not generally lead to accurate models of RNA-protein complexes, and that simultaneous optimization of RNA structure and rigid-body orientation is necessary to more accurately predict the structures of these complexes. Over a benchmark set of ten RNA-protein complexes, with the assumption of a few RNA-protein contacts, a Rosetta *RNP-denovo* fold-and-dock method recovered native-like RNA folds in all cases. The knowledge-based RNA-protein potential implemented in Rosetta enriched sampling of near-native models. In addition, models favored by potentials that do not include intramolecular RNA interactions were less accurate compared with those favored by the full Rosetta RNP potential for systems containing structured RNA, suggesting that it is necessary to balance consideration of both RNA-RNA and RNA-protein interactions to accurately fold protein-bound RNA structures. Finally, application of the *RNP-denovo* fold-and-dock method to three past modeling challenges of large RNA-protein systems resulted in improved accuracy compared with previously published models built with other methods and to prefold-then-dock methods, suggesting that *RNP-denovo* will be useful in real modeling scenarios. Overall, *RNP-denovo* appears to resolve a critical sampling bottleneck for *de novo* prediction of protein-bound RNA structures. We expect the method to be widely useful for structure determination, particularly because Rosetta already allows integration of numerous kinds of experimental data.

Although *RNP-denovo* represents an advance in our ability to predict the structures of RNA-protein complexes, fully *de novo* structure prediction without experimental data remains an unsolved challenge. The benchmark on the ten small RNA-protein systems presented here relied on having limited information about specific RNA-protein contacts; without this information, we do not yet have the tools to accurately predict structures of RNA-protein complexes. In addition, the tests on three larger RNA-protein systems suggest that *RNP-denovo* will be useful for real modeling challenges, but also highlight that high-accuracy RNA-protein modeling remains an unsolved problem. Our results suggest several possible reasons as to why RNA-protein structure prediction remains difficult. First, RNA-protein interactions, unlike RNA-RNA base pairing, are not highly stereotyped, making the development of a predictive low-resolution potential difficult. Second, the development of a statistical potential is hindered by the relatively small number of RNA-protein structures in the PDB; our non-redundant set of crystal structures contains only 154 systems. Finally, the overall conformation of an RNA-protein complex is determined by a balance of both inter- and intramolecular interactions, resulting in a folding landscape with many local energy minima in which just one of these sets of interactions may be optimized at the expense of the other. This scenario was observed in the tests of 1DFU structure prediction presented here: in addition to near-native models, there were many models generated in which the RNA-protein contacts were maximized at the expense of RNA structure. Efficiently sampling these conformations for large systems and in fully *de novo* tests will be a challenge.

The results described here suggest several additional areas for future improvement. First, the success of this method relies on having limited experimental data. Here for the benchmark of small RNA-protein systems, we simulated this situation by assuming specific RNA-protein contacts, and for the three large RNA-protein tests, we used FRET or crosslinking data, or information about highly conserved residues. However this method could be further generalized to include other types of experimental information such as nuclear magnetic resonance restraints (Zhang et al., 2018), contacts derived from evolutionary couplings (Weinreb et al., 2016), or small-angle X-ray scattering data (Schneidman-Duhovny et al., 2012; Schwieters et al., 2018), as we have accomplished recently for cryo-EM data, achieving models with near-atomic accuracy in blind challenges (Kappel et al., 2018). Second, improving the accuracy of this method and discriminating the top model (rather than the top 100) will require new high-resolution refinement methods that can be applied to RNA-protein complexes. Currently, conformations are scored exclusively with a low-resolution knowledge-based potential. Refining these low-resolution structures with a full-atom energy function will likely be necessary to improve the overall accuracy of the models and to improve discrimination of near-native versus non-native conformations. Preliminary tests show that full-atom refinement does not dramatically change the accuracy of *RNP-denovo* models and that it improves the RMSD of the best-scoring model in some but not all cases (Figure S6), suggesting that better refinement strategies are needed. Finally, this method is specific to RNA-protein complexes, but it could be generalized to also model DNA-protein complexes. This will require generating a DNA-protein score function analogous to the knowledge-based RNA-protein potential described here, and possibly also developing new sampling strategies appropriate for DNA.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - The Benchmark Set of Ten RNA-Protein Complexes
  - The Prefold-then-Dock Protocol
  - Assembling a Non-redundant Set of Crystal Structures of RNA-Protein Complexes
  - Development of the Rosetta Low-Resolution RNP Score Function
  - Scoring RNA-protein Interactions in the Plane of the Base (`rnp_base_pair`)
  - Scoring RNA-Protein Stacking Interactions (`rnp_stack`)
  - Scoring General RNA-Protein Interactions (`rnp_pair_dist`)
  - Scoring Interactions with the RNA Backbone (`rnp_aa_to_rna_backbone`)
  - Scoring Steric Clashes (`rnp_vdw`)
  - Simultaneous Folding and Docking in Rosetta
  - Excluding Homologous Fragments during RNP-denovo Modeling
  - Rescoring Rosetta Models with Docking Potentials
  - Telomerase Modeling
  - CAPRI Target 33/34 RNA Methyltransferase Modeling
  - Spliceosome Modeling
  - Full-Atom Refinement of RNP-denovo Models
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Modeling and RMSD Calculation
  - Calculating Target RMSD Accuracy for Larger RNA-Protein Complexes
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and three tables and can be found with this article online at https://doi.org/10.1016/j.str.2018.10.001.

