Published online 29 April 2016

Nucleic Acids Research, 2016, Vol. 44, No. 10 ***4565–4580***
doi: 10.1093/nar/gkw328

Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins

Isaure Chauvot de Beauchene, Sjoerd J. de Vries and Martin Zacharias*

Physics Department T38, Technical University of Munich, James-Franck-Str. 1, 85748 Garching, Germany

Received December 07, 2015; Revised April 08, 2016; Accepted April 12, 2016

ABSTRACT

Protein-RNA complexes are important for many biological processes. However, structural modeling of such complexes is hampered by the high flexibility of RNA. Particularly challenging is the docking of single-stranded RNA (ssRNA). We have developed a fragment-based approach to model the structure of ssRNA bound to a protein, based on only the protein structure, the RNA sequence and conserved contacts. The conformational diversity of each RNA fragment is sampled by an exhaustive library of trinucleotides extracted from all known experimental protein–RNA complexes. The method was applied to ssRNA with up to 12 nucleotides which bind to dimers of the RNA recognition motifs (RRMs), a highly abundant eukaryotic RNA-binding domain. The fragment based docking allows a precise de novo atomic modeling of protein-bound ssRNA chains. On a benchmark of seven experimental ssRNA–RRM complexes, near-native models (with a mean heavy-atom deviation of <3 Å from experiment) were generated for six out of seven bound RNA chains, and even more precise models (deviation < 2 Å) were obtained for five out of seven cases, a significant improvement compared to the state of the art. The method is not restricted to RRMs but was also successfully applied to Pumilio RNA binding proteins.

INTRODUCTION

The triptych DNA–RNA-protein constitutes the cornerstone of cell biology. Its central element, RNA, fulfills numerous essential functions such as transmission of the genetic information outside the nucleus (messenger RNAs, mRNAs), regulation of gene transcription, maturation of mRNA and their translation into protein, and inter-cellular signaling (1,2). Most of these functions are exercised in association with proteins. RNA is recognized by partner proteins and binds it to form functional protein–RNA complexes (3,4) that may also be crucial therapeutic targets (5–7). On the other hand, protein–RNA recognition can also be a therapeutic tool (8–10). Accurate, *in silico* prediction of the structure of protein-RNA complexes could be helpful for the rational, structure-based design of new therapeutic compounds.

Predicting the structure of a complex from the structure of the constituents is known as *docking*. It consists of two main tasks: to sufficiently sample the space of possible conformations and relative orientations (i.e. poses) of the components so as to include near-native structures, and to accurately score the poses for distinguishing near-native structures from decoys. To resolve a problem of two-body docking, the most intuitive approach is to assemble unbound structures of each molecule by rigid-body docking. Such methods work well when few conformational changes occur upon binding, as in most cases of protein–protein binding (11,12). However, RNA–protein docking is hampered by the high flexibility and conformational variability of RNA, including global rearrangements, changes of secondary structure elements and the flipping out of individual bases. Such non-linear large changes are extremely difficult to predict and model, which causes all current docking methods to fail in such cases (13).

Particularly challenging is the docking of single-stranded RNAs (ssRNA). There is no experimental structure for an unbound ssRNA: in the unbound state, ssRNA is either disordered or it adopts a secondary/tertiary (i.e. non-single-stranded) structure, and the specific single-stranded conformation is only induced or selected by binding to the protein (14). Therefore, this conformation must be modelled *ab initio*. The same difficulty arises in the frequent case of a structured RNA containing a single-stranded loop. While the global structure may change relatively little upon binding, the single-stranded loop can undergo significant conformational changes. Yet, these parts often carry the specificity of recognition. Consequently, the lack of methodology for modeling ssRNAs limits the accuracy of all current protein–RNA docking methods (15,16).

The resolution by experimental methods of hundreds of structures of RNA–protein complexes allowed identification of conserved RNA binding domains in proteins. The most abundant and best characterized is the RNA recognition motif (RRM), which binds exclusively ssRNA (17)

*To whom correspondence should be addressed. Tel: +49 89 289 12335; Fax: +49 89 289 12444; Email: martin.zacharias@ph.tum.de

with variable affinity (mM to μM per motif) and sequence specificity (from RNA rich in particular bases to a fully specific sequence). On the sequence level, RRM domains are easily recognized by two consensus sequences called RNP-1 (six residues) and RNP-2 (seven residues) (18). Proteins with RRMs represent ∼2 % of the human proteome (19). Most of these proteins contain two or even three modular RRMs that fix each a part of the RNA (18).

Around 200 experimental structures of ssRNA–RRM complexes have been deposited in the Protein Data Bank (PDB). The overall structure of the RRM domain is highly conserved, with the same layout of five beta-strands and two helices. However, the position and conformation of the ssRNA at the surface of each ssRNA–RRM complex is highly variable, as well as the exact binding site on the protein (Figure 1). This diversity is due to three main reasons. First, in multi-RRM proteins, the inter-domain linkers present different lengths, conformations and orientations toward the domains. Yet, this linker participates in the RNA binding in most complexes. Second, the relative orientation of the RRMs in the complex is variable. Finally, despite a high structural conservation among the RRM domains, the sequence identity is quite low (typically 20–30%, see Supplementary Information Table S1) and similar in the binding region (Supplementary Information Figure S1). These variabilities explain the differences in RNA sequence as well as in RNA conformation at the surface of the protein. Therefore, a direct homology modeling or threading of the RNA based on a homologuous RNA–RRM complex is typically not possible.

Despite these variabilities, the RNA–RRM binding mode follows a common structural archetype defined by four characteristic contacts (19). Through an aromatic stacking interaction, one nucleotide is anchored to a conserved aromatic residue at RNP-1 position 5 (Figure 2), and another nucleotide is anchored by a similar interaction to an aromatic residue at RNP-2 position 2. The first anchor nucleotide is further stabilized by a hydrophobic interaction with RNP-1 position 3 and a salt bridge with a positively-charged side chain on RNP-1 position 1. These four conserved interactions can in principle be used as anchors to aid the modeling of unknown RNA–RRM complexes.

We present here a validated method to model RRM–ssRNA complexes from the experimentally known structure of the protein, the sequence of the RNA, and a few conserved specific stacking contacts in the RRM–ssRNA binding mode. Our fragment-based method performs iterative anchor-driven docking of fragments conformational ensembles, growing RNA sub-chains of fragments starting from the vicinities of the known anchor nucleotides. We applied this new protocol on a benchmark of seven non-redundant structures of ssRNA-RRM complexes. With this method, we could model the RNA chain with 3 Å resolution for all complexes but one, and with 2 Å resolution for all but two. For two proteins, an unbound form is available: we applied the method also on the unbound forms with success comparable to the bound proteins. This suggests that the method is effective even when the bound structure of the protein is not known, but only the relative orientation of the RRM domains.

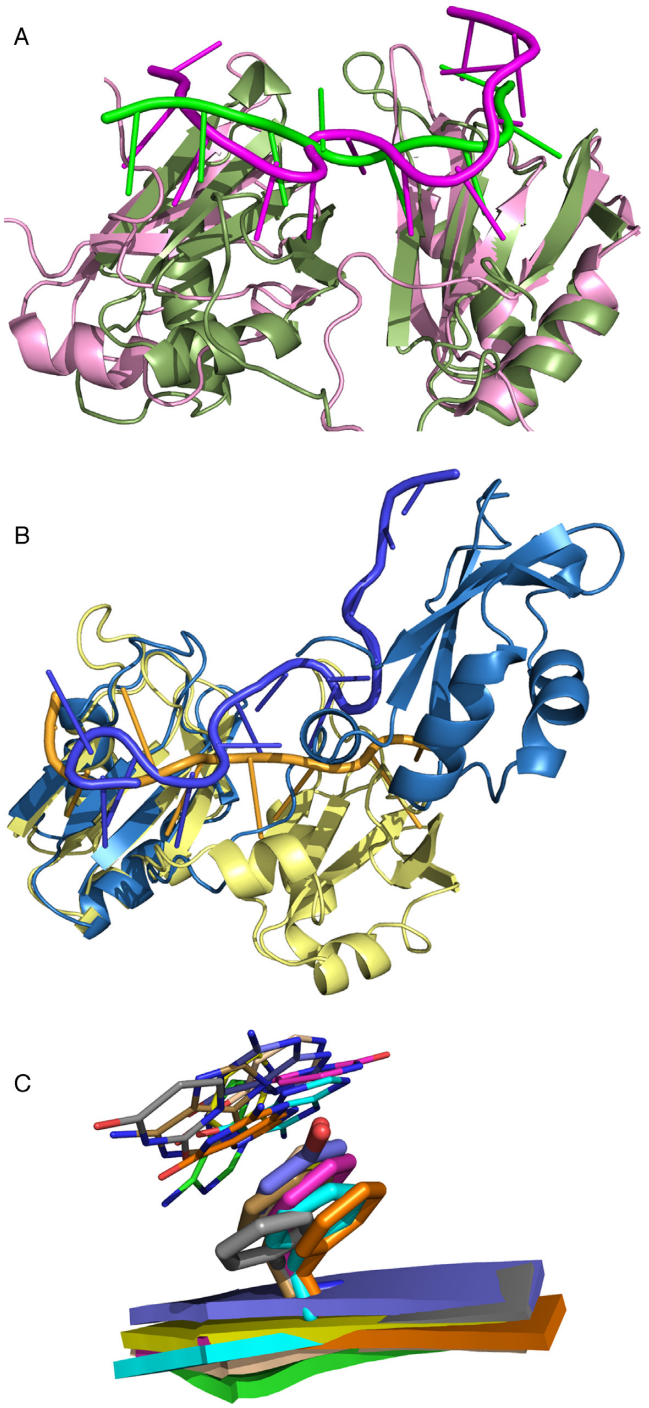

Figure 1. Diversity and conservation in structure and RNA binding mode of RRM domains. (A and B) Pairs of RNA-RRM complexes (see Table 1) represented as cartoons are superimposed on their N-terminal RRM, to illustrate the diversity in RRMs relative orientations in each protein and the large diversity in RNA binding mode. (**A**) Complexes with PDB code 2hy1 (green) and 2mgz (pink). (**B**) complexes 1cvj (yellow) and 4n0t (blue). For clarity, some coiled regions of the RRMs are not shown. (**C**) The N-term RRM of each of the RRM–RNA complexes from Table 1 are superimposed on their backbone. The β3 strand is represented as cartoon, with its conserved aromatic residue (sticks) that establishes a conserved stacking interaction with an RNA base (lines). The sugar and phosphate of the bound nucleotide are hidden for clarity.

