Fragment-based modeling of protein-bound ssRNA

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

Fragment-based modeling of protein-bound ssRNA

Isaure Chauvot de Beauchêne, Sjoerd J. De Vries, Martin Zacharias. Technische Universität München (Germany)

BIOLOGICAL CONTEXT

Single-stranded (ss)RNA — *Disordered in solution*

ssRNA-protein complex — *Key contacts for specificity*

ssRNA-binding protein

incorrect → **Disease**
- Viral reproduction
- Myogenic dystrophy
- Degenerative diseases

Correct → **Biological function**
- Transport of mRNA
- Maturation of mRNA
- Regulation of translation

Artificial → **Drug**
- Growth factor inhibitor
- Anti-virus (C-hepatitis)
- *In vitro* diagnosis

The structure of an RNA-protein complex is a key to:

(i) understand its function or malfunction

(ii) modulate or create it, for medicine or biotechnology

Experimental methods to obtain such structure (X-ray, NMR) are costly, time-consuming or limited to some complexes.

Therefore, it often requires computational modeling methods. Such methods exist for RNA-protein complexes, but fail at modeling **single-stranded RNA** because of their flexibility.

RATIONALE

Experimentaly known structures

Unbound protein

Unbound RNA

Unbound **ssRNA** Undefined

Classical Docking

Sampling of relative positions
(± ligand deformations)
Scoring of the final poses

Docking MODEL *versus* RNA-protein X-ray structure

***Ab initio* Modeling**

Modeling the RNA *in situ*, from its sequence

Solving experimentally the structure of an isolate protein or RNA is easier than of a complex.

On can use those isolate structures to model the complex via docking methods.

But if the structure of the free ligand is unknown, it has to be modelled from sequence.

For highly flexible objects like long ssRNA, possible conformations are too numerous to be all modelled

Our alternative approach consists in modeling RNA local conformations and assemble them on the protein surface.

12 DOF to explore per nucleotide
Fewer correlations (constraints) than in double helix
=> combinatorial explosion for > 5-6 nucl.

RNA adopt discrete local conformations (= *rotamers*)
=> they can be represented by a finit number of structural fragments

STRATEGY

RNA sequence

A U G G U C

Divided into overlaping trimers

Fragments library

AUG UGG GGU GUC

All possible local conformations.
Extracted from structure databases

Protein structure

X-ray, NMR ...

Predictable contacts
Conserved in protein family (optional)

Fragments docking

Parallel (if no known contacts)

OR

Sequential (if known contacts)

Positional restraints

Assembling

Identification of **chains**
= spatially overlapping poses

CONCLUSION

Achievements

Our fragment-based approach to model protein-bound ssRNA proved effective to sample fragment poses at the surface of the protein. This permits to **predict the RNA binding site** with same sensitivity and higher specificity than all other binding-site prediction methods based on protein structure [3].

We can predict the orientation of nucleotides binding to **RNA-protein conserved contacts,** in the most abundant RNA-binding domains of proteins (RRM and PUF) [4].

With those anchoring nucleotides, we could model bound **ssRNA up to 12-nucleotides** long, with a **resolution comparable to X-ray structures** [4].

Perspectives

We considered so far that we know which part of our RNA is single-stranded (*ss*) and binds the protein. In many real-cases, the RNA is partially structured (e.g. double-stranded, *ds*) or parts of it oscillate between *ss* and *ds* state. Moreover, the *ss* part binds the protein by only some of the nucleotides.

Therefore, we will include **RNA secondary structure prediction** methods together with ***in vitro* data** (e.g. SHAPE) to evaluate the likelihood of protein-binding for each nucleotide in the RNA of interest, before or within the docking process.

METHODS

Docking

ATTRACT docking engin [1,2]

1/ ~ 10^7 random starting states (position * orientation * conformation)
2/ Energy minimisation of bead-bead interactions in an empirical force field
3/ Elimination of redundant poses (converged on same local minima)
4/ Ranking of poses by score (= pseudo-energy)

=> For each fragment : best pose at 1 – 3 Å from X-ray structure among ~10^5 – 10^6 poses

protein

RNA

Assembling

Probabiliste

Up to 10^6 poses per fragment

Systematic

Forward – backward paths count
=> Selection of the most connected poses
(i.e. most probably part of the correct chain)

Pruning from each anchoring contact
=> Enumeration of all possible chains

Scoring

Weights in final scoring

Rank 6 Rank 100 Rank 56,000 Rank 1 Rank 1,200

Chains are scored by the geometric mean of the ranking* of the poses.

This enhances the weight of very well-ranked poses, to account for hot-spots** in the RNA.

$$\text{Score} = \left(\prod_{i \text{ in } [1,N]} \text{rank}(i) \right)^{1/N}$$

low score = good chain (hopefully)

* by pseudo-energy of protein-RNA interaction

** Fragments that bind with high energy (Often key parts for specific recognition)

Hierarchical clustering for efficient pruning

By distance

8 Å
5 Å
3 Å
...

By computing the distance between the centers of two clusters, we assess if they could contain overlapping poses.

If yes, their subclusters are evaluated pairwise, and so on.

This spares pose-pose comparisons for pairs of poses belonging to distant clusters.

By best ranks

300 rejected 120
2 accepted 10
160 100 10
2
830 100
60 20 10
2

By assembling a small subset of poses, we estimate an upper limit of the best possible score.

We consider the best rank contained in each cluster.

For each pair of distance-compatible clusters, we check if the best possible score is acceptable. If yes, we go down to their sub-clusters.

Score < 25
(300 x 120)$^{1/2}$ ~ 33
(2 x 120)$^{1/2}$ ~ 15

RESULTS

Without predicted contacts

We blind-tested this approach on 2 complexes of known structure [3].

Filtering the poses by their chain-forming propensity (connectivity) enriches the pool in correct poses (RMSD <5Å) more effectively than the docking score alone :

	Correct poses	Total
Docking score	1% - 4 %	6 . 10^5 – 1 . 10^6
Connectivity	10% - 13 %	3 . 10^3 – 8 . 10^3

X-ray
Docking

More than 10% of poses are correct

Binding site prediction

In white: RNA position in X-ray structure

Nb of poses contacting each amino-acid

With predicted contacts

We blind-tested this approach on 8 complexes of known structure [4].

We predicted the position and orientation of nucleotides establishing conserved contact, within 0.8 - 3.2 Å RMSD from the real structure (average 1.4 Å).

From those contacts, we modelled **5 to 7-nucl RNA within 2 Å RMSD** from the real structure, for 7 out of 8 complexes, among 130 to 270 proposed models.

By prolonging the RNA chain beyond the predicted contacts, we model **7 to 12-nucl RNA within 4 Å RMSD**.

MODEL *versus* X-RAY

Predictions (pink) versus experimental structure (X-ray, white) of RNA bound to either a RRM domain (green, pdb-code 2CVJ) or a PUF domain (cyan, pdb-code 3BX3)

[1] de Vries SJ, Schindler C, Chauvot de Beauchêne I, Zacharias M. (2015) Biophys J.
[2] Setny P, Zacharias M. (2013) *Nucleic Acids Reserch*
[3] Chauvot de Beauchêne I, de Vries SJ, Zacharias M. (2016) *PloS.Comput.Biol.*
[4] Chauvot de Beauchêne I, de Vries SJ, Zacharias M. (2016) *Nucleic Acids Research*

The authors thanks the **Deutsche Forschungsgemeinschaft (DFG)** for funding this work, and the **Leibniz Super-computing Centre (LRZ)** for providing computational time.