ACKNOWLEDGMENTS

We thank members of the Das lab for helpful discussions and members of the Rosetta community for sharing code. We thank Zhichao Miao and Eric Westhof for sharing the coordinates for their human spliceosome model. Calculations were performed on the Stanford BioX^3 cluster, supported by NIH Shared Instrumentation Grant 1S10RR02664701, and the Sherlock cluster. This work was supported by a Gabilan Stanford Graduate Fellowship (K.K.), an NSF GRFP (K.K.), NIH NIGMS MIRA R35 GM122579 (R.D.), NIH NIGMS R01 GM121487 (R.D.), and a RosettaCommons grant (R.D.).

AUTHOR CONTRIBUTIONS

K.K. and R.D. designed the computational approach. K.K. implemented the method and performed the tests and analysis. K.K. and R.D. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 7, 2018
Revised: August 27, 2018
Accepted: October 5, 2018
Published: November 8, 2018

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Anokhina, M., Bessonov, S., Miao, Z.C., Westhof, E., Hartmuth, K., and Luhrmann, R. (2013). RNA structure analysis of human spliceosomes reveals a compact 3D arrangement of snRNAs at the catalytic core. EMBO J. *32*, 2804–2818.

Beuth, B., Pennell, S., Arnvig, K.B., Martin, S.R., and Taylor, I.A. (2005). Structure of a *Mycobacterium tuberculosis* NusA-RNA complex. EMBO J. *24*, 3576–3587.

Carugo, O., and Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci. *10*, 1470–1473.

Chao, J.A., Patskovsky, Y., Almo, S.C., and Singer, R.H. (2008). Structural basis for the coevolution of a viral RNA-protein complex. Nat. Struct. Mol. Biol. *15*, 103–105.

Chauvot de Beauchene, I., de Vries, S.J., and Zacharias, M. (2016a). Binding site identification and flexible docking of single stranded RNA to proteins using a fragment-based approach. PLoS Comput. Biol. *12*, e1004697.

Chauvot de Beauchene, I., de Vries, S.J., and Zacharias, M. (2016b). Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. Nucleic Acids Res. *44*, 4565–4580.

Cheng, C.Y., Chou, F.C., and Das, R. (2015). Modeling complex RNA tertiary folds with Rosetta. Methods Enzymol. *553*, 35–64.

Crowder, S.M., Kanaar, R., Rio, D.C., and Alber, T. (1999). Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of sex-lethal. Proc. Natl. Acad. Sci. U S A *96*, 4892–4897.

Das, R., and Baker, D. (2007). Automated de novo prediction of native-like RNA tertiary structures. Proc. Natl. Acad. Sci. U S A *104*, 14664–14669.

Dong, G., Nowakowski, J., and Hoffman, D.W. (2002). Structure of small protein B: the protein component of the tmRNA-SmpB system for ribosome rescue. EMBO J. *21*, 1845–1854.

Fleishman, S.J., Corn, J.E., Strauch, E.M., Whitehead, T.A., Andre, I., Thompson, J., Havranek, J.J., Das, R., Bradley, P., and Baker, D. (2010). Rosetta in CAPRI rounds 13-19. Proteins *78*, 3212–3218.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845.

Gopal, B., Haire, L.F., Gamblin, S.J., Dodson, E.J., Lane, A.N., Papavinasasundaram, K.G., Colston, M.J., and Dodson, G. (2001). Crystal structure of the transcription elongation/antitermination factor NusA from *Mycobacterium tuberculosis* at 1.7 angstrom resolution. J. Mol. Biol. *314*, 1087–1095.

Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. *331*, 281–299.

Guilhot-Gaudeffroy, A., Froidevaux, C., Aze, J., and Bernauer, J. (2014). Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities. PLoS One *9*, e108928.

Gutmann, S., Haebel, P.W., Metzinger, L., Sutter, M., Felden, B., and Ban, N. (2003). Crystal structure of the transfer-RNA domain of transfer-messenger RNA in complex with SmpB. Nature *424*, 699–703.

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and Yokoyama, S. (1999). Structural basis for recognition of the tra mRNA precursor by the sex-lethal protein. Nature *398*, 579–585.

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., Li, H., and Xiao, Y. (2016). Using 3dRPC for RNA-protein complex structure prediction. Biophys. Rep. *2*, 95–99.