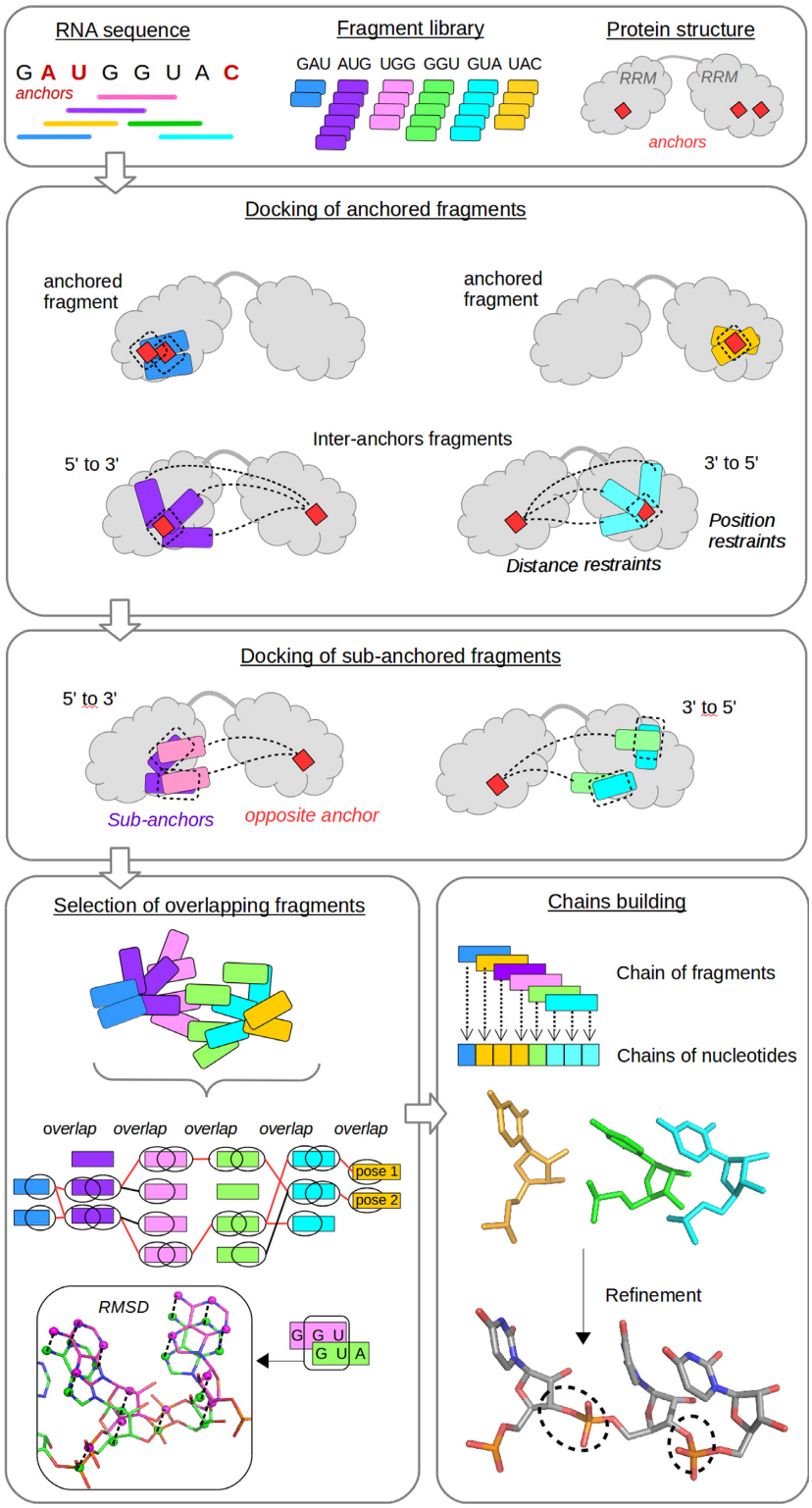

Figure 2. Flowchart illustrating the strategy for anchored ssRNA-protein docking with structural fragments. (1) The RNA sequence is cut in overlapping trinucleotides, and each trinucleotide is represented by an ensemble of conformers in an exhaustive fragment library. The structure of the protein is known, as well as contacts of the protein with 1 or 2 nucleotides per RRM, called 'anchors'. (2) Each conformer ensemble corresponding to a trinucleotide containing one or two anchor(s) is docked onto the protein. Weak position and distance restrains are applied toward the predicted anchor position of the contained anchor (to favour a correct position) and of the opposite anchor (to favour a correct orientation) respectively. (3) The docking poses are clustered and the 2 overlapping nucleotides of the most representative poses will be used as sub-anchors for docking the next fragment. Fragments are docked iteratively until the end of the chains (external fragments) or when two half-chains overlap (inter-anchor fragments). (4) The top-scored poses for each fragment are filtered by their capacity to form chains: pose-pose connectivity is determined by a strict overlap RMSD criterion. (5) Chains of connected poses are converted to chains of nucleotides, which are all-atom refined by energy minimization of the whole RNA-protein complex.

MATERIALS AND METHODS

General strategy

From the anchors, the docking of the first fragment explores the possible directions for the RNA chain at the surface of the protein. The positions energetically favoured, according to our scoring function, are retained. The docking of a second fragment attempts to place it in the prolongation of the retained positions of the previous fragment, and the new positions for the second fragment are scored. This process is repeated iteratively, until all the nucleotides in the RNA sequence have been docked. Between two anchors, the chains grown from each anchor in the direction of the other one have to be connected. This condition is fulfilled by measuring the overlap of the two extremities of the connecting chains.

Construction of a fragments library

We constructed a fragment-library of ssRNA trinucleotides extracted from all the available structures of RNA-protein complexes. The PDB was searched for X-ray structures and NMR structures of RNA–protein complexes containing at least three consecutive nucleotides resolved, without DNA or other large ligands at the RNA–protein interface. The main goal of this search was to collect all possible conformations to maximize the coverage of our library, even at the risk to include few non-realistic conformations from low-resolution complexes. The list of the 850 selected structures is provided as Supplementary Information Table S2. All trinucleotides where extracted from those structures and sorted by sequence.

To fill missing RNA atoms and to normalize modified nucleotides, we implemented an automatic completion and normalization procedure: (i) four libraries of mononucleotides, one per base, were built by extracting all the canonical and complete bases from the ssRNA-protein complexes, and clustering them at 0.5 Å ; (ii) a residue-names mapping pointed from each name of a modified base to the name of the chemically closest canonical base ; (iii) the abnormal absence or presence of atoms compared to the corresponding canonical base (unresolved or from modified nucleotide, respectively) was detected and the abnormal atoms were discarded ; (iv) the remaining atoms were superimposed to each conformer of the corresponding mononucleotide library ; (v) the closest nucleotide was selected and the missing atoms were extracted from this template.

To increase the number of conformations for each sub-library (i.e. each possible sequence), we mutated each conformer by all the possible purine => purine (Pu) and pyrimidine => pyrimidine (Py) mutations, in a three-step procedures. (i) All fragments composed of the same Pu/Py arrangement were pooled together and mutated into G/U. Each mutation consists of modifying one or two atom(s) from the base without changing the overall conformation of the fragment: atoms that were not common to the two purines/pyrimidines were discarded and replaced by the same procedure as for addition of missing atoms. (ii) Each ensemble of U/G–U/G–U/G conformers was clustered at 1 Å and only the center of each cluster was kept. (iii) The eight non-redundant U/G–U/G–U/G ensembles were then redistributed to all the corresponding Pu/Py sequences, by the same mutation protocol as described in (i). This procedure permitted to multiply by eight the total number of conformers in our library, ending up with 64 sub-libraries of 5000–11 000 non-redundant conformers. Finally, each ensemble was converted into ATTRACT coarse-grain representation.

Analysis of the library

The exhaustiveness of the library was tested by computing, for each fragment in our benchmark (*see next paragraph*), the all-atom RMSD between the bound fragment and the closest conformer in the corresponding sub-library, after withdrawal of the bound form if present. For each pair of sub-libraries with compatible sequences XYZ–YZQ, all pairwise overlap RMSDs were computed, and a binary matrix was created that stores the compatibility of two conformers according to an overlap cutoff of 3 Å RMSD. These compatibility matrices were used to reduce the ensemble of conformers to be docked for a sub-(sub)-anchored fragment on each anchor obtained by the previous docking.

Creation of a benchmark

Among all experimental structures of RNA-RRM complexes from the PDB (July 2015), we selected non-redundant complexes containing at least two RRM domains and 6 non-paired nucleotides bound to the protein. We discarded structures with resolution worse than 3 Å, and NMR structures with highly variable positions of most nucleotides among the different models. For each of the 7 resulting structures, we discarded the nucleotides with no contacts to the protein (4 Å cutoff). For NMR structures, we selected the model with the highest number of protein-RNA contacts, and we deleted the 5′- or 3′-terminal nucleotides that presented a highly variable position among the different NMR models.

The final benchmark contains 7 non-redundant ssRNA-protein complexes, among which four X-ray structures and three NMR structures (Table 1). The RRM of two different proteins share only 7–36 % sequence identity (average 22%, Supplementary Information Table S1), which is also representative for the RNA binding site (Supplementary Information Figure S1), and the relative orientation of RRMs in the multi-RRM complexes is highly variable (Figure 1), providing a very diverse set of test-cases.

Each complex contains two RRMs, each RRM containing one or two highly conserved aromatic residues (‘anchoring residues’) that each binds a nucleotide (‘anchored nucleotide’), and which position can be obtained from the homology to other RRMs.

Prior to any docking, we assessed for each fragment in our benchmark if the bound form of that fragment was one of the conformers of our library. For each fragment, the RMSD (all atom) of all conformers from the corresponding library ensemble towards the bound fragment was computed. Any library conformer with RMSD <0.1 Å was discarded and, if possible, replaced by the closest conformer in the same cluster, from the 1 Å clustering mentioned in a previous paragraph (*Construction of a fragment library*).

Table 1. Benchmark of RNA-RRM complexes structures

	Nb nucl.	Exp. Met.	NMR model	Resolution	Nucl. indexes	RNA sequence	Protein
1B7F	9	X-ray	–	2.60	3–12	U**G**UUUUUU**U**	Sex-lethal (Sxl) [h]
1CVJ	8	X-ray	–	2.60	1–8	A**A**AAAAA**A**	poly(A)-BP [h]
2MGZ	12	NMR	2	–	1–12	UGCA**UG**G**UG**U**GC**	RBFOX + sup-12 [h]
2YH1	8	NMR	1	–	601–608	U**U**UUUUU**U**	U2AF65 [h]
3NNH	10	X-ray	–	2.75	2–11	UUG**U**UUU**GU**U	CUG-BP 1 [h]
4BS2	10	NMR	1	–	1–10	GUGU**G**AA**UG**A	TDP-43 [h]
4N0T	13	X-ray	–	1.70	40–53	AAA**C**AAUAC**AG**AG	U6 snRNP [S]
3BX3	8	X-ray	–	3.00	1–8	**U**GUAUAU**A**	PUF-4 [S]
5BZV	9	X-ray	–	2.35	1–9	U**G**UACUAU**A**	PUF-5p [S]

Anchors are distinguished in bold. * The 4N0T complex contains three RMMs, among which only two establish canonical contacts with the RNA. X} or {X: paired nucleotides. [S]: Saccharomyces [h]: human [D]: Drosophila.

