HIPPO: HIstogram-based Pseudo-POtential for scoring ssRNA-protein fragment-based docking poses

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

HIPPO: HIstogram-based Pseudo-POtential for scoring ssRNA-protein fragment-based docking poses

Anna Kravchenko[1], Sjoerd Jacob de Vries[1], Malika Smaïl-Tabbone[1] and Isaure Chauvot de Beauchene[1]
[1] Université de Lorraine, CNRS, Inria, LORIA F-54000 Nancy, France

1 MOTIVATION

Single-stranded (ss) RNA-protein complexes are very challenging to model:
1. The *absence of an unbound ssRNA structure* due to its disorder prevents the use of classical docking methods;
2. *ssRNA flexibility* does not allow for exhaustive conformational sampling for long chains;
3. The relatively *small number of experimental ssRNA-protein structures* prevents the use of end-to-end deep learning for this problem.

Fragment-based docking approach using ATTRACT coarse-grained model [1] tackles this flexibility issue by splitting the ssRNA sequence into fragments small enough to allow their conformations to be exhaustively sampled within a given accuracy threshold, including near-bound conformation. However, this approach suffers from the **scoring problem**: the frequent inability of the ATTRACT scoring function [2] (ASF) to recognize correct poses among many incorrect poses. Since ASF parameters are not ssRNA-specific and were determined in 2010, we see a substantial opportunity for improvement.

2 APPROACH

We built a new scoring parameters set, based on the **ratio of bead-bead[1] distances in correct vs incorrect poses**, with the following procedure (see details in 6 PIPELINE):
1. Dock data cases (protein[2] and RNA fragment) with known 3D structure to generate a pool of docking poses;
2. Distinguish correct poses (LRMSD<5Å) from incorrect poses (LRMSD>7Å) based on the known 3D structure;
3. Count occurrences of each <protein bead type; RNA bead type> contact in the pool of correct and incorrect poses;
4. Derive a set of parameters as the occurrences ratio, in a form of a histogram set $\mathcal{H}$ (one bar per distance range) for each <bead type; bead type> pair.

The final $\mathcal{H}$ gives, for each <bead type; bead type> pair, for each bead-bead distance range, **the propensity of this distance to be seen in correct rather than incorrect pose**. In a real docking case, each pose is scored by summing, over all its RNA-protein contacts, the propensity of that distance to be correct for that pair of bead types.

As 1 set of histograms is not enough to cover the variety of ssRNA binding modes, we derived a **collection of 4 $\mathcal{H}$**. We apply each set independently to select a fraction of best-scored poses, then pool the 4 fractions together.

[1] groups of atoms in coarse-grained representation [2]
[2] to streamline the process, we focused on the RNA recognition motifs (RRMs) as this domain is especially relevant for ssRNAs and present in ~65% of ssRNA-binding proteins

3 RESULTS

Here we present HIPPO, a composite coarse-grained ssRNA-protein scoring potential derived analytically from contact frequencies in correct versus incorrect docking poses. We validated (see 6 PIPELINE, b) it on a benchmark of 57 experimentally solved RRM-ssRNA complexes, which consists of 217 non-redundant RRM-fragment[1] cases. HIPPO achieved a **3-fold or higher enrichment[2] in the 20% best-scored poses for half of the fragments**, versus only a quarter with ASF.

For most fragments, one $\mathcal{H}$ out of the collection of 4 picks up most of the correct selected poses. Thus, this enrichment would be greatly improved if one could predict which $\mathcal{H}$ to select for a given fragment. In particular, HIPPO would drastically improve the chance of a very high enrichment (12-fold or higher) of the best-scored fragment in the complex, a scenario where the incremental modelling of the entire ssRNA chain from one fragment becomes viable [3]. However, for the latter result, more research is needed to make it directly practically applicable. Regardless, our approach already improves upon the state of the art in ssRNA-protein modelling and is **extendable to other types of protein-nucleic acid interactions**.

[1] a fragment is a trinucleotide
[2] the list of 20% top-ranked poses contains 60% or more of all sampled NNs

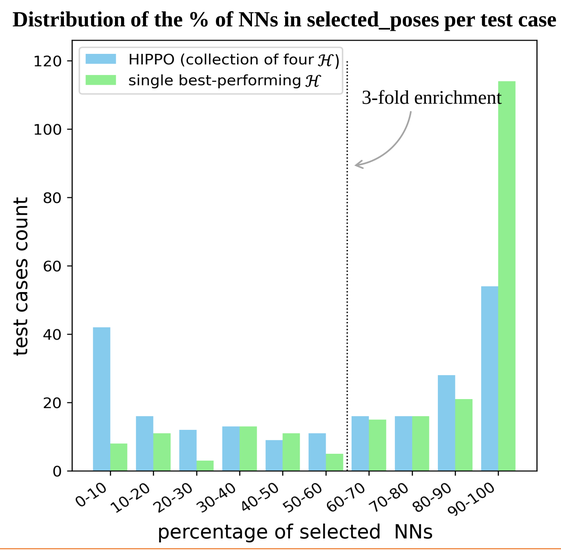
4 PERFORMANCE

HIPPO application entails independent scoring of all docking poses with each of four $\mathcal{H}$, followed by pooling together the top 5% of each scoring list.