Huang, Y.Y., Liu, S.Y., Guo, D.C., Li, L., and Xiao, Y. (2013). A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Sci. Rep. *3*, 1887.

Iwakiri, J., Tateishi, H., Chakraborty, A., Patil, P., and Kenmochi, N. (2012). Dissecting the protein-RNA interface: the role of protein surface shapes and RNA secondary structures in protein-RNA recognition. Nucleic Acids Res. *40*, 3299–3306.

Janin, J. (2010). The targets of CAPRI rounds 13-19. Proteins *78*, 3067–3072.

Jiang, J., Wang, Y., Sušac, L., Chan, H., Basu, R., Zhou, Z.H., and Feigon, J. (2018). Structure of telomerase with telomeric DNA. Cell *173*, 1179–1190.e13.

Kappel, K., Liu, S., Larsen, K.P., Skiniotis, G., Puglisi, E.V., Puglisi, J.D., Zhou, Z.H., Zhao, R., and Das, R. (2018). De novo computational RNA modeling into cryoEM maps of large ribonucleoprotein complexes. bioRxiv. https://doi.org/10.1101/332791.

Kondo, J., and Westhof, E. (2011). Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. Nucleic Acids Res. *39*, 8628–8637.

Kumarevel, T., Mizuno, H., and Kumar, P.K. (2005). Structural basis of HutP-mediated anti-termination and roles of the Mg2+ ion and L-histidine ligand. Nature *434*, 183–191.

Lee, T.T., Agarwalla, S., and Stroud, R.M. (2004). Crystal structure of RumA, an iron-sulfur cluster containing *E. coli* ribosomal RNA 5-methyluridine methyltransferase. Structure *12*, 397–407.

Lee, T.T., Agarwalla, S., and Stroud, R.M. (2005). A unique RNA fold in the RumA-RNA-cofactor ternary complex contributes to substrate selectivity and enzymatic function. Cell *120*, 599–611.

Lensink, M.F., Velankar, S., and Wodak, S.J. (2017). Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. Proteins *85*, 359–377.

Lensink, M.F., and Wodak, S.J. (2010). Docking and scoring protein interactions: CAPRI 2009. Proteins *78*, 3073–3084.

Li, C.H., Cao, L.B., Su, J.G., Yang, Y.X., and Wang, C.X. (2012). A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins *80*, 14–24.

Lu, M., and Steitz, T.A. (2000). Structure of *Escherichia coli* ribosomal protein L25 complexed with a 5S rRNA fragment at 1.8-A resolution. Proc. Natl. Acad. Sci. U S A *97*, 2023–2028.

Maris, C., Dominguez, C., and Allain, F.H.T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. *272*, 2118–2131.

Miao, Z., Adamiak, R.W., Antczak, M., Batey, R.T., Becka, A.J., Biesiada, M., Boniecki, M.J., Bujnicki, J., Chen, S.J., Cheng, C.Y., et al. (2017). RNA-puzzles round III: 3D RNA structure prediction of five riboswitches and one ribozyme. RNA *23*, 655–672.

Miao, Z.C., and Westhof, E. (2017). RNA structure: advances and assessment of 3D structure prediction. Annu. Rev. Biophys. *46*, 483–503.

Miller, M.T., Higgin, J.J., and Hall, T.M.T. (2008). Basis of altered RNA-binding specificity by PUF proteins revealed by crystal structures of yeast Puf4p. Nat. Struct. Mol. Biol. *15*, 397–402.

Mitchell, S.F., and Parker, R. (2014). Principles and properties of eukaryotic mRNPs. Mol. Cell *54*, 547–558.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)—Round XII. Proteins *86*, 7–15.

Nguyen, T.H.D., Tam, J., Wu, R.A., Greber, B.J., Toso, D., Nogales, E., and Collins, K. (2018). Cryo-EM structure of substrate-bound human telomerase holoenzyme. Nature *557*, 190–195.

Parks, J.W., Kappel, K., Das, R., and Stone, M.D. (2017). Single-molecule FRET-Rosetta reveals RNA structural rearrangements during human telomerase catalysis. RNA *23*, 175–188.

Perez-Cano, L., and Fernandez-Recio, J. (2010). Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins *78*, 25–35.

Perez-Cano, L., Solernou, A., Pons, C., and Fernandez-Recio, J. (2010). Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Pac. Symp. Biocomput. 293–301, https://doi.org/10.1142/9789814295291_0031.

Rahman, M.M., Muhseen, Z.T., Junaid, M., and Zhang, H.J. (2015). The aromatic stacking interactions between proteins and their macromolecular ligands. Curr. Protein Pept. Sci. *16*, 502–512.

Rau, M., Stump, W.T., and Hall, K.B. (2012). Intrinsic flexibility of snRNA hairpin loops facilitates protein binding. RNA *18*, 1984–1995.