Statistical analysis of the position of the anchored nucleotides

We extracted all the structures of RRM-RNA complexes in the PDB (except for NMR structures with poor convergence of the models), resulting in 230 RRM–RNA anchoring interactions: 85 RNP-2/Py anchors, 17 RNP-2/Pu anchors, 27 RNP-1/Py anchors and 75 RNP-1/Pu anchors. Each of these four anchoring modes was analyzed separately. For the first three anchoring modes, all anchors were superimposed on the backbone atoms of a three-residue segment, consisting of the anchor residue and one on either side. The RNP-1/Pu anchors, which conformations are more diverse, were instead superimposed on the side chain of the anchoring residue. After this superposition, for each anchoring mode, all nucleotides were clustered with 2 Å cutoff (in coarse-grain). This resulted in two clusters for each pyrimidine anchoring mode and three clusters for each purine anchoring mode. The nucleotides in each cluster were averaged into an average anchor.

Prediction of the position of the anchored nucleotides

Based on these average anchors, we predicted the position of each anchored nucleotide from either the backbone conformation or the rotamer of the corresponding anchoring amino-acid. For simplicity, we assume knowledge of which anchor average is the correct one. In a fully blind situation, the anchor average can often be predicted from the anchoring rotamer (a direct correspondence exists, e.g. for the RNP-1/pyrimidine, results not shown) or from the surrounding rotamers (some anchor averages would induce clash and can therefore be discarded). If the position of an anchor is unclear, the docking could be repeated for each of the two or three anchor averages. The predicted positions were used as the basis for a weak positional restraint of the beads of the coarse-grain anchored nucleotide, defining a harmonic energy term that is zero as long as the bead is within a certain cutoff of the predicted position, and starts to rise slowly beyond that. The cutoff was adjusted according to the positional diversity of this bead in the corresponding cluster.

Docking: general settings

All docking was done with ATTRACT (20,21), using a coarse-grained protein–RNA force field (22). Both the bound protein and the fragment were in coarse-grained representation. Each pyrimidine/purine was represented by six or seven beads and each amino acid by three or four beads (20,22). If not otherwise specified, all RMSD are given in ATTRACT coarse-grained representation (22). The coarse-grained model defines almost exactly the position of all the atoms of the base (by two or three beads, the base being considered as planar and rigid), and exactly the position of the phosphate (by one bead). The positions of the other atoms are approximated by two beads located at the center of mass of two or three atoms. The receptor (protein) and ligand (RNA fragments) were considered as rigid during docking, the receptor was fixed and each ligand explored three rotational and three translational degrees of freedom.

When growing the RNA chain from the first to the second RRM (and *vice versa*), the fragments were docked with two types of restraints:

(i) Short-range restraints toward the corresponding anchors of the first RRM, or on the two 5′ nucleotides toward the poses of the previous fragment used as sub-anchors, to guide the position of the fragment. The restraints were applied to each bead of the anchoring nucleotide(s) toward the corresponding bead of the (sub-)anchor.
(ii) Long-range restraints of the 3′ nucleotide toward the closest anchor of the second RRM, to guide the direction of the fragment (Figure 2). The restraints were applied to the sugar bead GS2 of the 3′-terminal nucleotide towards the phosphate bead GP1 of the anchor.

When docking fragments containing nucleotides upstream from the first anchor or downstream from the last anchor, only short-range restraints were applied.

Docking modes and anchors

Different modes for each docking run were defined, with the docking protocol depending on the distance from the anchor (Figures 2 and 3). The easiest docking runs are those where the fragment directly overlaps with two or one anchor residue (anchored fragments). Within those docking runs, there are cases where the anchors correspond to the middle nucleotide of the fragment (mode AM for anchor-middle), to the first or three nucleotide of the fragment

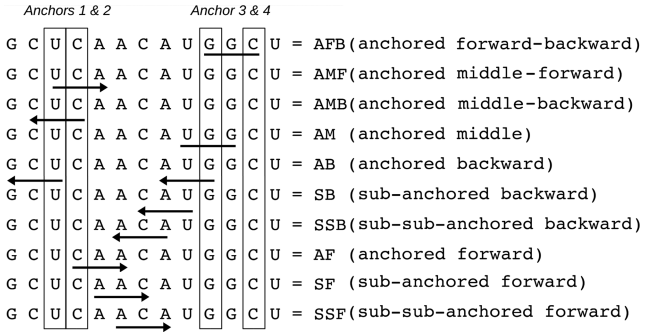

Figure 3. Terminology used for the different docking modes.

(docking forward or backward in the sequence: mode AF or AB), to both to middle nucleotide and the first/third (mode AMF/AMB), or to the two external nucleotides first and third (mode AFB) . The results of the these docking runs are used as sub-anchors to dock the adjacent fragments (sub-anchored fragments). Finally, the result of docking of the sub-anchor fragments can be used for the next adjacent fragment (sub-sub-anchored fragment). In the following sections, the docking protocols for anchored, sub-anchored and sub-sub-anchored fragments are detailed.

Anchored fragments

For each anchored fragment, 1 million random starting positions and orientations of the fragment towards the protein were produced by the 'randsearch' procedure of ATTRACT (21). The conformers from the ensemble of suitable sequence from our library were randomly distributed to these starting positions, resulting in only ∼100–200 starting positions per conformer. A first docking stage consisted of minimization until convergence without any protein–RNA force field, in order to orient each pose toward the anchor by distance restraints alone. The minimal distance for short-range restraint towards the predicted position of each bead of the anchor was set to 4 Å, to allow some diversity in the orientation and location of the resulting poses. A minimal distance for long-range restraint of the 3′ extremity of the fragment (sugar bead GS2 in ATTRACT coarse-grained representation) towards the predicted position of the opposite anchor (Figure 3) was extrapolated from the maximal possible length of a trinucleotide according to our fragment library, with 2 Å margin. In a second docking stage, 50 minimization steps were performed in ATTRACT force-field, the long-range pairwise interactions between ligand and receptor being approximated on a pre-calculated receptor grid. The distance to the current anchor was set to 2 Å, to account for a slight variability of the positioning of the anchoring nucleotides in different RRMs, with a harmonic constant of 20 kcal/mol/Å^2. A final re-scoring was performed without grid, pairwise interactions being considered until a distance of 7 Å.

The final poses were sorted by ATTRACT score, and the redundant poses (within 0.05 Å from a better scored pose) were discarded. The 10 000 top-scored poses were retained and clustered at 0.5 Å resolution to serve as sub-anchors for the next docking run. The clustering was based on the RMSD of only the two nucleotides to be used as 'sub-anchor' for docking of the next fragment (nucleotides 2 and 3 if docking from 5′ to 3′, nucleotides 1 and 2 if docking from 3′ to 5′). The two nucleotides from the top-scored pose of each cluster were used as a sub-anchors.

Sub-anchored fragments

The procedure to dock the sub-anchored fragments was globally similar to the previously described procedure for docking of the anchored fragment. But we used here an ensemble of receptors, each receptor containing the protein and one of the selected poses from the previous fragment to be used as sub-anchor. The previous docking run provided ∼5000–7000 sub-anchors for docking the sub-anchored fragments. The atom beads of the sub-anchor were turned into dummies ('ghost' beads) in the ATTRACT force-field and used as new reference for the position restraints. Those restraints were set to 0 Å for the first docking stage that orients the fragment without force-field, in order to preserve the information provided by the previous docking. Then the restraints were changed to 2 Å for the second docking stage, to enlarge the sampling and account for inaccuracies in the previous docking run.

When finally assembling the poses into chains, we will only keep poses of consecutive fragments with low overlap RMSD of the two common nucleotides (Figure 2). Two conformers XYZ and YZQ which common nucleotides YZ have highly different structures could never be associated in a chain. Therefore, we did not dock all combinations {sub-anchors XYZ—conformers YZQ}, but only compatible conformers, i.e. with a minimal overlap RMSD under 3 Å after fitting. Such compatibility was tested for all possible pairs in our library. This filtering retained between 37% and 54% of XYZQ combinations, depending on the sequence. The total number of starting positions was set to 30 million, to which were distributed the {sub-sub-anchor, conformer to dock} combinations, in a way that harmonizes in priority the number of starting positions per conformer, then the number of starting positions per sub-sub-anchor.

(Sub-)sub-sub-anchored fragments

The diversity in the poses after docking increases with the distance of the docked fragment toward the initial anchors (anchored < sub-anchored < sub-sub-anchored . . .). Therefore, the number of poses to keep and the clustering cutoff to apply were adapted accordingly. According to the ranks of good solutions in the previous docking run and to some testing of clustering parameters, the 10^5 top-scored poses from SF/B mode docking were retained and clustered at 1 Å. The top-scored poses in each of the 10 000 first clusters were used as sub-anchors for SSF/B docking, and the total number of starting positions was set to 45 million.

Unbound docking

The structures of PDB code 2YH0 (23) and 3SXL (24) were used as the unbound form of the U2AF65 and sex-lethal protein present in complexes 2YH1 (23) and 1B7F (25) respectively. The unbound protein 3SXL contains three

seleno-methionines, which were converted into methionines in order to fit the 1B7F sequence, and residues 203–209 (3SXL) and 232–257, 237–242 (2YH0) of the coiled inter-RRMs linker and/or C-term parts were removed. For each protein, each unbound RRM was fitted on the corresponding bound RRM with PyMOL (26). The third anchor (F170) is located in close vicinity (<5 Å) of the flexible hinge. Even if not knowing the bound form, one could easily anticipate contacts of the anchored fragment with the hinge, which would not be reproduced in the unbound docking where the hinge was removed. This difficulty was addressed by increasing the number of starting positions for the docking of that particular anchored fragment from 1 to 10 million. The rest of the protocol was identical to the protocol for docking with the bound protein.

Assembly of fragment-chains

The top-scored poses were selected from each docking. The number of docking poses kept for chain building was adjusted to the distance of the fragment to an anchor, the sampling becoming more difficult when the distance increases. Test runs indicated that selecting 10^3 (ASF and AM runs), 2×10^3 (AF/AB runs) and 2×10^6 (AF/AB external runs and S(S)F/S(S)B runs) top-scored poses provided a good compromise between accuracy of the best solutions and number of decoys. Additionally, at least one of two AF/AB non-external runs gave good scored poses for each complex, thanks to at least one of the two RRMs having very accurately predicted anchor position(s). Therefore, and to keep as low as possible the number of pose*pose combinations to test, we pooled together chains obtained by combining $2 \times 10^3 \times 2 \times 10^3$, $400 \times 1 \times 10^4$ or $1 \times 10^4 \times 400$ poses for AF/AB runs (leading to 4×10^6 combinations tested for each pattern). For a few fragments, results come from two docking runs at each extremity of two half-chains grown from both anchors (runs SF and SB, or SSF and SSB) (Figure 2). In those cases, the results from the two docking runs were pooled together prior to clustering.

