Protein-RNA docking with ATTRACT

Download all input files and scripts on the website of the MBI-DS4H platform

Type in terminal: export MAIN=/home/user/.../practical_course

---

exo 1: Rigid docking

In this exercise, you will assess the impact of RNA and protein flexibility on results obtained by rigid docking, using as test-case a complex between the ribosomal protein TL5 and a double-strand RNA. The structures of the protein unbound, the RNA unbound and the protein-RNA complex are all experimentally known (PDB IDs 2J01, 364D, 1FEU).

You will first run an artificial docking test, using the **bound** structures of the RNA and protein (i.e the same structure as in the experimental complex). You will then run a real-case docking, using the **unbound** structures of RNA and protein. The difference between the bound and unbound structures correspond to conformational changes that the RNA and protein undergo prior to or during the binding process.

1. Bound docking

go on: http://www.attract.ph.tum.de/services/ATTRACT-devel/standard.html and fill up the parameters below. The pdb files to be uploaded are in inputs/exo1/.

Partner (left menu)

- **Receptor:** structure file: upload protein_b.pdb
  RMSD calculation : on
  Reference RMSD PDB file: upload protein_b.pdb
- **Ligand:** structure file: upload rna_b.pdb
  RMSD calculation : on
  Reference RMSD PDB file: upload rna_b.pdb
  What kind of molecule are you docking? : choose

Analysis: Calculate interface RMSD after docking: on
Number of structures to collect as PDB file : 20
Maximum number of structures to analyze : 1000

Computation: Name of the docking run : *bound*
Nb of CPU : 4 *if* you run it via jupyter, your number of CPU if you run it on your machine

Click on **Get configuration** (bottom left). Download bound.tgz, unzip, go in the main directory and run the docking, by typing:
tar xzf bound.tgz ; cd bound; ./bound.sh). [3']

When the computation is done, you should get a file *results.irmsd* that contains a list of solutions with their rank (1st column) and iRMSD (2nd column). To print the solutions with

iRMSD < 3 A in the top-100 ranked solutions (if any), type:
awk '$1<100 && $2<3' results.irmsd

Plot the results: $MAIN/scripts/plot-rmsd-rank.sh result.irmsd
How is the sampling[1]? The scoring[2]?
Visualize the top-20 ranked solutions: pymol results.pdb receptor.pdb ligand.pdb
Check the consistency of the iRMSD values by comparing the models obtained to the bound molecules.

2. Unbound docking

Have a look in pymol at the unbound structures of the RNA and protein ([protein/rna]_**ub**.pdb), that have been fitted on the bound structures ([protein/rna]_**b**.pdb). Look at an all-atom representation (e.g. "lines").
What difficulty will the rigid unbound docking face?

Close and re-open the web-interface (not just refresh, else some parameters are not re-initialized). Perform the docking with the unbound forms. As you still want to compare your results with the experimental structure of the complex, keep the bound form of each molecule as the Reference RMSD PDB file to compute the iRMSD. Download, run the docking and analyze the results like previously.
How are the sampling and scoring affected by the molecule flexibility ?

exo 2: Flexible docking

To account for molecules flexibility during docking, we will test and compare two forms of flexible docking and their combinations:

- Pre-compute **harmonic modes** (i.e. energetically favorable directions of deformation) based on the structure of each molecule and its internal atomic forces. The modes are then used as additional degrees of freedom along the minimization. This is very demanding in term of computational time, so here we will limit ourselves to up to 4 modes, while in principle a dozen is recommended.
- Use a **conformational ensemble**, i.e. a set of different unbound structures. Those can come from different X-ray experiment, or NMR experiment, or can be obtained by MD simulations (cf lecture by S. Pasquali). Here we provide a set of 10 RNA structures coming from NMR, and 3 protein structures coming from X-ray.

We will use as a test-case a complex between the transcription factor NF-kappaB(p50) and a kinked stem-loop RNA. The structures of the unbound protein, unbound RNA and protein-RNA complex are all experimentally known (PDB IDs 1LES, 2JWV, 1OOA).
On the ATTRACT web-interface, use protein_ub.pdb and rna_ub.pdb in inputs/exo2/ as receptor and ligand. Perform two docking runs:

---

1 Capacity of the docking to find at least one near-native structure, e.g. with low iRMSD.
≤ 1 Å is excellent, ≤ 2 Å is good, ≤ 4 Å is acceptable.

2 Capacity of the scoring function to discriminate (give low rank to) the near-native structures

1. Rigid docking

No flexibility, same as previously. This will be our reference test to assess the improvement obtained by using flexible docking.

2. Flexible docking

Each of you will choose one combination of flexible options, then put a cross in the corresponding box in the shared table *here* (to avoid that all groups use the same options). Use only one type of flexibility per molecule. Same settings as exo1 in **Analyses**.

- **"hm" : harmonic modes** (for receptor and/or ligand)
  Generate harmonic modes : on
  number of modes: 1-2 if hm on both molecules, ≤ 4 else-wise (to save time)
- **"ens" : conformational ensemble** (for receptor and/or ligand)
  Receptor/Ligand: Use [protein/rna]_ub_ens[3/10].pdb

!!! Using hm on the receptor is very time consuming. If you want to use it on the protein and not the RNA, better use the RNA as receptor and the protein as ligand. Use it on both molecules only if you run the docking on your laptop on 8 or more CPUs.