Schlundt, A., Tants, J.N., and Sattler, M. (2017). Integrated structural biology to unravel molecular mechanisms of protein-RNA recognition. Methods *118*, 119–136.

Schneidman-Duhovny, D., Kim, S.J., and Sali, A. (2012). Integrative structural modeling with small angle X-ray scattering profiles. BMC Struct. Biol. *12*, 17.

Schwieters, C.D., Bermejo, G.A., and Clore, G.M. (2018). Xplor-NIH for molecular structure determination from NMR and other data sources. Protein Sci. *27*, 26–40.

Selmer, M., and Su, X.D. (2002). Crystal structure of an mRNA-binding fragment of *Moorella thermoacetica* elongation factor SelB. EMBO J. *21*, 4145–4153.

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

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209–225.

Stoldt, M., Wohnert, J., Gorlach, M., and Brown, L.R. (1998). The NMR structure of *Escherichia coli* ribosomal protein L25 shows homology to general stress proteins and glutaminyl-tRNA synthetases. EMBO J. *17*, 6377–6384.

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., Matelska, D., Magnus, M., Chojnowski, G., Kasprzak, J.M., Kozlowski, L.P., Dunin-Horkawicz, S., and Bujnicki, J.M. (2014). Computational modeling of protein-RNA complex structures. Methods *65*, 310–319.

Wan, R.X., Yan, C.Y., Bai, R., Huang, G.X.Y., and Shi, Y.G. (2016). Structure of a yeast catalytic step I spliceosome at 3.4 Å resolution. Science *353*, 895–904.

Ward, A.B., Sali, A., and Wilson, I.A. (2013). Integrative structural biology. Science *339*, 913–915.

Weinreb, C., Riesselman, A.J., Ingraham, J.B., Gross, T., Sander, C., and Marks, D.S. (2016). 3D RNA and functional interactions from evolutionary couplings. Cell *165*, 963–975.

Yang, X.J., Gerczei, T., Glover, L., and Correll, C.C. (2001). Crystal structures of restrictocin-inhibitor complexes with implications for RNA recognition and base flipping. Nat. Struct. Biol. *8*, 968–973.

Yang, X.J., and Moffat, K. (1996). Insights into specificity of cleavage and mechanism of cell entry from the crystal structure of the highly specific *Aspergillus* ribotoxin, restrictocin. Structure *4*, 837–852.

Yoshizawa, S., Rasubala, L., Ose, T., Kohda, D., Fourmy, D., and Maenaka, K. (2005). Structural basis for mRNA recognition by elongation factor SelB. Nat. Struct. Mol. Biol. *12*, 198–203.

Zhang, K.M., Keane, S.C., Su, Z.M., Irobalieva, R.N., Chen, M.Y., Van, V., Sciandra, C.A., Marchant, J., Heng, X., Schmid, M.F., et al. (2018). Structure of the 30 kDa HIV-1 RNA dimerization signal by a hybrid cryo-EM, NMR, and molecular dynamics approach. Structure *26*, 490–498.e3.

Zheng, S., Robertson, T.A., and Varani, G. (2007). A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. FEBS J. *274*, 6378–6391.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
RNP-denovo models	This paper	https://purl.stanford.edu/gt072md8147
Structure of the sex-lethal RRM complex	RCSB PDB (Handa et al., 1999)	PDB: 1B7F
Structure of unbound sex-lethal RRM protein	RCSB PDB (Crowder et al., 1999)	PDB: 3SXL
Structure of the ribotoxin-restrictocin – SRL analog complex	RCSB PDB (Yang et al., 2001)	PDB: 1JBS
Structure of ribotoxin-restrictocin – SRL analog unbound protein	RCSB PDB (Yang and Moffat, 1996)	PDB: 1AQZ
Structure of the HutP antitermination complex	RCSB PDB (Kumarevel et al., 2005)	PDB: 1WPU
Structure of HutP antitermination unbound protein	RCSB PDB (Kumarevel et al., 2005)	PDB: 1WPV
Structure of mRNA binding domain of SelB elongation factor	RCSB PDB (Yoshizawa et al., 2005)	PDB: 1WSU
Structure of SelB unbound protein	RCSB PDB (Selmer and Su, 2002)	PDB: 1LVA
Structure of NusA transcriptional regulator complex	RCSB PDB (Beuth et al., 2005)	PDB: 2ASB
Structure of NusA unbound protein	RCSB PDB (Gopal et al., 2001)	PDB: 1K0R
Structure of methyltransferase RumA in complex with rRNA	RCSB PDB (Lee et al., 2005)	PDB: 2BH2
Structure of methyltransferase RumA unbound protein	RCSB PDB (Lee et al., 2004)	PDB: 1UWV
Structure of PP7 coat protein and viral RNA	RCSB PDB (Chao et al., 2008)	PDB: 2QUX
Structure of PP7 unbound protein	RCSB PDB (Chao et al., 2008)	PDB: 2QUD
Structure of Puf4 bound to 3' UTR of target transcript	RCSB PDB (Miller et al., 2008)	PDB: 3BX2
Structure of Puf4 unbound protein	RCSB PDB (Miller et al., 2008)	PDB: 3BWT
Structure of SmpB-tmRNA complex	RCSB PDB (Gutmann et al., 2003)	PDB: 1P6V
Structure of SmpB unbound protein	RCSB PDB (Dong et al., 2002)	PDB: 1K8H
Structure of *E. coli* L25-5S rRNA	RCSB PDB (Lu and Steitz, 2000)	PDB: 1DFU
Structure of L25 unbound protein	RCSB PDB (Stoldt et al., 1998)	PDB: 1B75
Human telomerase structure (cryo-EM)	Nguyen et al., 2018	N/A
Spliceosome model	Anokhina et al., 2013	N/A
Software and Algorithms		
RNP-denovo	This paper	www.rosettacommons.org
FARFAR	Cheng et al., 2015	www.rosettacommons.org
RPDock	Huang et al., 2013	http://biophy.hust.edu.cn/download.html
blastclust	Altschul et al., 1990	https://blast.ncbi.nlm.nih.gov/Blast.cgi
DARS-RNP	Tuszynska and Bujnicki, 2011	http://genesilico.pl/software/stand-alone/statistical-potentials
3dRPC	Huang et al., 2016	http://biophy.hust.edu.cn/download.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Rhiju Das (rhiju@stanford.edu).

