New strategy for optimizing knowledge-based docking parameters: application to ssRNA-RRM docking

Anna Kravchenko, Malika Smaïl-Tabbone, Sjoerd Jacob de Vries, Isaure Chauvot-De-Beauchêne

Histogram-based approach for docking parameters optimization

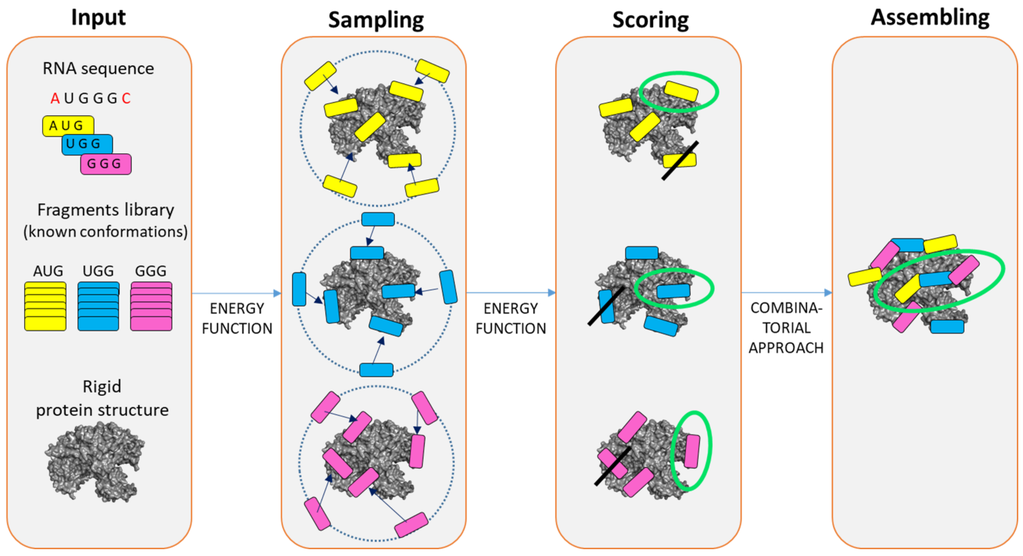
Anna Kravchenko[1], Malika Smaïl-Tabbone[1], Sjoerd Jacob de Vries[1,2] and Isaure Chauvot-de-Beauchêne[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France [2]Ressource Parisienne en Bioinformatique Structurale (RPBS), Paris, France

1. Fragment-based docking

Docking is a computational prediction of a 3D structure of a molecular complex. A fragment-based docking approach was developed to tackle highly flexible ligands [1], e.g. ssRNA bounded to protein. It works by:
1/ splitting the ligand into overlapping fragments;
2/ *docking* them onto the receptor separately;
3/ assembling compatible poses back into the whole ligand.

Docking consists of a) sampling - generation of the poses by energy minimization from random starting points around the protein with the help of the **differentiable energy function**; and b) scoring - filtration of the poses by their RANK (keep poses with low energy, remove poses with high energy).



Current challenges of such a docking approach for ssRNA:
1/ **Sampling**: at least one near-native pose (correct pose, LRMSD under 3Å) has to be generated per fragment;
2/ **Scoring**: at least one near-native pose per fragment has to remain after the scoring.

These challenges can be addressed by the optimization of the docking parameters.

In the foundation of this work lies the assumption that information about common contact distances in the binding interfaces of protein-ssRNA complexes can be used to: a) identify near-native poses; b) sample more near-native poses.

2. Histogram-based approach

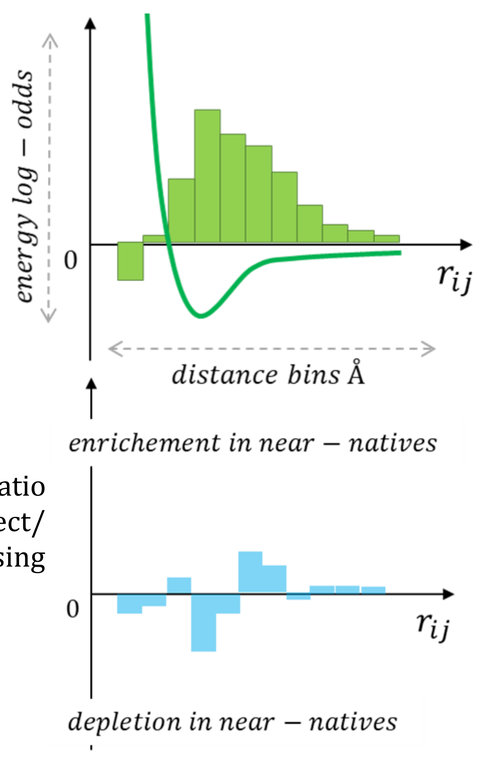
The histogram-based approach requires the creation of log-odds and residual histograms for each unique pair of beads (atoms in coarse-grained representation) type *(i,j)*.

Log-odds histogram is built by converting the energy function (Lennard-Jones curve with the soft potential) into relative probabilities of each bead-bead distance (discretized into bins) in

$$log\big(P(r_{ij})\big) = -kTE_{ij}$$

protein-RNA interfaces using the next formula:

where k – the Boltzmann constant, T – absolute temperature, r_{ij} – the distance between a pair of beads type *(i,j)*, E_{ij} – energy approximation between a pair of beads type *(i,j)*.