Compatible poses of adjacent fragments were selected according to the coarse-grain RMSD of their two common nucleotides, within an overlap-cutoff of 1.4 Å. The poses were then arranged into full chains of fragments, by joining pairs *i–j* and *j–k* into *i–j–k* triplets, etc. To accelerate the procedure, the poses were clustered by RMSD with different clustering cutoffs, and comparison was made with adapted overlap-cutoff at each clustering level. If necessary, the overlap-cutoff was iteratively increased by 0.1 Å until forming at least 3000 nucleotide-chains (see below), or to a maximum of 2.5 Å. The fragment poses of each docking run being ranked by their ATTRACT score, we computed for each chain the geometric mean of the ranks of the poses and kept the chains with the lowest mean rank. The choice of a geometric over an arithmetic mean aims at allowing an energetically unfavorable position of a fragment if it allows another fragment to reach a highly favorable position.

From fragment-chains to nucleotide-chains

In the chains made of fragments (fragment-chains), each residue is represented by one, two or three nucleotide(s) belonging to overlapping poses of fragments (Figure 4). The fragments were merged by retaining for each residue only the nucleotide belonging to the fragment the closest to a mid-anchor, in the order: AFB > AMF/B > AM > AB/F > SB/F > SSB/F. In case of indecision, we selected the fragment the most converged (i.e. showing the lowest diversity of poses for that position in the chains). For some complexes, the nucleotides are extracted only from a subset of the fragments. For instance, in 3NNH, the fragments come from docking runs AM-AF-SF-SB-AB-AMB for fragments 3 to 8, respectively. The nucleotides will be extracted from the poses of fragment 1 (first to third nucl), fragment 2 (fourth nucl), fragment 5 (fifth nucl) and fragment 6 (sixth to eighth nucl). The poses from fragments 3 and 4 are thus not used in the final building of the chain, but served at selecting connecting poses for the other fragments. Consequently, two fragment-chains differing only by the pose(s) of fragment(s) 3 and/or 4 result in the same chain of nucleotides (nucleotide-chain), and the number of nucleotide-chains is inferior or equal to the number of fragment-chains. The ratio fragment-chains / nucleotide-chains is highly dependent on the docking scheme from each complex (i.e. the positions of the anchors).

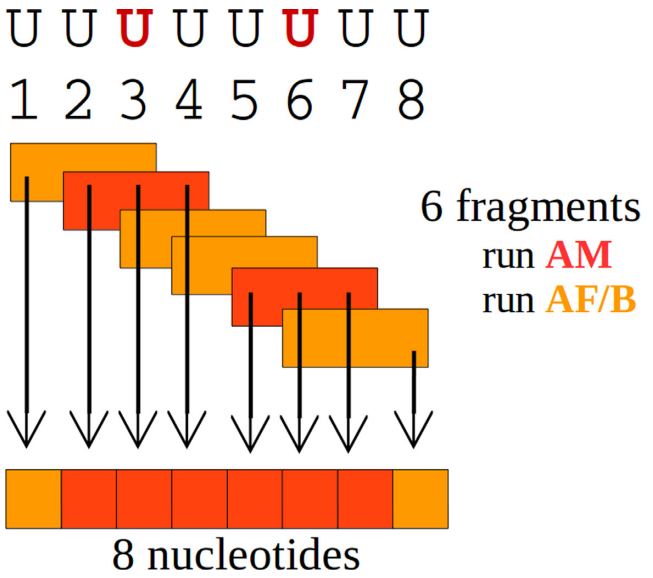

Figure 4. Schematic representation of fragment-chains to nucleotide-chains conversion, by selecting the nucleotide from fragments close to an anchor.

Each nucleotide-chain was sorted by the mean rank of the fragment-chain from which it was extracted. If the same nucleotide-chain was extracted from different fragment-chains, only the lowest mean rank was taken into account. The 1000 top-ranked nucleotide-chains were retained to be converted into all-atom representation.

All-atom refinement

The nucleotide-chains are made of disconnected coarse-grained nucleotides. To be transformed into a realistic model, each chain was converted from ATTRACT coarse-grained representation to all-atom representation *via* our mono-nucleotides library. Each nucleotide in the chain was

compared to all the conformers in our mono-nucleotides coarse-grain library by computing the RMSD after fitting. The conformer with lowest RMSD was kept and the corresponding conformer in our mono-nucleotides all-atom library was selected. The 1000 top-ranked nucleotide chains were then clustered at 0.5 Å. The quality of the models was assessed by computing the all-atom RMSD toward the bound form, as well as the fraction of native contacts (according to the CAPRI criterion (27)).

For each anchor-to-anchor chain in each complex, the 1000 all-atoms nucleotide-chain were then minimized with AMBER, in order to resolve clashes due to coarse-grain to all-atom conversion and to connect the nucleotides. The hydrogens were added to the protein and RNA of each all-atoms model using the Leap module of AMBER12 with leaprc.ff10 parameters. The models were minimized using the sander module of AMBER12, by 400 steps of steepest descent followed by 1000 steps of conjugate gradient, with a non-bonded cutoff a 12 Å, the non-bonded list updated every five steps, the initial step length set to 0.0001 Å, and other sander parameters as defaults. All protein atoms were restrained with a harmonic constant of 1000 kcal/mol/Å^2. The final models were clustered at 0.5 Å and evaluated by computing the RMSD on all heavy atoms of the RNA compared to the experimental structure, after fitting the protein.

Docking on the pumilio domain

To assess if our method is applicable to other RNA-binding domains than the RRM, we chose as an example the Pumilio domain (PUF) and we selected the test structures of PDB code 3BXV (28) and 5BVZ (29) containing seven and eight bound nucleotides respectively. To create conditions similar to our study of the RRM domain, we considered that the binding mode of the PUF domain is unknown but for the position of two terminal nucleotides which we used as anchors. As there are only 25 highly redundant structures of PUF–RNA complex in the PDB, we did not perform a statistical analysis of the anchors position as for RRM–RNA domains, but predicted the positions of the anchors by homology. For each complex, we supposed the structure of the other complex as known, we superimposed the two proteins, and we took the position of the corresponding nucleotides in the known complex as reference positions for the anchors of the unknown complex. To account for protein deformation between the two complexes, we superimposed separately part of the complex for each anchor, selected as the consecutive residues including all the residues located at 12 Å from the anchor in the known complex. The docking and chain-forming procedures were strictly identical to for the RRM-RNA complexes.

Computation times

Each fragment docking took between a few minutes (anchored fragments) and ∼8 h (sub-sub-anchored fragments) on eight CPUs, depending on the number of starting positions and the number of distance restraints. The clustering of the docking poses took a few seconds (10,000 poses) to a few minutes (100 000 poses) on one CPU. The selection of chains-forming fragments took a dozen of minutes (five anchored fragments) to ∼24 h (10 fragments) on one CPU. The conversion from fragment- to nucleotide-chains took a dozen of minutes on one CPU. The all-atom refinement of 1000 chains took ∼1 h on 10 CPUs, and the analysis of the results ∼$\frac{1}{2}$ h on one CPU. Globally, the full RNA modeling process took between ∼65 CPU hours (complex 2yh1) and ∼350 CPU hours (complex 2mgz), depending on the length of the chain and the position of the anchored nucleotides in the sequence.

RESULTS

General strategy

We developed a new iterative protocol to model a RRM–bound RNA, starting from a few (three or four) anchor nucleotides that establish conserved contacts with the protein. Using each (pair of) anchor nucleotide(s) as a seed, we grow the RNA chain in either direction (5′ or 3′) by adding a trinucleotide fragment. For each fragment, a large number of candidate poses are obtained by docking with ATTRACT with a position (overlap) restraint of the extremity of the fragment towards the seed anchor(s), as well as a distance restraint towards the nearest non-seed anchor. The conformations of the fragments are provided by a conformational library that we built by extracting all trinucleotides from the published experimental structures of protein-RNA complexes. The docking process is iterative: the best candidate poses are kept and used as sub-anchors for a subsequent docking run to obtain candidate poses for the next fragment. The complete fragment-based docking protocol is illustrated in Figure 2. Once all fragments in the sequence have been docked, pairs of spatially overlapping fragments are identified and merged into overlapping chains. Finally, the overlapping chains are converted and refined into single-nucleotide chains (Figures 2 and 3).

Identification of anchor nucleotides

Throughout the study, we have used one or two anchor nucleotide(s) per RRM, assuming knowledge of their position within the RNA sequence. The RNA–RRM binding mode contains four conserved interactions involving two anchor nucleotides, but they are not always present together (19). Still, the aromatic residues RNP-1 position 5 and RNP-2 position 2 are particularly conserved. We verified this by a sequence analysis of the RRM Pfam (30) family (PF00076, RP15 multiple sequence alignment, October 2015, analyzing all 11 159 sequences without gaps at either position), showing that in 76.1 % of the RRMs, both of these residues are aromatic, and in 93.8 %, at least one of them is. To define the anchor during docking, we determined its most probable position from the rotamer of the anchoring residue, by the analysis of all known structures of RRM-RNA complexes. We verified that in our benchmark, the presence of an aromatic residue at that position indeed indicates the presence of an anchor nucleotide. This was true for all 25 anchors (see Supplementary Information Table S3), except for 3NNH RNP-1. In that case, however, the anchoring residue has a highly unusual rotamer and the predicted anchor nucleotide would heavily clash with the protein (results not

shown). For an eighth complex, 4N0T, the anchoring nucleotide is in a non-canonical flipped state but since this is not apparent from the rotamer, the anchor was kept.

With this strategy, a total of 24 anchoring nucleotides (three or four per complex) could be predicted for the seven complexes in the benchmark, with a precision of 0.8–3.2 Å RMSD (average 1.4 Å) in coarse-grain representation.

Completeness of the fragment library

We constructed a fragment-library of ssRNA trinucleotides extracted from all the available structures of RNA-protein complexes in the PDB. To model correctly the RNA in our benchmark, the library must contain for each bound RNA fragment at least one conformer close to the bound conformation after superposition. Among the 57 bound fragments in our benchmark (∼8 fragments per complex), 79% were approximated by <1.0 Å toward the bound form (Supplementary Information Table S4), and only one was best approximated by conformers with RMSD above 2.0 Å (4N0T frag. 11). Apart for that fragment, we presumed that this accuracy was sufficient to model all the test complexes.