METHOD DETAILS

The Benchmark Set of Ten RNA-Protein Complexes

Ten systems were chosen from the non-redundant set of RNA-protein complexes with corresponding unbound protein structures available, described in (Perez-Cano and Fernandez-Recio, 2010). The specific systems were selected manually to represent a diversity of types of RNA-protein interactions (unbound protein structures listed in parentheses): 1DFU (1B75) (Lu and Steitz, 2000; Stoldt et al., 1998), 1B7F (3SXL) (Crowder et al., 1999; Handa et al., 1999), 1JBS (1AQZ) (Yang et al., 2001; Yang and Moffat, 1996), 1P6V

(1K8H) (Dong et al., 2002; Gutmann et al., 2003), 1WPU (1WPV) (Kumarevel et al., 2005), 1WSU (1LVA) (Selmer and Su, 2002; Yoshizawa et al., 2005), 2ASB (1K0R) (Beuth et al., 2005; Gopal et al., 2001), 2BH2 (1UWV) (Lee et al., 2004, 2005), 2QUX (2QUD) (Chao et al., 2008), and 3BX2 (3BWT) (Miller et al., 2008). Fixed residues for tests in which some RNA residues were held fixed relative to the protein were selected as follows: for systems containing only single stranded RNA (1B7F, 1WPU, 2ASB, 3BX2), the first and last RNA residues were kept fixed; for 1DFU, 1WSU, and 2QUX, the first three base pairs were held fixed; for 1P6V the first two base pairs of both helices were held fixed (residues 19-20, 37-38, 41-42, 56-57); for 2BH2, the first three base pairs of the RNA helix and the 5' nucleotide were kept fixed; for 1JBS, the 5' and 3' residues were held fixed.

The Prefold-then-Dock Protocol

For each of the ten structures in the benchmark set, the RNA was folded with the FARFAR method (Cheng et al., 2015). 5000 structures were generated for each system. The resulting structures were clustered in Rosetta, with a clustering radius of 2.0 Å. The centers of the ten most populated clusters were then docked to the unbound protein structures with RPDock (Huang et al., 2013). RMSDs were calculated over RNA heavy atoms after alignment based on the protein coordinates. The best RMSD was selected from the top ten scoring models for each of the ten docked RNA cluster centers (100 structures in total for each system).

To dock the native bound conformations, the RNA and protein chains were extracted from the bound complex, then docked together with RPDOCK. The best RMSD was selected from the top 100 models for each system.

For tests with fixed RNA residues, the RNA residues described above were held fixed during RNA folding in FARFAR. RMSDs were calculated over RNA residues that were not held fixed after alignment to the fixed RNA residues. The best RMSD was selected from the top 100 models for each system (clustering was not performed).