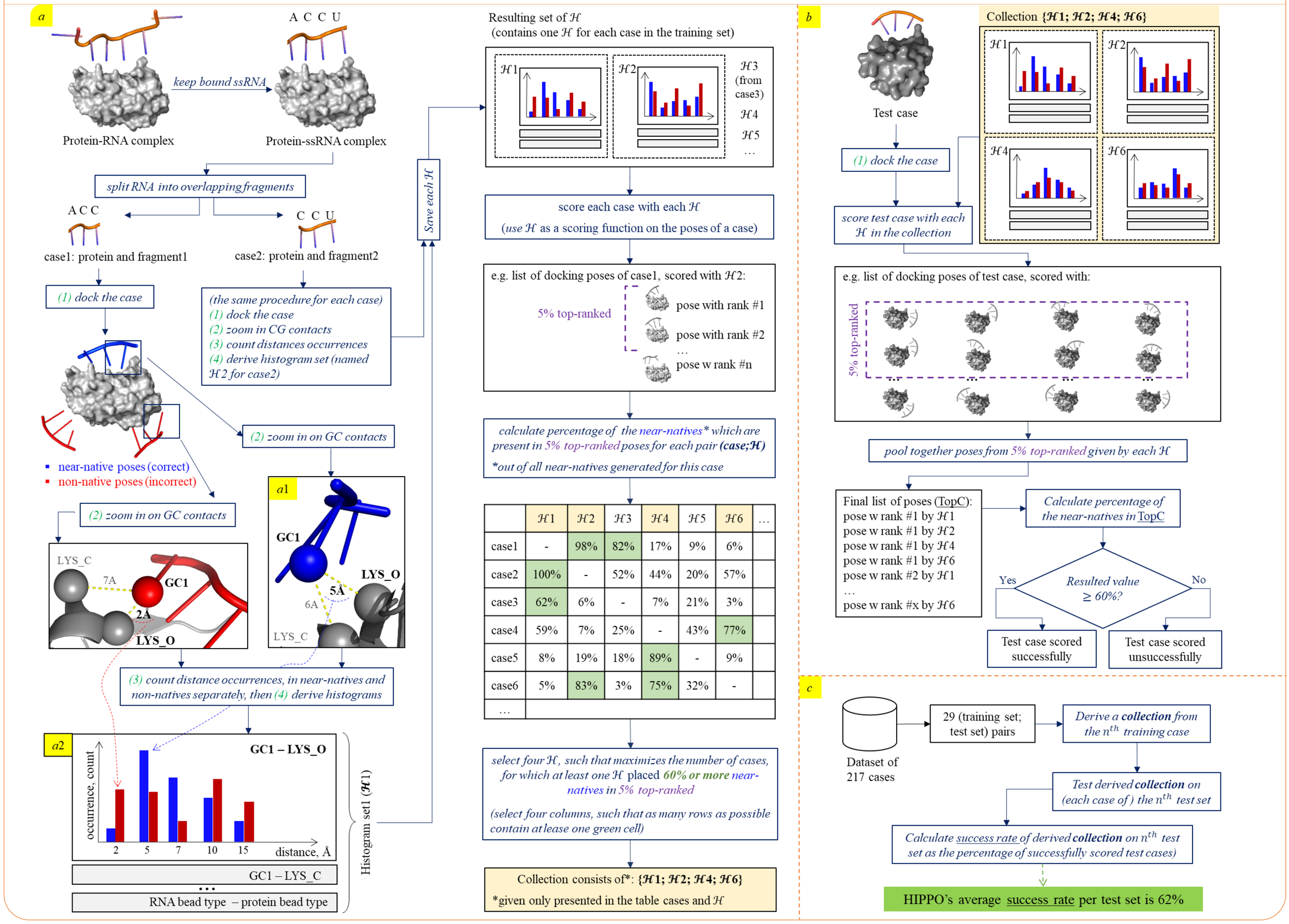
	ASF	HIPPO
% of cases with 60% of more correct poses selected	26	53
% of cases with 80% of more correct poses selected	7	38
% of correct poses selected, averaged over all test cases	43	55
Nb of cases with 80% or more correct poses selected	15	75
Highest % of correct poses selected among the cases of a complex, averaged over all test cases	60	72
Nb of complexes with 60% or more correct poses selected for at least one fragment	54	75
Nb of complexes with 80% or more correct poses selected for at least one fragment	9	33
Nb of complexes with higher % of correct poses selected for the best-scored fragment	18	39

5 PERSPECTIVES

- Use in incremental modelling approach
- Build classifier to select best $\mathcal{H}$ per case
- Use as a *sampling* function
- Apply to all-atom models
- Apply to general protein-ssRNA benchmark
- Apply to protein-ssDNA benchmark



6 PIPELINE



References:
[1] Chauvot de Beauchene I, de Vries SJ, Zacharias M. (2016). *NAR*
[2] Setny P, Zacharias M. (2011). *NAR*
[3] Chauvot de Beauchene I, de Vries SJ, Zacharias M. (2016). *NAR*

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