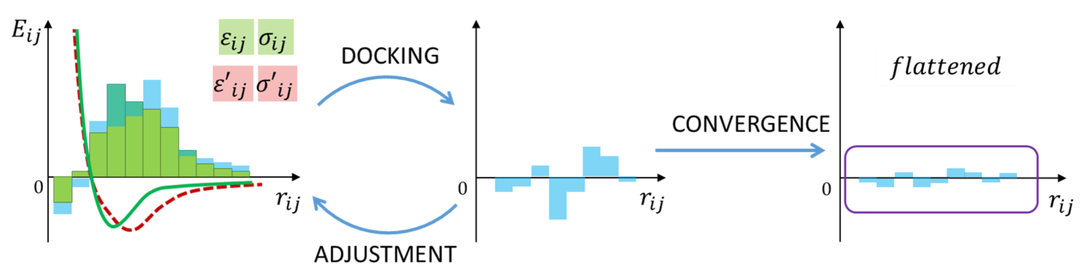


Residual histogram is built by measuring the ratio of occurrences of bead-bead distances in correct/ incorrect docking poses (obtained using ATTRACT docking engine [2]):

$$log\left(\frac{\%\, N_{near-native}(r_{ij})}{\%\, N_{non-near-native}(r_{ij})}\right)$$

which corresponds to the residual error of the energy function.

Next, we sum log-odds and residual histograms and fit the docking parameters $[\sigma_{ij}, \varepsilon_{ij}]$ to the resulting histogram. With the new parameters $[\sigma'_{ij}, \varepsilon'_{ij}]$, we re-dock the benchmark and obtain a new residual histogram. Repeat until convergence - until the residual



histogram is **flat**:

After convergence, this procedure should generate **equal distributions** of bead-bead distances in correct and incorrect poses, which are thus indistinguishable by bead-bead distances criteria. The number of correct poses will then have been completely optimized by bead-bead distances.

3. Preliminary results

Current dataset consists of 131 data cases (1 protein + 1 RNA fragment).

We ran preliminary tests for this approach by making a set of histograms for all beads of the randomly selected data case 1M5K-UGC, 337 histograms in total. Then we re-ranked docking poses for the remaining data cases with this set of histograms. We consider such a re-ranking to be successful if at least 75% of the near-native poses (LRMSD < 5Å) ended up in top 20% of all ranked poses; to fail if less than 20% of near-native poses were in top 20% of all ranked poses.

Then we repeat the same experiment for data case 3MOJ-UUC (randomly selected from the set of failed cases for 1M5K-UGC histograms) obtaining a set of 196 histograms. The results are shown in Table 1:

Table 1: Results of test-cases re-ranking

overlap of cases

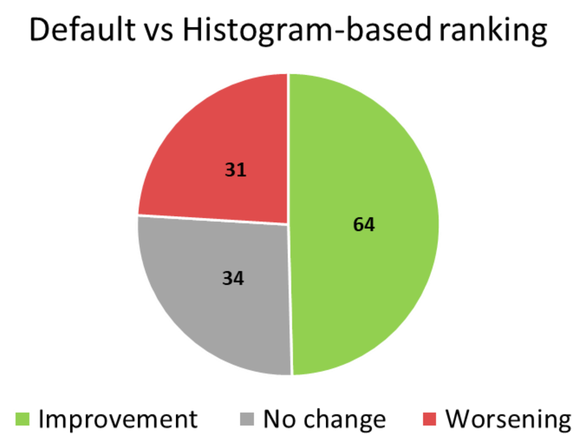
	Number of cases, ranked by 1M5K-UGC histo		Number of cases, ranked by 3MOJ-UUC histo	Number of cases, ranked by default
Success	54	*10*	35	6
Failure	48	*27*	63	39
Other	28		32	85

To compare performance of both sets of histograms with performance of the current docking parameters for each data case we:

1. Pooled 10% of top-ranked by 1M5K-UGC histograms poses with 10% of top-ranked by 3MOJ-UUC histograms poses; removed redundant poses; calculated percentage P_{histo} of remaining in this pool near-native poses;
2. Calculated percentage P_{def} of near-native poses in top 20% of ranked by default parameters docking poses;
3. Calculated $\Delta = P_{histo}-P_{def}$. Here the evaluation criteria are the next:

- If $\Delta \geq 15\%$, histogram sets outperform default parameters (improvement);
- If $\Delta \leq -15\%$, default parameters outperform histogram sets (worsening);
- Otherwise, $-15\% < \Delta < 15\%$, there is no significant changes (no change);

Two histogram sets, each based on a single data case, outperform the default ranking by docking parameters on 49% of the dataset by bringing at least 15% more near-native poses into the top 20% of all docking poses.



4. Conclusions

We are developing a new method to optimize the parameters of a knowledge-based energy function for the coarse-grained fragment-based ssRNA-protein docking. Our first results of application to a few protein-fragment cases of known structures are very promising.

Perspectives

- With the current procedure, we aim to obtain a small number of sub-optimal histogram sets, which cover the whole protein-fragment dataset. Alternatively, it can be a larger number of sub-optimal histograms, in this case we need to have a way to determine which histogram set is suitable for a given protein-fragment case (e.g. by the protein family, ligand sequence, experimental information). This will address the **scoring** problem;
- We plan to develop an algorithm to convert a set of histograms into a set of the docking parameters, which will allow to address the **sampling** problem;
- Upon optimization, the correct poses could be further identified on other criteria than bead-bead distance (e.g. angles between atoms) with the help of machine learning;
- With updated parameters, ssRNA-protein docking can become a powerful tool for solving ssRNA-protein complexes. Combined with AlphaFold for the determination of protein structure, ssRNA-protein docking can be performed by a researcher with little to no knowledge of structural biology, making it accessible to the wide scientific community.

[1] Chauvot de Beauchene I, de Vries SJ, Zacharias M. (2016). Nucleic Acids Research
[2] Setny P, Zacharias M. (2011). Nucleic Acids Research
[3] Chauvot de Beauchene I, de Vries SJ, Zacharias M. (2016). Nucleic Acids Research

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 813239