A hybrid combinatorial method for docking single stranded RNA on proteins at the thermodynamic equilibrium

Chinmay Singhal, Yann Ponty, Isaure Chauvot de Beauchêne

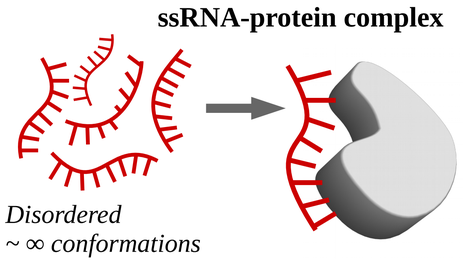
A hybrid combinatorial method for docking ssRNA on proteins at the thermodynamic equilibrium

Chinmay Singhal [1], Yann Ponty [2], Isaure Chauvot de Beauchene [3]

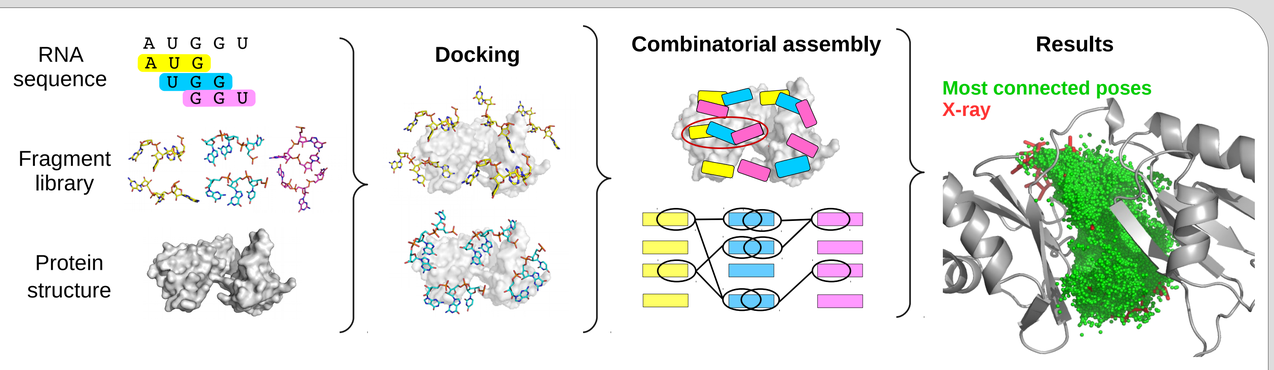
1. AMIBio team, Inria Saclay 2. CNRS - LIX 3. CNRS - LORIA

Introduction

ssRNA-protein complex



Protein-RNA complexes participate in many aspects of cell regulation. Their atomistic structural description is crucial to understand the recognition mechanism, but the experimental resolution of their structure is arduous. **Computational docking** methods aim at modeling a 3D assembly, by assembling structures of each isolated constituent. Yet for highly flexible objects like **single-stranded RNA (ssRNA)**, the isolated structure of the whole molecule can adopt an ensemble of conformations too large to be experimentaly solved or computationaly modeled.



We recently proposed a fragment-based approach to **model ssRNA local conformations and assemble them on the protein surface**. By docking then re-assembling overlapping fragments, we could select a pool of most-connected poses that delineates the RNA-binding site (Fig. right) [1]. But **the number of possible assemblies is beyond the reach of brute force approaches.**

Here, we present an improved method capable of **modeling the ssRNA solely based on the protein structure and ssRNA sequence**. Improvements include **(A)** a **new docking protocol** for systematic sampling in deep pockets, **(B)** a **stochastic backtracking algorithm** for unbiased sampling of assemblies, and **(C)** a **combination of filters** of the ensemble of sampled models based on biophysical caracteristics of the binding site.

As a proof-of-principle, we applied this method on a **11-mer poly-U ssRNA inserted in the deep cavity of an exonuclease**

