An integrative approach for predicting the RNA secondary structure for the HIV–1 Gag UTR using probing data

Afaf SAAIDI[1] supervised by Yann PONTY[1] and Bruno SARGUEIL[2]
[1] AMIBio team, Laboratoire d'informatique de l'école polytechnique, Inria Saclay, 91120, Palaiseau, France
[2] Laboratoire de cristallographie et RMN Biologiques, Faculté de pharmacie Paris Descartes, 75006, Paris, France

Corresponding author: yann.ponty@lix.polytechnique.fr

Structure modeling is key to understand the mechanisms of RNA retroviruses such as HIV. Many *in silico* prediction approaches suggesting structural models of moderate to good accuracies are available. However, the prediction methods could be further improved by taking advantage of both next generation sequencing technologies and different experimental techniques such as enzymatic and SHAPE probing data [1]. In a published article [2], we introduce and use a structural modeling method based on the integration of many experimental probing data to direct predictions with the aim to find the most accurate structure lying in the intersection of disjoint sources of experiments.

Method. High-throughput experimental data can be derived from two sources of experimental data: SHAPE [1] and enzymatic probing. We used the stochastic sampling [3] mode of `RNASubopt` to sample structural models from the Boltzmann distribution in a way that favors/forces compatibility with the derived constraints. Namely, SHAPE reactivity profiles were used as soft constraints, meaning that observed reactivity values were translated into pseudo-energies. Position-specific susceptibilities to RNAses cleavage were used as hard constraints by setting arbitrary cut-offs above which specific base are forced to be paired and unpaired. Both types of constraints reduce the space of possible conformations, leading to a set of structures that are maximally compatible with the provided data.

We posited that the optimal structure(s) should be energetically stable and supported by several experimental data. Thus, for each type of probe, we generate a set of structures compatible with experimentally-derived constraints. We merge those sets, and performed a structural distance-based clustering across experimental conditions, to generate several sets of structural models that are well-supported by experimental data. Clusters were scored using three criteria, namely, their stability, coherence and diversity (recurrence across structural conditions). Within the set of clusters returned by our clustering algorithm we elect clusters on the Pareto frontier, ie clusters that are not strictly dominated with respect to these three criteria. Finally, representative structures (centroids), corresponding to Maximum Expected Accuracy structures are built and returned.

Results. The HIV-1 sequence probed in this study corresponds to the 5'UTR preceding the gag coding region from the NL-4.3 strain (Genbank: AF324493.2). A sample set of 12 000 structures, covering 6 sources of experimental data, was generated. The clustering step led us to elect two optimal clusters, whose corresponding centroids were then assessed in the light of their compatibility with specific SHAPE data. This allowed us narrow down our proposed models to a single candidate, whose base pair conservation/covariation was confirmed by comparative analysis.

Conclusion and perspectives. Our integrative approach allowed us to implement a consensus structure compatible with many different experimental probing data. As further work, we project to use our approach with other sources of experimental probing data and to exploit the alignment to build an additional constrained sample instead of using it for validation.

References

[1] K. A Wilkinson, E. J Merino, and K. M Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols*, 1(3):1610–6, 2006.

[2] J. Deforges, S. de Breyne, M. Ameur, N. Ulryck, N. Chamond, A. Saaidi, Y. Ponty, T. Ohlmann, and B. Sargueil. Two ribosome recruitment sites direct multiple translation events within HIV1 Gag open reading frame. *Nucleic Acids Research*, 2017.

[3] Y. Ding and C Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.