Assembling a Non-redundant Set of Crystal Structures of RNA-Protein Complexes

All crystal structures containing both RNA and protein chains, with resolution better than 2.5 Å were downloaded from the Protein Data Bank in August 2016. Complexes containing only protein chains of length less than 20 amino acids or RNA chains of length less than four nucleotides were discarded. All protein chains were clustered using blastclust (Altschul et al., 1990) with a 30% sequence identity cutoff using the command `''blastclust -i fasta.txt -o blastclust_output.txt -b T -S 30''`. Because the protein chains were clustered individually, different protein chains of the same RNA-protein complex could end up in different clusters after this step. This was addressed by merging clusters with members from the same RNA-protein complex. A single representative structure from each cluster was then selected based on a priority score defined as $1/Resolution - R_{free} + Number\ of\ protein\ chains$, where higher priority scores were prioritized. This score favors structures with more protein chains so that complete large RNA-protein complexes are selected over smaller subsets of these complexes (e.g. this will select a complete ribosome structure over a structure of a single ribosomal protein and a fragment of ribosomal RNA). The final structures were visually inspected to check that they actually reflected the biological assemblies and that multiple copies of the same complex were not present in a structure (e.g. in the case of a virus capsid bound at every protein subunit by the same RNA hairpin). Final PDB IDs are listed in Table S3. Additionally, we generated a second list of structures that does not contain structures that are homologous to any of the structures in the benchmark set of ten RNA-protein complexes described above. For this purpose, we again used a 30% sequence identity cutoff for protein chains.

Development of the Rosetta Low-Resolution RNP Score Function

Distances between RNA and protein atoms for each of the structures in the non-redundant set were calculated in Rosetta. Distributions were analyzed (below) to inform the final pairwise score terms. The RNA-protein score function is a linear combination of the previously described RNA score terms and the five RNP score terms described here. Weights for the RNP score terms were adjusted so that the magnitudes of the final score values were similar. The final score function is available within Rosetta as `''rna_lores_with_rnp_aug.wts''`.

Scoring RNA-protein Interactions in the Plane of the Base (`rnp_base_pair`)

A coordinate system was set up on each base as described previously (Das and Baker, 2007). Distributions of each of the protein side chain centroids around each of the four RNA bases were analyzed for $0 < |z| < 3$ Å, $0 < |x| < 10$ Å, and $0 < |y| < 10$ Å, and binned into 2 Å by 2 Å boxes in the x-y plane. The resulting two-dimensional histograms were smoothed with a 1.2 Å Gaussian filter (using the gaussian_filter function in python (scipy)) and normalized by the total counts of interactions across all protein side chains and RNA bases in a given `bin`. `rnp_base_pair` scores were calculated as the negative logarithm of these frequencies. The scores were renormalized by subtracting the maximum score value across all bins, so that the maximum value of this score was equal to zero.

Scoring RNA-Protein Stacking Interactions (`rnp_stack`)

Distributions of protein side chain centroids around each of the four RNA bases were analyzed as described for `rnp_base_pair` above, but above and below the plane of the base, for $3.0 < |z| < 6.5$ Å. The distributions showed clear stacking interactions between each of the four RNA bases and tryptophan, tyrosine, phenylalanine, valine, and leucine. The `rnp_stack` score term rewards these interactions by applying a bonus of -1.0 when any of these five protein side chain centroids falls within $\sqrt{x^2 + y^2} < 4.0$ Å and $3.0 < |z| < 6.5$ Å of an RNA base.

Scoring General RNA-Protein Interactions (rnp_pair_dist)

Distances between RNA base centroids and P, C5', C1', C3' atoms and protein side chain centroids and N, C, CA, O, C atoms were binned from 0 to 20 Å in 2 Å intervals. rnp_pair_dist scores were calculated as described previously for protein-protein docking (Gray et al., 2003) (S_{pair}), but because the potential here is a function of distance rather than binary interaction, the potential was normalized for interaction volume. This normalization was performed for each distance range by dividing by the total number of interactions between any protein atom and any RNA atom within that distance. Scores were adjusted so that the maximum score value was equal to zero, and to reduce noise, resulting values between 0 and -1.0 were set to 0.

Scoring Interactions with the RNA Backbone (rnp_aa_to_rna_backbone)

Distances between RNA phosphate atoms and protein side chain centroids were binned for distances between 3 and 10 Å in 1 Å intervals. Counts in each bin were normalized by the volume of the bin and by the counts of any protein side chain in a given distance bin. The rnp_aa_to_rna_backbone scores were calculated as the negative logarithm of these frequencies. Scores were normalized to zero at 10 Å.

Scoring Steric Clashes (rnp_vdw)

Interactions between each of nine RNA atoms (P, C5', C1', C3', N1, and for adenosine N6, N7, N3, O2'; for cytidine N4, C6, O2, C2'; for guanosine O6, N7, N2, O2'; for uridine O4, C6, O2, C2') and protein side chain centroids, C, CA, O, N, and CB atoms are penalized when they come within the minimum distance observed in the non-redundant set of crystal structures. Like for the RNA and protein steric penalties (Das and Baker, 2007; Simons et al., 1997), the rnp_vdw score is calculated as $(d_{i,j}^2 - x_{i,j}^2)^2 / d_{i,j}^2$ where $x_{i,j}$ is the distance between the RNA atom i and protein atom j, and $d_{i,j}$ is the minimum observed distance.