A. RNA fragment docking w. deepATTRACT

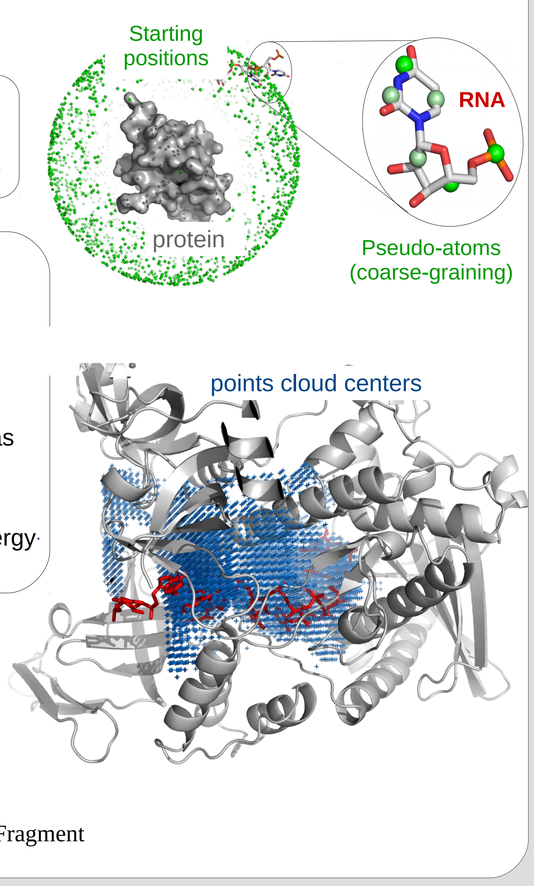
For the fragment assembly to succeed, all positions in the sequence must have been sampled correctly. We previously used the ATTRACT software [3] to dock RNA on the protein surface. Here **we adapted ATTRACT for docking inside deep cavities**. As we deal with a poly-U, one single docking of a UUU fragment was performed.

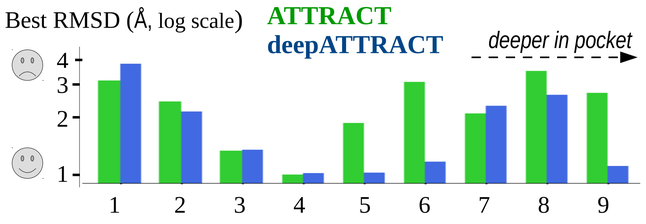
ATTRACT

search space : 10^7 states (position, orientation, conformer)
Energy minimisation in empirical force field
Select 10^6 poses of lowest energy
fragment library

deepATTRACT

Detection of pocket points (POCASA [4], 2 Å probe, 1 Å grid)
Selection of points cloud centers (> 500 points within 7Å)
Cluster fragment library to keep few representative
Coarse filter: Sample orientations * points * few conformers
Retain low-energy poses
Fine filter: For each pose, test all conformers in same cluster as the representative
=> search space **10^9**
Keep low-energy poses, minimize further, keep **10^6** lowest energy
Cluster w. 2 Å cutoff





B. Fragments assembly by stochastic backtracking

We then search for **chains of compatible docking poses with low total energy** (approximating poses energy as additive). 10^6 docking poses have to be kept to ensure retaining enough correct solutions at each sequence position.

=> **10^{72} possible assemblies !**

=> **How can we retrieve the most probable assemblies while avoiding a brute-force enumeration?**

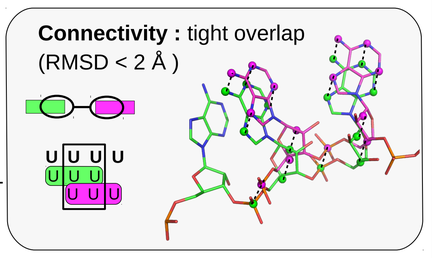
By an unbiased sampling of paths: For each sample, at each position in sequence, the next pose is chosen among all poses with a probability given by the **partition function Z**, which defines the probability of each pose to participate in a downstream path. Z is pre-calculated by **dynamic programming**.

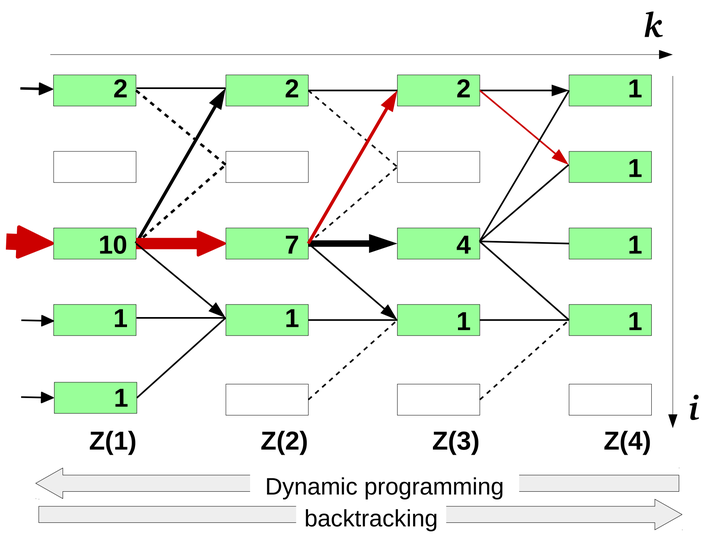
Partition function Z — Docking scores (~ binding energy)

$$Z_{fwd}(k,i) = \sum_{j/connect(j,i)} \exp\left(\frac{E(i,j)}{RT}\right) \times Z_{fwd}(k-1,j)$$