While the docking runs, you can start preparing (but not run yet!) the docking in exo3.

Download and analyze the results as previously. Fill up the shared table. Remove the cross and add the data on your best-RMSD solution: iRMSD (rank). It should look like this:

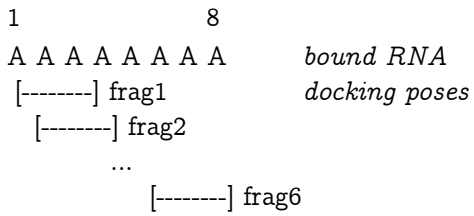
<table>
<tr><th></th><th>protein
rigid</th><th>protein
ens</th><th>protein
1 hm</th><th>protein
2 hm</th><th>protein
3 hm</th><th>protein
4 hm</th></tr>
<tr><td>RNA rigid</td><td></td><td>X</td><td></td><td></td><td>0.5 (2)</td><td></td></tr>
<tr><td>RNA ens</td><td></td><td>1.2 (20)</td><td>X</td><td></td><td></td><td></td></tr>
<tr><td>RNA 1 hm</td><td>X</td><td>5.2 (20)</td><td colspan="2" rowspan="2">> 10' on 4 CPU</td><td colspan="2" rowspan="4">Run on your laptop
if > 4 CPU</td></tr>
<tr><td>RNA 2 hm</td><td></td><td></td></tr>
<tr><td>RNA 3 hm</td><td></td><td></td><td colspan="2" rowspan="2">Run on your laptop
if > 4 CPU</td></tr>
<tr><td>RNA 4 hm</td><td>X</td><td></td></tr>
</table>

Does the flexible docking improve the sampling ?

exo 3: Fragments docking

In the case of single-stranded RNA, the unbound form is too flexible to be observable experimentally. We cannot use classical docking as previously. We will use a fragment-based approach to dock the RNA from its sequence. Here, we assume that we know that the RNA binds in a single-stranded state, as most protein families that bind ssRNA are well identified.

We will use as an example a poly-A RNA of 8 nucleotides binding to a 2-domains poly-A-binding protein. The structure of the bound complex has been solved experimentally (PDB ID 1CVJ). We will dock AAA fragments, compare the poses to each 3-nucleotide part of the bound RNA, then assemble chains of poses overlapping by 2 nucleotides (so 6 poses in total, to get an 8-nucl chain).



One should in principle use a full-size fragment library (conformational ensemble) of thousands of conformers, to cover the structural diversity of a trinucleotide. To make computation times compatible with the time constraints of this workshop, we will use a reduced ensemble of 12 conformers.

1. Fragments docking

On the web-interface, use one of the protein_ub-1.pdb as receptor and rna_ens12.pdb as ligand, in inputs/exo3/. Do **not** use any harmonic modes. Deactivate RMSD calculation. Run the docking.

After docking is done, first check if each fragment has been correctly sampled. From the docking directory, run: $MAIN/scripts/rmsd.sh. In the terminal, you will see the 5 best solutions obtained for each fragment, with its rank by energy.

2. Fragments assembly

Assembly the docking poses into chains of up to 6 fragments (up to 8 nucl). Adjacent poses in a chain must have their common nucleotides at the same position on the protein (concretely, have an overlap RMDS below a cutoff). We will use the branch-and-bound method that allows you to sample only chains with a geometric mean of the ranks below a given cutoff.

We will adjust the overlap and mean-rank cutoffs in order to obtain correct solutions. This is a calibration exercise. Use:

$MAIN/scripts/assemble.sh

nfrag	number of tri-nucleotides to assemble
cutoff	overlap cutoff (in Å). Start with a small one (e.g. 1.0). **LIMIT: 5.0**
nposes	number of top-ranked poses to assemble. Recommended [100 - 10000]
meanrank	max geometric mean of pose ranks in each chain (≤ npose)
maxchains	max number of chains in output. Recommended ≤ 100000

ex: $MAIN/scripts/assemble.sh 6 1 100 100 1000
nfrag cutoff npose meanrank maxchains

When you have found a suitable set of parameters that give you at least an acceptable solution (iRMSD < 5A), visualize it with pymol result.pdb chains-*.pdb rna_b.pdb protein_b.pdb

Fill up the shared $MAIN/exo3/table *:*

nfrag	**cutoff**	**nposes**	**meanrank**	**maxchains**	**best RMSD**	**rank**
6	1	100	100	1000	4.2	1024

In **real life docking**, you would create a benchmark of 3D structures of RNA-protein complexes that have properties similar to the real case you want to dock (such as size of the protein, size of the ligand, approx binding affinity, existing data for data-driven docking...). Then you would train your docking method on that benchmark in order to determine the best parameters, i.e. those that give you correct solutions among as few sampled chains as possible, for as many cases in your benchmark as possible. Then, you would apply those parameters on your real-case system.

To determine which of your sampled chains could be a correct model, you will test what is the smallest set of **low-resolution experiments** that could distinguish between them. For instance, mutagenesis at 10 different residues that are at the interface or not in the different chains can in theory distinguish among up to 2**10 solutions.

Additionally, you can test the validity of models by running **molecular dynamics** (at least 3 simulations for reproducibility). Check the stability of the model to discard unstable ones, and/or compute the binding energy, to compare with experimental affinity (more time-consuming).