Simultaneous Folding and Docking in Rosetta

To create *RNP-denovo*, the FARNA method was updated to include protein binding. Specifically, rigid-body docking moves are included along with the standard RNA fragment insertion moves in the Monte Carlo simulation. For runs where some RNA residues were held fixed relative to the protein (the benchmark set of ten complexes described above and telomerase described below), these rigid-body docking moves were not used, but the center of mass of the RNA could still change dramatically during modeling. Conformations are scored with the RNA-protein score function described above. This method is freely available to academic users as part of the Rosetta software package (www.rosettacommons.org). An example Rosetta command line is as follows:

```
rna_denovo -f fasta.txt -secstruct_file secstruct.txt -s protein_structure.pdb -minimize_rna false -nstruct 100 -new_fold_tree_initializer true -rna_protein_docking true -convert_protein_CEN false -FA_low_res_rnp_scoring true
```

where fasta.txt lists the full RNA and protein sequence, secstruct.txt contains the secondary structure in dot-bracket notation with the protein represented as dots, and protein_structure.pdb is the structure of the protein.

For each system, the RNA was built in the presence of the unbound protein structure or a homology model for telomerase and the CAPRI RNA methyltransferase. Coordinates for the fixed RNA residues for the benchmark set of ten complexes were taken from the bound complex. The orientation of the fixed RNA residues relative to the unbound protein was determined by aligning the unbound protein structure to the bound complex. For the benchmark set of ten complexes, 7000 models were generated for each system and RMSDs were calculated over the RNA heavy atoms after alignment of the protein coordinates. Further details for the three large RNA-protein systems are described below.

Excluding Homologous Fragments during RNP-denovo Modeling

For the models built without using fragments from homologous structures, 6-nucleotide fragments that match a 6-nucleotide sequence of purines and pyrimidines in the native structure and have the same secondary structure, with RMSDs less than 1.0 Å to the native RNA structure were excluded from the final fragment library. This was performed using the following flags during *RNP-denovo* modeling: ''-fragment_homology_rmsd 1.0 -exclusion_match_type MATCH_YR -exclude_fragment_files native_structure.pdb''.

Rescoring Rosetta Models with Docking Potentials

For each of the ten systems in the benchmark set, all 7000 Rosetta *RNP-denovo* models were rescored with the DARS-RNP and 3dRPC potentials using the following commands:

For 3dRPC:

```
3dRPC -mode 8 -system 9 -par scoring.par
```

For DARS-RNP:

```
python ~/src/DARS-RNP_v3/DARS_potential_v3.py -f list.txt -n
```

Telomerase Modeling

Models of the telomerase core RNP with an 11-nucleotide template hybrid were built without enforcing the formation of the P1 stem exactly following the procedure described previously, but using the *RNP-denovo* fold-and-dock method described here instead of the smFRET-Rosetta protocol previously described (Parks et al., 2017). As before, the template hybrid and CR4/5 RNA were kept fixed relative to the homology model of the TERT protein. Distance restraints based on FRET data were applied as described previously. As before, approximately 2500 models were built. Two additional sets of models were built following the same procedure, but without the FRET distance restraints: one with the full Rosetta RNA-protein score function and the other with only the RNA score terms and the score term penalizing RNA-protein steric clashes (rnp_vdw). RMSDs were calculated between Rosetta models and the published 7.7 Å cryo-EM structure (Nguyen et al., 2018) at the ends of the pseudoknot motif over two pseudoatoms defined as the centroids of the C1' positions of residues 96 and 118, and residues 108 and 182. RMSDs were not calculated over all atoms because the cryo-EM structure was determined at 7.7 Å resolution, which is not high enough to confidently resolve positions of individual atoms. Additionally, positions of the highly conserved pseudoknot motif differ considerably between the recently determined *Tetrahymena* (Jiang et al., 2018) and human structures (Nguyen et al., 2018), and previous single-molecule FRET studies suggest conformational flexibility (Parks et al., 2017). Predicted FRET values were calculated by converting the distances between the C5 atoms in respective atom pairs to FRET with the following equation:

$$FRET = \frac{1}{1 + (r/R_0)^6}$$

where the Förster radius, R_0 = 56 Å and r is the distance between atoms.

CAPRI Target 33/34 RNA Methyltransferase Modeling

Models of the RNA methyltransferase complex were built using either the *RNP-denovo* fold-and-dock or the prefold-then-dock method. In each case, we used the previously built protein homology model bound to the SAM cofactor (Fleishman et al., 2010) and assumed the correct RNA secondary structure. Fixed ideal A-form helices were used to model all helical elements. In both cases, constraints were applied matching those used in the original CAPRI modeling (Fleishman et al., 2010). 10000 models were built using the *RNP-denovo* fold-and-dock method. For the prefold-then-dock method, 10000 models of the RNA alone were built with FARNA. The top scoring 1000 models were clustered as described in the prefold-then-dock section above. The cluster centers of the ten most populated clusters were then docked in Rosetta to the protein homology model (using the *RNP-denovo* fold-and-dock method described here, but treating the whole RNA as a rigid body) with constraints (described above) applied. 100 models of the complex were generated for each cluster center, and the top 10 scoring were taken from each to give a final pool of 100 top scoring models. RMSDs were calculated over all RNA heavy atoms after alignment over either all protein heavy atoms or all RNA heavy atoms.