Position in sequence — Index of docking pose

Connectivity : tight overlap (RMSD < 2 Å)



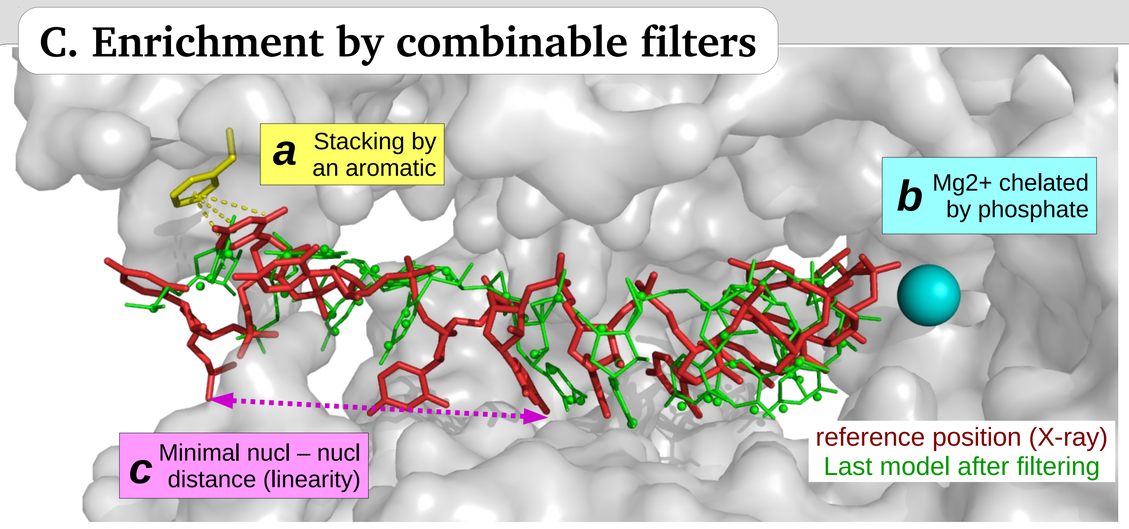


Example with T = ∞ (simply sum over paths)

- Compute all pairwise connectivities
- Eliminate non-connected poses (propagate)
- Compute partition function Z backward, by dynamic programming
- Sample **paths** forward, w. proba = Z(k,i) / ∑Z(k)
- Iterate 100,000 times

This stochastic backtracking results in a **canonical sub-ensemble of assemblies**

C. Enrichment by combinable filters



The stochastic backtracking provied 100,000 models containing **0.3% correct solutions**, and a **best solution at 2.2 Å** from the reference position (Xray).

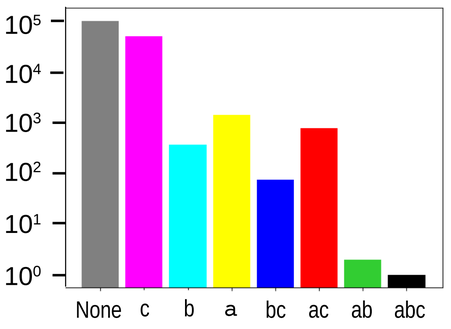
The scoring function of ATTRACT is not precise enough to select the best models. Instead, we used some knowledge on the system to define **geometric constraints used as filters**:

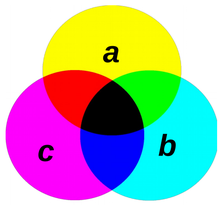
a. Aromatic rings are known for establishing stacking interactions with RNA bases. One is present at the entrance of the protein pocket → we impose the 4th base to be at < 5 Å from the aromatic ring.

b. Mg^{2+} ions are known for establishing ionic interactions with RNA phosphate groups. One such ion is present in the bottom of the protein pocket → We impose the last phosphate to ba at < 7 Å from the Mg^{2+}.