Docking of the anchored fragments

Anchored fragments consist of one/two anchor nucleotide(s) with a position predicted by homology and two/one nucleotide(s) of unknown position. The anchored docking runs were successful for most fragments (Supplementary Information Table S5). It resulted in hits (RMSD ≤ 2.0 Å) for 32 out of 42 anchored fragments, and near-hits (RMSD ≤ 3.0 Å) for 9 of the 10 other fragments. The lower sampling quality for five fragments in complexes 4N0T and 1CVJ (lowest RMSD in [2.2 Å ; 3.1 Å]) is due to the imprecision of the library, the closest conformer in the library being at 1.5–2.2 Å RMSD when fitted on the bound form, instead of 0.3–1.4 Å for the other anchored fragments (Supplementary Information Table S5). The lower sampling quality for three fragments in complex 4BS2 is probably due to the error on the predicted position of the only anchor available for those fragments.

Docking of the sub-anchored fragments

In the next docking runs, we predict the position of three unanchored nucleotides. The position of two nucleotides are only approximated by the previous docking runs, which implies an accumulation of errors. The increased difficulty was addressed by increasing the sampling (i.e. the number of docking poses). The top-scored poses from each AF/B docking run were clustered to be used as sub-anchors for the docking of the adjacent fragments. The SF/B docking produced hits for 7 of the 11 sub-anchored fragments, near-hits for all but one fragment (3NNH frag.1), and good solutions (RMSD < 4 Å) for all fragments. The absence of hits in five cases correlates with low quality of the sub-anchors (RMSD > 1.8 Å). More generally, for all (sub)-sub-anchored runs, the RMSDs of the closest pose are correlated with the quality of the sub-anchor (correlation coefficient 0.77) rather than with the quality of the library (Figure 5).

In most cases, the closest docking poses had a lower RMSD than the corresponding sub-anchor, revealing the convergence of the poses toward the bound form after divergence of the poses from the previously docked adjacent fragment.

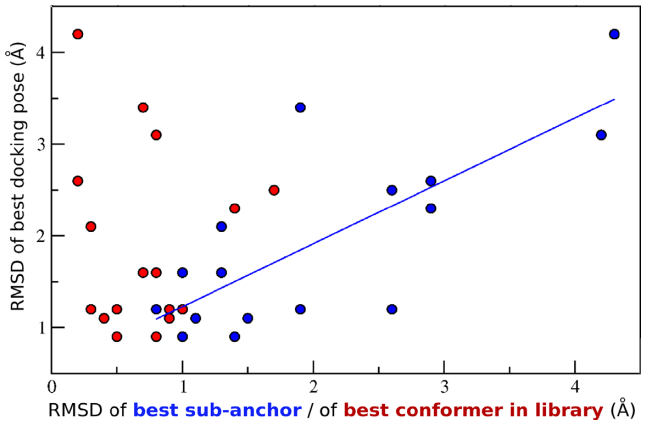

Figure 5. Impact of the quality of the library and of the sub-anchors on the sampling for SF/B docking modes.

Docking of the (sub)-sub-sub-anchored fragments

In those runs, the docked fragments are separated from the initial anchors by one or two nucleotides which position has been predicted. Nonetheless, the results still show reasonable sampling precision for most runs, with ∼120–270 hits found for three out of six docking runs (Supplementary Information Tables S6 and S7). No near-hits were found for docking runs {4BS2 fragment 1} and {4N0T fragment 6—docking mode SSF}. Those failures are again due to the imprecisions of the sub-anchors (best at 4.2 or 4.3 Å RMSD). But fortunately, one is compensated by near-hits found for the same fragment 6 of 4N0T by docking from the other anchor (mode SSB, fragment 7 used as sub-anchor). After those last runs, the two half-chains grown from each anchor covered the full sequence between the two anchors, with some overlap.

Selection of chains-forming fragments

The top-scored poses obtained for the fragments with overlapping sequence (Supplementary Information Table S7) were joined into chains of spatially overlapping poses, by assessing the overlap-RMSD of poses for each pair of adjacent fragments (Figure 3). All poses of one fragment A were compared to all poses of the adjacent fragment B, and pairs A_i–B_j with RMSD of the two overlapping nucleotides under a chosen overlap-cutoff were kept. The poses of fragment B belonging to an A_i–B_j pair where then compared to the poses of the next adjacent fragment C, and so on, until the end of the chain was reached. The presence of a weakly bound fragment in an RNA chain can be explained by the necessity to accommodate for an adjacent 'hotspot-binding' fragment with a highly favorable binding energy. Taking ATTRACT ranking of a pose as indicator of its binding energy, we considered that a pose can have a bad rank if compensated by good ranks of other poses in the chain. We thus computed for each chain the geometric mean of the ranks

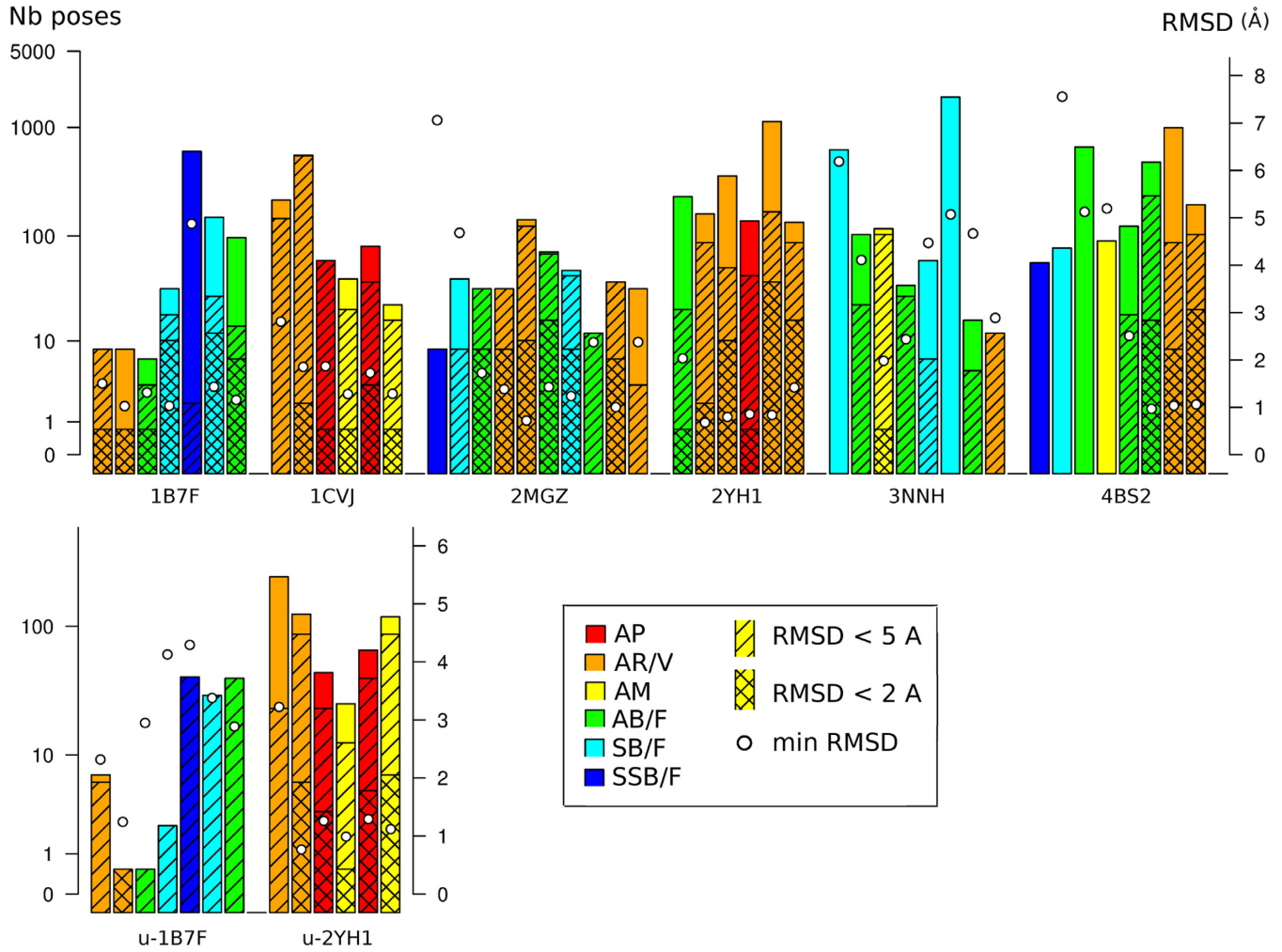

Figure 6. Results per fragment of the selection of chain-forming poses in the 1000 top-ranked chains. Each of the nine bar-plots represents one complex. For each complex, each fragment is represented by one colored bar. The height represents the number of poses (log scale) and the color describes the docking mode. The percentage of poses under 2 or 5 Å RMSD is represented, respectively, by the black and dashed areas of the bar. The RMSD of the closest pose is represented by a white dot. Results of unbound (u-) docking are presented in the lower panel.

(mean rank) of the poses, and kept the 3000 chains of lowest mean rank. Finally we assessed for each fragment the total number of poses kept, and the number of hits among those poses. Results per fragment are presented in Figure 6. The effectiveness of the poses filtering by chain-forming criteria is illustrated by Figure 7.

Out of 45 fragments, 39 were sampled with accuracy better than 5 Å RMSD, and most of them with accuracy better than 2 Å (27/45 fragments). The length of the chain does not have a major impact on the quality of the results at the fragment level: complex 2MGZ (10 fragments) has most fragments very well sampled at 2 Å accuracy, whereas complex 3NNH displays much worse sampling, with only three fragments with an accuracy better than 4 Å. Anchored fragments, especially containing two anchors, are in general better sampled than others, as expected.

From fragment-chains to nucleotide-chains

After assembling fragments into fragment-chains, two steps are still needed to convert the chains into models of an ssRNA strand. First, the overlaps between adjacent fragments have to be resolved and the fragments merged into chains of single nucleotides (nucleotide-chains). Second, the result of this merging process is a chain of disconnected nucleotides, which have to be connected in a sterically correct way.