Spliceosome Modeling

Models of the spliceosome core RNA bound to Prp8 were built starting from the published RNA model (Anokhina et al., 2013). This model was docked to the Prp8 crystal structure (using the same region as for the previously published model). For the *RNP-denovo* fold-and-dock runs, the position of U5 bound to the 5' exon was allowed to move relative to the rest of the RNA model by assuming that the 3' end of the 5' exon is still covalently connected to the 5' end of the intron and allowing the torsions at this connection to move freely. For the prefold-then-dock runs, the published RNA model was treated as a rigid body. The three crosslinking constraints used to build the published model of the RNA docked to Prp8 were also included here. Specifically, a score penalty was applied to conformations where none of the C1' atoms of the U5 RNA residues were within 5.0 Å of residues 1281-1414 of Prp8. An additional penalty was applied when there were no RNA branchpoint helix residues within 5 Å of Prp8 residues 1575 or 1598 (the ends of the disordered loop in Prp8). Finally, a penalty was also applied when there were no residues within the ACAGA/pre-mRNA loop within 30 Å of residue 1826 of Prp8 (the connection to the RNaseH-like domain, which was not included in the models, meant here to represent its approximate location). All of these restraints were implemented as ambiguous flat harmonic restraints in Rosetta with standard deviations of 1 Å. Because these restraints were so stringent, they were considered "satisfied" in the final models when the total restraint score was less than 60.0 Rosetta energy units and as an additional check, U5 residue 39 was within 60 Å of the center of mass of Prp8 residues 1281-1323+1326-1413, the fourth intron residue in the branchpoint helix was within 25 Å of the center of mass of Prp8 residues 1575 and 1598, and intron residue U+4 was within 60 Å of Prp8 residue 1826. Approximately 1500 models were built. RMSDs were calculated over phosphorus atoms in the last (3') residue of the 5' exon, the first residue (5') of the intron, and the residue right after (3' of) the branchpoint A with respect to the corresponding residues from the published cryo-EM structure of the yeast C complex (Wan et al., 2016) after alignment over Prp8.

Full-Atom Refinement of RNP-denovo Models

For the full-atom refinement of *RNP-denovo* models, first RNA residues and protein side chains within 20 Å of any RNA residue were minimized, then protein side chains within 20 Å of RNA were repacked. Next, small rigid body perturbations and single residue RNA fragment insertions were attempted in a Monte Carlo simulation. Finally, RNA residues and protein side chains within 20 Å of any RNA residue were minimized again. The `''rna_hires_with_protein.wts''` full-atom score function, available in Rosetta, was used to score the structures during each of these stages.

QUANTIFICATION AND STATISTICAL ANALYSIS

Modeling and RMSD Calculation

For the benchmark set of small RNA-protein complexes, 7000 models were generated in all cases. For telomerase, 2500 models were built as was done previously (Parks et al., 2017). For CAPRI Target 33/34 and spliceosome modeling, 10000 and 1500 models were built, respectively (see details in Method Details section). RMSDs were calculated over RNA residues after alignment over protein residues as described in detail for each system in the Method Details section.

Calculating Target RMSD Accuracy for Larger RNA-Protein Complexes

The size independent RMSD metric, $RMSD_{100}$ proposed in (Carugo and Pongor, 2001) is given by:

$$RMSD_{100} = \frac{RMSD}{-1.3 + 0.5\ \ln(N)}$$

where N is the number of residues. This relationship was originally determined for proteins, but we assumed that it would also hold for RNA-protein complexes. $RMSD_{100}$ values for the initial benchmark set and the larger RNA-protein complexes should be the same for models exhibiting the same degree of structural similarity. We then used the average number of RNA residues in the initial benchmark set (N_{small} = 21) and the larger RNA-protein complexes (N_{large} = 117), as well as the average RMSD value for the initial benchmark set ($RMSD_{small}$ = 4.3 Å) to calculate the expected RMSD value ($RMSD_{large}$) for the larger RNA-protein complexes:

$$RMSD_{large} = \frac{RMSD_{small}\left(-1.3 + 0.5\ \ln\left(N_{large}\right)\right)}{-1.3 + 0.5\ \ln(N_{small})}$$

DATA AND SOFTWARE AVAILABILITY

RNP-denovo is freely available to academic users as part of the Rosetta software package at www.rosettacommons.org. Documentation is available at https://www.rosettacommons.org/docs/latest/application_documentation/rna/rnp-modeling and a demo is available at https://www.rosettacommons.org/demos/latest/public/rnp_structure_prediction/README. All models described here are available at https://purl.stanford.edu/gt072md8147.