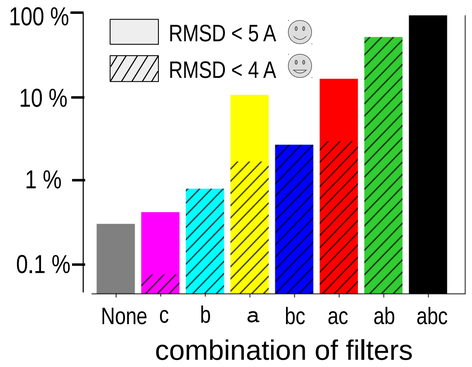
c. We consider the single-stranded RNA as linear → nucleotides not in same fragment are at > 6 Å from each other

Size of the ensemble of models





Fraction of correct models



Applying each filter separately leads to an enrichment of x **1.6** to x**34** in correct solutions. Combining two filters leads to an enrichment of x **2.7** to x **163.** Combining the three filters retains only one model, which is at **4.0 Å** from the reference position (cf Figure above).

Conclusion

We present an improved method capable of modeling protein-bound ssRNAs solely based on the protein structure and ssRNA sequence. The main improvements include:

- A **new docking protocol** for docking RNA fragments inside deep pockets with systematic sampling. This shows improved sampling quality (lower RMSDs) for the nucleotide fragments buried inside the binding pocket.
- A **stochastic backtracking algorithm** that can perform unbiased sampling from the connectivity graph of the docked fragments. This algorithm generates near-native RNA chains among 100,000 samples.
- An efficient and effective **filtering procedure** to incorporate experimental knowledge on the protein-ssRNA system. The filters can be combined, leading to an enrichment of up to x163 after filtering

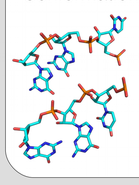
As a first proof of principle, the method could model *ab initio* a protein-bound ssRNA with a length of 10 nucleotides, an unprecedented length far beyond the reach of standard small-molecule docking programs.

Perspectives

Combine probabilities

Weight the edges and nodes of the graph by :

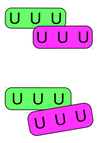
Conformational energy



E(i) = -1 Kcal/mol
P(i) = 0.2

E(i) = - 2 Kcal/mol
P(j) = 0.54

Connectivity

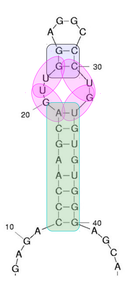


RMSD 1.0 Å
P(i, j) = 0.9

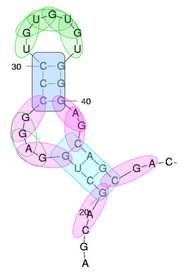
RMSD 2.5 Å
P(i, j) = 0.5

Fold the RNA on the protein

From RNA sequence, predict a Boltzmann ensemble of secondary structures (2D), dock all fragment of possible structure at each position in sequence, then incorporate the 2D structure probability in the probabilistic assembly.

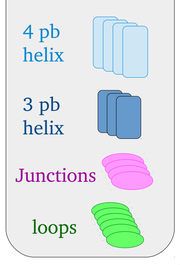


P (2D_a) = 5



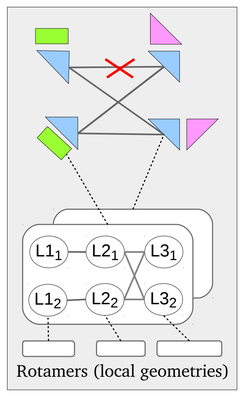
P (2D_b) = 1

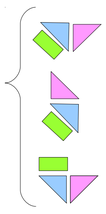
2D-specific fragment libraries



Co-assemble w. protein conformations

Account for protein flexibility : Proteins also change conformation upon binding, albeit less drastically than RNAs. To avoid sampling many of their possible global conformations and use all of them for docking, one can apply to the protein the same principle of decoupled local sampling as for RNA.

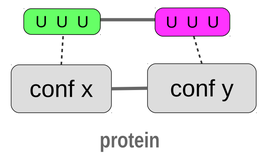




Domains orientation : Discretize the ensemble of possible pairwise orientations, then assess which are geometrically compatible

Loop conformations : Enumerate local conformations inside domain, check compatibilities pairwise and with domains orientation.

Each “local” set of conformations can be used for docking, then the compatibility of adjacent conformations can be assessed together with the connectivity of RNA poses.



[1] Isaure Chauvot de Beauchêne, Sjoerd J. de Vries, Martin Zacharias (2016) PloS.Comput.Biol.
[2] Martin Zacharias (2003) Proteins [3] Sjoerd J. de Vries, Christina E.M Schindler, Isaure Chauvot de Beauchêne, Martin Zacharias (2015) Biophys. J.