In chains of overlapping fragments, each nucleotide is represented by one to three nucleotides from different fragments. We tested three procedures to merge the fragments. For each residue we kept the nucleotide either (i) with the lowest ATTRACT score after re-scoring the single nucleotides from each pose on the protein, or (ii) from the fragment with the lowest diversity of the poses, considering the convergence of the docking as an indicator of the accuracy of the poses, or (iii) from the closest fragment to two or one anchor(s) (AFB if present, either AMF/B if present, either AM, etc.) (Figure 4). The last procedure gave the best results in terms of percentage of good chains (data not shown), and was further applied. The results obtained after fragment-chains to nucleotide-chains conversion, all-atom conversion

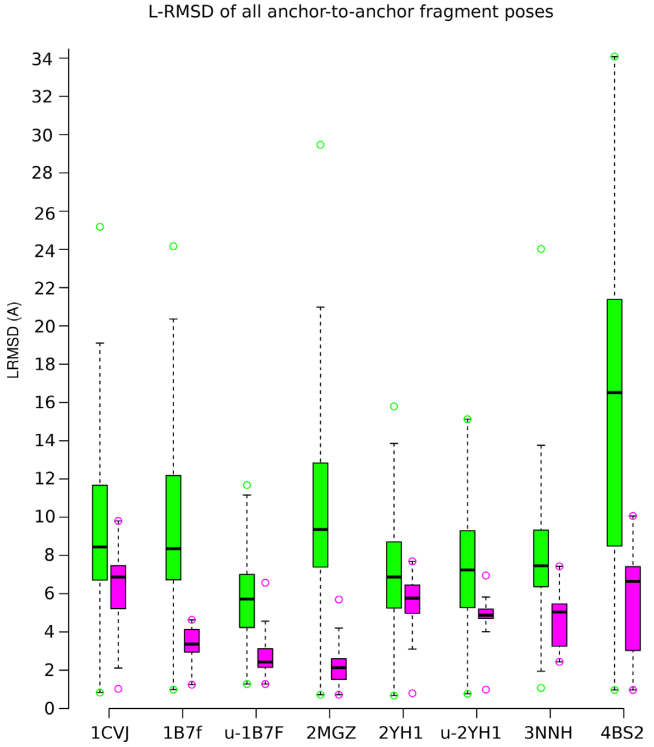

Figure 7. Effect of selection of chain-forming poses on the poses quality. For each complex are represented as box plots the L-RMSD of the poses for all fragments together, before (green) and after (magenta) selection of chain-forming poses. The best and worst poses are represented by circles.

and clustering at 0.5 Å, are presented in Table 2 and Figures 8 and 9. Results obtained before and after all-atom refinement are compared in Supplementary Information Table S8.

Chain building and refinement

When growing the RNA chain from the first to the second (pair of) anchor(s), the fragments are docked with long distance restraints of the 3′ extremity toward the second anchor, to guide the direction of the fragment. In contrast, when docking 'external fragments' upstream of the first anchor or downstream of the second anchor, no such long distance restraints can be applied (Figure 3). The sampling is then more difficult, which impacts also the chain building with external fragments. We built chains step by step with increasing difficulty: first from first anchor to last anchor, then with one external nucleotide on each side (if existing), then two external nucleotides, etc.

The building and refinement of nucleotide-chains from anchor to anchor allowed to retain models with the bound RNA modeled with a precision under 2 Å (all-atoms RMSD) with respect to the native structure for all but one complexes, and under 3 Å for all complexes (after removal of complex 4N0T). The precision of the best model in the 100 top-ranked models was even better than 2 Å for four of the six complexes. Moreover, for each complex, between 1 and 98% of the proposed models were good (RMSD $\leq$ 3 Å), and between 16 and 99 % were correct (RMSD $\leq$ 4 Å), with 54 % correct models on average. Even more impressive are the fractions of native contacts (fnat: nucleotide-amino acid pairs with at least two atoms at less than 5 Å from each other) that are reproduced in our models. All models of all complex have fnat higher than 0.50, between 13 and 52% of the models higher than 0.80, with a maximal fnat ranging from 0.84 to 0.98.

When growing the chains with one external nucleotide at each anchor (if present in the bound complex), the overall quality of the results was only slightly affected. The percentage of good or correct models was even improved in some cases, and the precision of the top-ranked model in two cases (3NNH, 1B7F). The quality of the results in term of ligand RMSD decreases when adding three or four external nucleotides, yet all models keep fnat >0.50, due to mis-orientation of nucleotides but at a correct position. Finally, we could model a 10-nucleotide chain with a precision of 2.1 Å, and 11- and 12-nucleotides chain with a precision of 2.5–4.6 Å. The success of our method is particularly remarkable for the long 11–nucleotides chains of 2MGZ (nucl. 2–12). In this case, the first nucleotide is sampled only by the first fragment, corresponding to a sub-anchored external fragment, i.e. remote from the first anchor and with no opposite anchor to orient the fragment correctly (Figure 3). Nonetheless, accurate models could still be sampled for the whole chain, with the first nucleotide modeled at 3.1 Å precision.

The refinement by minimization with AMBER proved effective to improve the quality of most of the models in most complexes. Between 50 and 94% of the models showed a lower RMSD after refinement, with improvements up to 3.9 Å. The clustering reduced the number of models from 1000 to between 11 and 217, depending on the complex (Supplementary Information Table S8). The number of clusters did not correlate with the length of the chain.

Docking with unbound RRM domains in the bound relative orientation

Up to now, we used the bound form of the protein in all docking runs. The use of an unbound form of the protein introduces additional difficulties in the method, due to misplacement of side chains, varied protein loops conformation (in particular the inter-domain linker) and changes in orientation of the RMM domains. However, we hypothesized that the side chains and loops flexibility would be partly neutralized by our coarse-grain representation, smoothing the surface of the protein, and the weakness of our anchor restraints. This assumption is valid for small-amplitude flexibility, which seems compatible with the relatively high rigidity of RRM structures. To test this hypothesis, we performed unbound docking on two complexes. Each protein is made of two RRMs linked by a flexible hinge of ∼10–30 residues. Each RRM domain shows very similar structures in the bound and unbound form, with similar orientations of the side chains, especially at the binding site. Yet the orientation of the two domains is very different (Supplementary Information Figure S2). Moreover, in the sex -lethal protein (bound complex 1B7F), the flexible linker is partially unresolved in the unbound form, and the resolved parts do not superimpose onto the bound form. In this study, we considered the orientation of the two RRM

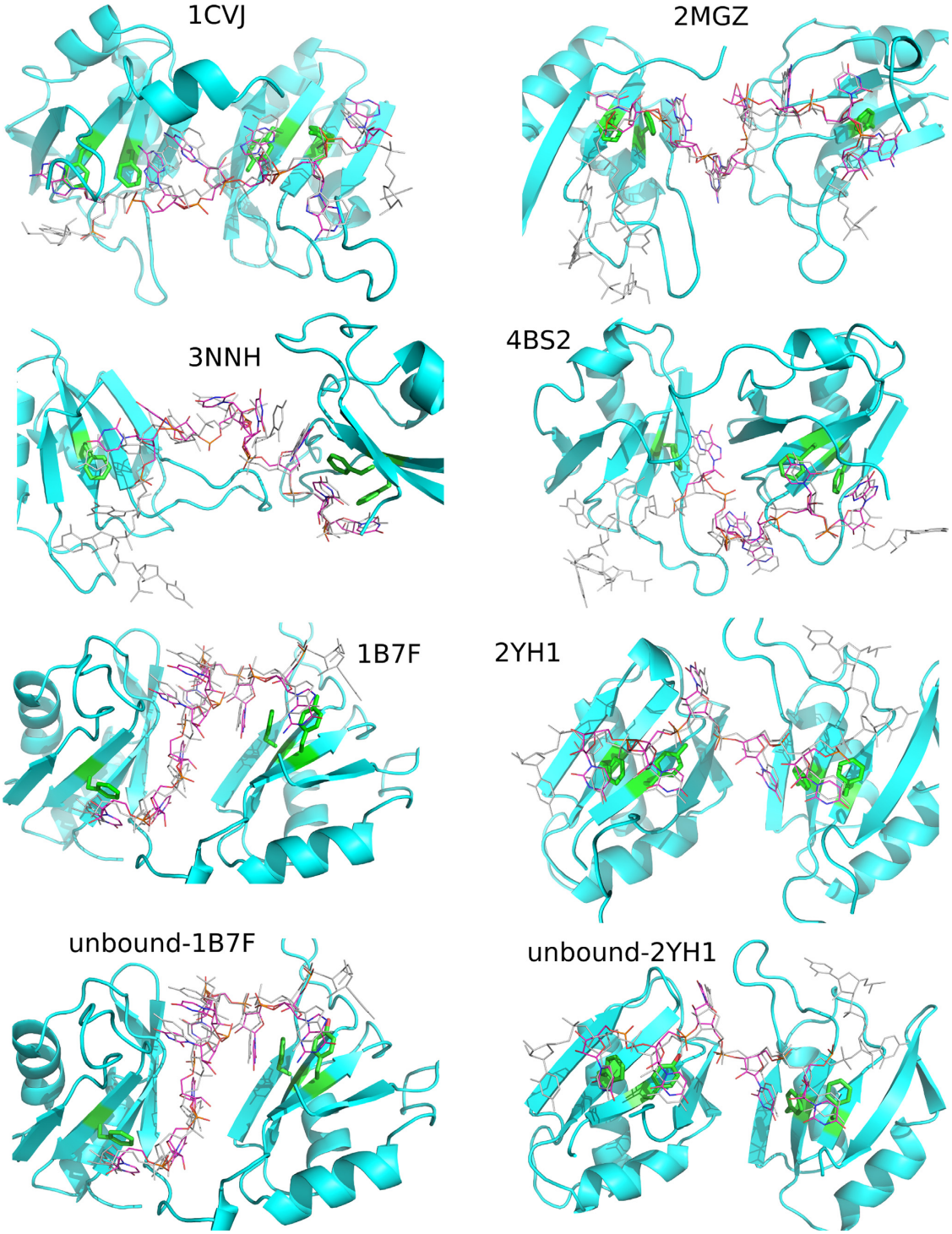

Figure 8. Best solution (min RMSD) for anchor-to-anchor chains obtained for each RRM–RNA complex after all-atom conversion and clustering. The protein is represented as cyan cartoon, with each anchor as green sticks. The bound and modeled RNA are represented in white and magenta sticks respectively.

Table 2. Results obtained by chain building, fragment-chain to nucleotide-chain conversion, coarse-grain to all-atoms conversion and clustering

	Nucl.	Min RMSD (Å) [Fnat (%)] in			% poses with RMSD			% poses with Fnat			Nb chains
		Top 1	Top 10	Top 100	≤2Å	≤3 Å	≤4 Å	≥80%	≥75%	≥50%	
{Anchor-to-anchor}											
1b7f	2-9	**3.6** [79]	3.4 [85]	2.1 [90]	1	13	**50**	23	52	100	250
1cvj	2-7	**2.0** [75]	**1.8** [83]	**1.8** [84]	**21**	**98**	**98**	13	70	100	224
2mgz	5-11	**2.2** [88]	**2.2** [92]	**1.6** [98]	10	27	**99**	52	97	100	182
2yh1	3-7	**1.2** [98]	**1.2** [98]	**1.2** [98]	5	8	36	19	35	100	132
3nnh	4-10	4.5 [78]	**3.7** [90]	**3.4** [90]	0	**1**	26	15	68	100	151
4bs2	5-9	**2.7** [90]	**1.8** [98]	**1.8** [98]	**1**	**11**	**16**	17	42	100	272
{Anchor-to-anchor} + 1 nucleotide											
1b7f	1-9	**3.6** [73]	**1.6** [90]	**1.4** [90]	**7**	**14**	32	15	17	100	107
1cvj	1-8	2.7 [75]	2.7 [81]	2.5 [83]	0	73	90	18	82	100	236
2mgz	4-12	2.3 [88]	2.3 [91]	2.1 [92]	0	**41**	96	86	100	100	182
2yh1	2-8	2.2 [95]	2.0 [98]	2.0 [98]	**22**	**98**	**100**	100	100	100	45
3nnh	3-10	**4.3** [75]	**3.7** [85]	3.7 [90]	0	0	**50**	37	76	100	62
4bs2	4-10	4.7 [72]	4.6 [77]	4.0 [84]	0	0	1	9	69	100	110
{Anchor-to-anchor} + two nucleotides											
2mgz	3-12	2.3 [88]	**2.2** [91]	2.1 [92]	0	**47**	97	85	100	100	146
2yh1	1-8	4.7 [82]	1.8 [97]	1.8 [97]	7	14	45	64	100	100	58
3nnh	2-10	>5 [73]	4.4 [81]	4.4 [89]	0	0	1	30	80	100	93
4bs2	3-10	>5 [70]	> 5 [73]	> 5 [77]	0	0	0	0	3	100	121
{Anchor-to-anchor} + three nucleotides											
2mgz	2-12	3.1 [80]	3.1 [81]	2.5 [86]	0	4	73	42	94	100	124
3nnh	1-10	>5 [75]	5.0 [81]	5.0 [84]	0	0	0	25	96	100	51
4bs2	2-10	>5 [72]	>5 [75]	>5 [76]	0	0	0	0	28	100	50
{Anchor-to-anchor} + 4 nucleotides											
2mgz	1-12	3.7 [77]	3.7 [78]	3.7 [80]	0	0	74	0	76	100	38
4bs2	1-10	> 5 [68]	> 5 [69]	> 5 [70]	0	0	0	0	0	100	23
Docking on unbound protein											
1b7f	2-9	3.6 [79]	3.4 [85]	2.1 [90]	2	12	52	25	54	100	215
	1-9	**3.5** [73]	**1.6** [90]	**1.6** [90]	**8**	**20**	**55**	22	27	100	83
2yh1	3-7	**4.1** [83]	**1.2** [100]	**1.2** [100]	**11**	**33**	**64**	63	94	100	126
	2-7	4.6 [76]	3.2 [83]	2.4 [91]	0	5	34	25	66	100	411
	1-7	> 5 [71]	3.8 [80]	3.1 [89]	0	0	7	7	23	100	393
Docking on a pumilio domain											
3bx3	1-6	2.7 [97]	1.9 [98]	1.9 [100]	40	100	100	100	100	100	20
5bzv	2-7	2.6 [88]	2.2 [93]	2.2 [95]	0	97	100	100	100	100	31
	1-7	2.2 [95]	1.9 [95]	1.9 [95]	36	100	100	100	100	100	14

For each complex, the best results among the different chain lengths are in bold. Chains for which less than 1000 fragment-chains could be assembled are underlined.
RMSD and Fnat are computed in all-atom resolution, Fnat was calculated with a 5 Å cutoff.

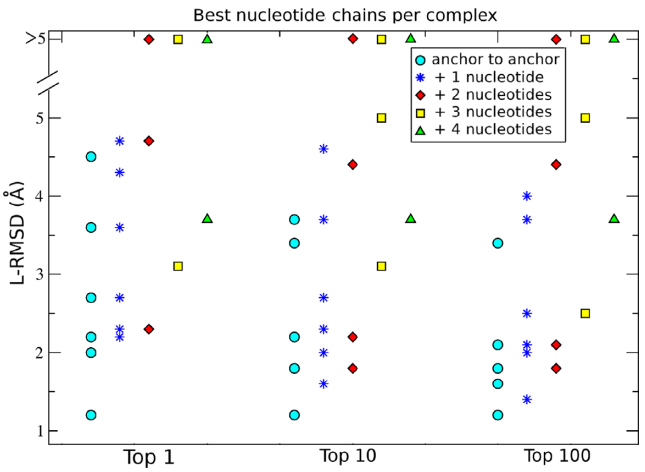

Figure 9. L-RMSD of the best solution of bound docking per RRM complex, for chains of different lengths.

as experimentally known, and hence we superimposed each unbound domain onto the bound domain prior to docking. The resolved parts of the linker were removed. After this procedure, the two proteins have an all heavy-atom RMSD of 1.7–1.5 Å, both globally and at the RNA binding site (5 Å cutoff), for 1B7F and 2YH1 respectively.

Predictions of the position of the anchor nucleotides based on the rotamers of the conserved motifs in the unbound form reached an RMSD towards the real position quite close (−0.4 Å to +1.2 Å) to what was observed for predictions based on the bound form. This results of the high conservation of those rotamers in the bound and unbound form. The docking of the fragments sampled near-native and closest-to-native solutions with a resolution comparable to the bound docking: we could sample all fragments with 2 Å precision (Table 3). The difference in lowest RMSD for each fragment between bound and unbound docking ranged from −0.9 to +0.5 Å. The number of hits compared to what was obtained by bound docking is higher for 1B7F, whereas it remains almost constant for 2YH1. The anticipation of increased difficulty for docking on the unbound structure had led us to increase the sampling for the anchors by a factor 10. This changes seems to overcompensate the inaccuracy induced by the unbound form, and indicates that higher sampling would probably lead to more accurate models for the other complexes as well.

Chain assembling with unbound domains

The poses were assembled in 1000 anchor-to-anchor chains, among which 4–13% approximated the RNA within 2 Å RMSD in 1B7F and 2YH1 respectively, the best model reaching 1.6 – 1.0 Å RMSD respectively. Clustering the all-atoms chains at 0.5 Å reduced the number of models from 1000 to 215/126 while keeping a best model at 2.1/1.2 Å, respectively, and similar percentages of hits. The all-atom

Table 3. Results at the fragments level of bound and unbound docking

	Poses from unbound docking			Nb poses	Poses from bound docking		
	Min RMSD (Å)	Hits	Near-hits		Min RMSD (Å)	Hits	Near-hits
1B7F							
Frag 1 AMB	1.5	3	268	1×10^3	1.3	3	251
Frag 2 AMF	1.0	20	94	1×10^3	1.0	14	74
Frag 3 AF	1.3	22	105	1×10^4	1.0	15	74
Frag 4 SF	0.9	321	1155	2×10^6	1.0	204	770
Frag 5 SF	1.2	73	346	1×10^6	1.2	11	182
Frag 5 SB	1.3	29	118	1×10^6	2.2	0	39
Frag 6 SB	1.2	413	3313	$2Cx10^6$	1.5	33	1030
Frag 7 AB	0.8	73	235	1×10^4	2.1	0	90
2YH1							
Frag 1 AB	1.6	14	87	1×10^4	1.2	10	72
Frag 2 AMB	0.8	13	113	1×10^3	0.7	13	105
Frag 3 AMF	1.3	6	78	1×10^3	0.8	7	141
Frag 4 AFB	1.0	1	10	1×10^3	0.9	1	9
Frag 5 AMB	1.3	15	70	1×10^3	0.8	23	60
Frag 6 AMF	1.1	23	177	1×10^3	1.4	28	210

refinement of all the 1000 models barely changed the quality of these results.

Applicability to other RNA-binding domains

We further tested if our method could be directly applicable to other RNA-binding domains than the RRM. We chose as an example the Pumilio domain (PUF), artificially considering that the binding mode was unknown but for the position of two terminal nucleotides which we used as anchors. For this family, we selected the test structures of PDB code 3BXV (27) and 5BVZ (28) (Table 1). The RNA interface of the two proteins have 58% sequence identity and 2.4 Å backbone RMSD. The RNA sequences are similar but for the insertion of a C base in the middle in 5BVZ. This insertion is accommodated by torsions of the RNA backbone to bind globally the same binding site as in 3BXV, resulting in 2.9 Å all-atoms RMSD between the common nucleotides of the two complexes. The docking and chain-buiding on both pumilio domains proved highly successful, with the whole chain (6 and 7 nucleotides) being modelled with 1.9 Å RMSD precision, and with 8-5 hits among the only 20–15 final models (Table 2, Figure 10). Those very encouraging preliminary results suggest that our method, developed for RNA-RRM complex, would be generalisable to other RNA-binding domains, as long as the positioning of at least two anchors can be approximated at ∼2 Å RMSD.

DISCUSSION

Predicting the most probable conformation of a protein-bound ssRNA of a given sequence is made highly difficult by the flexibility of the ligand. Here, we show that fragment-based protein-RNA docking is an efficient strategy to address the huge conformational diversity of ssRNA. With our method, highly accurate ssRNA conformations can be generated from the RNA sequence. In the ensemble of models generated (a few dozen up to a few hundred), there is typically at least one chain with an RMSD better than 2 Å. This constitutes an excellent starting point for binding free energy calculations, and expert knowledge and limited experimental information may be used to identify the correct chain among the ensemble. The accuracies (up to 1.2 Å RMSD) of the predicted ssRNAs are comparable to the results classically obtained by protein – small ligand docking, even though the ssRNAs are very flexible and much larger than any small ligand. Also on the protein side, flexibility can be tolerated by our method. Good results were obtained, not only with bound receptor structures, but also with two unbound structures with substantial conformational change (1.7–1.5 Å RMSD towards the bound form). High-quality homology modelling can often generate structures of similar proximity to the bound form. Combined with the large number of known RRM structures, this opens up the possibility of obtaining receptor structures by homology-modelling of RRM domains. In other words, to successfully predict the protein-ssRNA structure, it might be sufficient to know the sequences of the RNA and of the RRM domains, if certain assumptions are fulfilled (see below). However, we like to emphasize that the scoring function has not been designed to distinguish between binding and non-binding sequences or nucleotides.

Our fragment-based ssRNA docking paradigm is fundamentally different from existing protein–RNA docking methods. For protein-bound ssRNA, prediction cannot be based on a structure of the unbound RNA (or the RNA bound to another protein) for two main reasons. First, RNA is disordered and/or unstable in a single-stranded form in solution, which prevents the resolution of any experimental structure. Second, even if experimental structures existed, the high flexibility of ssRNA would prevent an efficient coverage of its conformational landscape starting from one or a few particular conformations. This is a typical issue that can be addressed by a fragment-based approach. Fragment-based docking allows to avoid conformational sampling of the ligand as a whole, and facilitates the insertion of a large ligand in a tight binding pocket. More importantly, it permits the docking of a ligand for which no structure is known, but for which the structure of all fragments can be modeled and their conformational space extensively sampled.

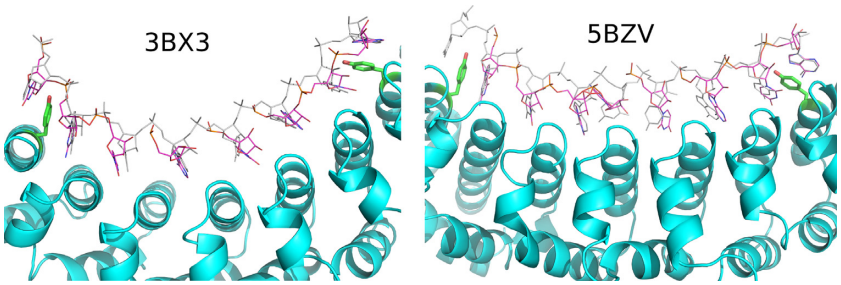

Figure 10. Best solution (min RMSD) for anchor-to-anchor chains obtained for each PUF–RNA complex after all-atom conversion and clustering. Same color-code as in Figure 8.

Still, our fragment-based approach faces certain limitations. To build the correct pose of the whole ligand, all the fragments have to be correctly sampled. This is made difficult by a non-homogeneous distribution of the binding energy: some fragments may contribute less than others to the binding, their weak binding being compensated by other 'hotspot-binding' fragments with high binding energy. In practice, one has to sample larger number of poses than in classical rigid-body docking to include such non-optimal poses. To obtain correct poses within a still reasonable amount of decoys, the scoring function has to be reasonably efficient for each fragment. This is only possible if three major assumptions are fulfilled.

The first assumption is that each fragment, even weakly bound, establishes sufficient favorable contacts with the receptor. In our benchmark this is fulfilled for all complexes. The presence of such weakly bound nucleotides in a complex could be assessed by NMR studies (e.g. by differences in bound-unbound chemical shifts, relaxation times, line broadening or an absence of NOEs), and this information could be added to predict the structure of a complex were such data are available but the NMR data are insufficient to resolve the complete structure of the RNA (31).

The second assumption is that the conformational library has to be sufficiently exhaustive to contain a structure close enough to the bound form for each fragment. This is true for most fragments in our benchmark. Yet most fragments in 4N0T have a closest conformer at 1.1–2.2 Å, which decreased the docking performance and made chain-building impossible for that complex. Such cases are not predictable, but their occurrence should diminish with regular updates of the library, thanks to a rapidly growing number of experimentally solved structures of RNA-protein complexes.

Finally, the surface of the receptor has to be known or modeled with sufficient accuracy. Our method proved highly efficient when bypassing this assumption, either entirely by docking on the bound protein, or partially by imposing the correct orientation of the two RRMs. The last procedure is relevant for real cases, as knowledge on the domain orientation could be provided by low-resolution experimental methods, such as SAXS or NMR data on RNA-protein complexes with unresolved RNA (e.g. using RDCs or paramagnetic restraints). Such low-resolution experimental data is considerably easier to obtain than a high-resolution structure. The inherent inaccuracy could be addressed by selecting several models to be used for the docking. However, given the very high precision reached by our 'biased' unbound docking (2.1 Å for eight nucleotides, 1.2 Å for five nucleotides), we hypothesize that such experimental inaccuracy could be tolerated by our protocol at the cost of some loss in precision. Inaccuracy in the prediction of the position of the anchor nucleotides would also negatively impact our results, and the procedure would probably fail for predicted positions beyond ∼3 Å RMSD. However, the conservation of rotamers in the conserved motifs between the bound and unbound protein for our 2 unbound cases permitted accurate enough predictions of the position of the anchors. Moreover, in the present study, the flexible hinge was removed, which might have had some negative impact on the results. In further studies, the hinge could be modeled, and an ensemble of hinge models could be used for docking. The increase in the number of decoys could be counter-balanced by improvements in the scoring function. The ATTRACT force-field used here for scoring was built in 2013 (22), based on only 109 non-redundant crystallographic structures of RNA-protein complexes. The number of such non-redundant structures has doubled since then (Januar 2011–October 2015). The continuing growth in the number of experimental protein-RNA structures allows future improvements in the ATTRACT force field, with a positive impact on the docking and scoring of individual fragments. Likewise, the scoring function at the whole-chain level will benefit from optimizations in our all-atom refinement process, including solvation effects. Finally, the current procedure completely neglects the RNA internal energy, both in the library building (no filtering of unrealistic conformations) and in the chain building. Taking this energy into account in future development could reduce the number of conformers to dock and thereby the number of decoys, as well as improving the evaluation of the chains. Together, such improvements would reduce even more the number of generated chains, and help identifying the correct chain by targeted experiments.

CONCLUSION

We provide a unique method to model up to 11 consecutive nucleotides of a ssRNA bound to one or two RRMs, from the structure of the protein and the sequence of the RNA. Tested on seven cases, the method modeled RNA chains with accuracy up to 1.2 Å (five nucleotides) to 2.5 Å (11 nucleotides). The method reached comparable success when tested on a case where only an unbound form of the protein and the relative orientation of the RRM domains were

known. Meanwhile, with integration of low-resolution experimental data on the orientation of the RRM domains in the bound protein, our method would permit to model new structures of RRM-RNA complexes of unknown structure with high accuracy. It is also directly applicable to other protein-ssRNA complexes. These models could serve as starting points for lead optimization for RNA-based drug design.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Deutsche Forschungsgemeinschaft (DFG) [Za153/19-2]. Computing time was by provided by the Leibniz Supercomputing Centre (LRZ) within grant pr48po. Funding for open access charge: DFG [Za153/19-2].
Conflict of interest statement. None declared.

REFERENCES

1. Geisler,S. and Coller,J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.*, **14**, 699–712.
2. Huang,Y., Zhang,J.L., Yu,X.L., Xu,T.S., Wang,Z.B. and Cheng,X.C. (2013) Molecular functions of small regulatory noncoding RNA. *Biochem. Mosc.*, **78**, 221–230.
3. Varani,G. and Nagai,K. (1998) Rna Recognition by Rnp Proteins During Rna Processing. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 407–445.
4. Hogg,J.R. and Collins,K. (2008) Structured non-coding RNAs and the RNP Renaissance. *Curr. Opin. Chem. Biol.*, **12**, 684–689.
5. Vanderweyde,T., Youmans,K., Liu-Yesucevitz,L. and Wolozin,B. (2013) Role of stress granules and RNA-binding proteins in neurodegeneration: a mini-review. *Gerontology*, **59**, 524–533.
6. Derrigo,M., Cestelli,A., Savettieri,G. and Di Liegro,I. (2000) RNA-protein interactions in the control of stability and localization of messenger RNA (review). *Int. J. Mol. Med.*, **5**, 111–134.
7. Kandil,S., Biondaro,S., Vlachakis,D., Cummins,A.-C., Coluccia,A., Berry,C., Leyssen,P., Neyts,J. and Brancale,A. (2009) Discovery of a novel HCV helicase inhibitor by a de novo drug design approach. *Bioorg. Med. Chem. Lett.*, **19**, 2935–2937.
8. Hugo,N. and Bouvet,P. (2003) Assemblage des complexes ribonucléoprotéiques. *MS Médecine Sci.*, **19**, 1271–1279.
9. Drolet,D.W., Nelson,J., Tucker,C.E., Zack,P.M., Nixon,K., Bolin,R., Judkins,M.B., Farmer,J.A., Wolf,J.L., Gill,S.C. *et al.* (2000) Pharmacokinetics and safety of an anti-vascular endothelial growth factor aptamer (NX1838) following injection into the vitreous humor of rhesus monkeys. *Pharm. Res.*, **17**, 1503–1510.
10. Peer,D. and Lieberman,J. (2011) Special delivery: targeted therapy with small RNAs. *Gene Ther.*, **18**, 1127–1133.
11. Bonvin,A.M.J.J. (2006) Flexible protein-protein docking. *Curr. Opin. Struct. Biol.*, **16**, 194–200.
12. Zacharias,M. (2010) Accounting for conformational changes during protein-protein docking. *Curr. Opin. Struct. Biol.*, **20**, 180–186.
13. Lensink,M.F. and Wodak,S.J. (2010) Docking and scoring protein interactions: CAPRI 2009. *Proteins Struct. Funct. Bioinforma.*, **78**, 3073–3084.
14. Leulliot,N. and Varani,G. (2001) Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry (Mosc.)*, **40**, 7947–7956.
15. Fulle,S. and Gohlke,H. (2010) Molecular recognition of RNA: challenges for modelling interactions and plasticity. *J. Mol. Recognit. JMR*, **23**, 220–231.
16. Morozova,N., Allers,J., Myers,J. and Shamoo,Y. (2006) Protein|NA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinforma. Oxf. Engl.*, **22**, 2746–2752.
17. Messias,A.C. and Sattler,M. (2004) Structural basis of single-stranded RNA recognition. *Acc. Chem. Res.*, **37**, 279–287.
18. 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.
19. 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.
20. Zacharias,M. (2003) Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci. Publ. Protein Soc.*, **12**, 1271–1282.
21. de Vries,S.J., Schindler,C.E.M., Chauvot de Beauchêne,I. and Zacharias,M. (2015) A Web Interface for Easy Flexible Protein-Protein Docking with ATTRACT. *Biophys. J.*, **108**, 462–465.
22. Setny,P. and Zacharias,M. (2011) A coarse-grained force field for Protein–RNA docking. *Nucleic Acids Res.*, **39**, 9118–9129.
23. Mackereth,C.D., Madl,T., Bonnal,S., Simon,B., Zanier,K., Gasch,A., Rybin,V., Valcárcel,J. and Sattler,M. (2011) Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, **475**, 408–411.
24. Lee,A.L., Kanaar,R., Rio,D.C. and Wemmer,D.E. (1994) Resonance assignments and solution structure of the second RNA-binding domain of sex-lethal determined by multidimensional heteronuclear magnetic resonance. *Biochemistry (Mosc.)*, **33**, 13775–13786.
25. 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.
26. Delano,W. (2002) The PyMOL Molecular Graphics System.
27. Méndez,R., Leplae,R., Lensink,M.F. and Wodak,S.J. (2005) Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins*, **60**, 150–169.
28. 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.
29. Wilinski,D., Qiu,C., Lapointe,C.P., Nevil,M., Campbell,Z.T., Hall,T.M.T. and Wickens,M. (2015) RNA regulatory networks diversified through curvature of the PUF protein scaffold. *Nat. Commun.*, **6**, 8213,
30. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–222.
31. Phelan,M.M., Goult,B.T., Clayton,J.C., Hautbergue,G.M., Wilson,S.A. and Lian,L.-Y. (2012) The structure and selectivity of the SR protein SRSF2 RRM domain with RNA. *Nucleic Acids Res.*, **40**, 3232–3244.