Towards structural models for the Ebola UTR regions using experimental SHAPE probing data

Afaf Saaidi, Delphine Allouche, Bruno Sargueil, Yann Ponty

Towards structural models for the Ebola UTR regions using experimental SHAPE probing data

Afaf Saaidi[1,2], Delphine Allouche[3], Bruno Sargueil[*,3] and Yann Ponty[*,1,2]

[1]CNRS/LIX, Ecole Polytechnique, Palaiseau
[2]AMIB team, Inria Saclay
[3]CNRS UMR 8015, Laboratoire de cristallographie et RMN Biologiques, Université Paris Descartes, Paris

Abstract

Next-Generation Sequencing (NGS) technologies have opened new perspectives to refine the process of predicting the secondary structure(s) of structured non-coding RNAs. Herein, we describe an integrated modeling strategy, based on the SHAPE chemistry, to infer structural insight from deep sequencing data. Our approach is based on a pseudo-energy minimization, incorporating additional information from evolutionary data (compensatory mutations) and SHAPE experiments (reactivity scores) within an iterative procedure. Preliminary results reveal conserved and stable structures within UTRs of the Ebola Genome, that are both thermodynamically-stable and highly supported by SHAPE accessibility analysis.

Keywords RNA, Ebola, secondary structure prediction, SHAPE chemistry, High-throughput sequencing, compensatory mutations.

1 Context

The Ebola virus causes an acute, serious illness which is often fatal if untreated. A promising research direction would be to selectively interfere with the expression of Ebola genes. This requires a mechanical understanding, at a molecular level, of the processes underlying gene regulation. Many such processes revolve around the presence of specific secondary structure elements [7] in untranslated regions [2].

RNA structure prediction can either be performed computationally from thermodynamics principles in the presence of only the sequence [10], or via a comparative approach when an aligned set of homologous sequences is available [1].

Experimentally, such data can be supplemented by chemical probing data, leading to more accurate predictions in the context of RNA structure determination [3].
Our approach is based on a pseudo-energy minimization, incorporating additional information from evolutionary data (compensatory mutations) and SHAPE experiments (reactivity scores) within an iterative procedure.

*To whom correspondance should be addressed

RNA folding can be approached in a less deterministic setting, by considering the outcome of the folding process as a dynamic ensemble of structures. In this vision, a sequence actively fluctuates between many alternative conformations through time, leading to a probability distribution over the set of structures.
The challenge is to take advantage from those probabilities to determine the most probable structures that verify the set of experimental constraints and thus help in revealing the Ebola virus replication mechanism.

SHAPE probing principles SHAPE (Selective 2'Hydroxyl Acylation analyzed by Primer Extension) chemistry is a prominent experimental method, when used in combination with structural modeling methods, it can lead to finding reliable structures [9]. This experimental technique is based on the addition of reagents that interact with the phosphate-sugar backbone on particularly flexible positions, inducing the formation of an adduct. Upon exposure to the reverse-transcriptase (RT), the adduct either causes the mutation of the nucleotide (SHAPE-MAP [9]), or causes the early detachment of the RT (SHAPE-CE [3] , SHAPE-Seq [5]). A quantification of these effects allows to assign a Reactivity score to each position to refer to its ability to interact with the reagent. Unpaired nucleotides are predominantly reactive compared to the paired bases, the latter being constrained by their Hydrogen bonds. Hence, the reactivity of a nucleotide can be used to inform *ab initio* structure modeling, or to validate homology-based predictions.

SHAPE MAP reactivities In the context of SHAPE-Map, reactivities are computed by comparing three mutation rates observed in different experimental conditions: presence (Shape)/absence (Control) of SHAPE reagent, and in the absence of structure (Denatured). The reactivity of a position n is calculated as :

$$\mathsf{Reactivity}(n) = \frac{mut_{\mathsf{Shape}}(n) - mut_{\mathsf{Control}}(n)}{mut_{\mathsf{Denatured}}(n)}$$

2 Material and methods

2.1 Dataset

The Ebola virus genome is 19 kb long, with seven open reading frames in the order 3' NP-VP35-VP40-GP-VP30-VP24-L 5', encoding for structural proteins, envolope glycoprotein, nucleoprotein, non structural proteins and viral polymerase. For each frame, we study consecutively the 5' and 3' non coding regions, this results into 14 sequences ranging from 3'NP to 5'L.

The Shape experiments are performed using 1M7 as reagent (weeks et al [8]). We also evaluated the differential SHAPE experiment that uses an additional reagent NMIA, in order to increase the accuracy of RNA structural models.

The experiments are followed by a sequencing and mapping process [8]. Each position in the RNA sequence is characterized by its occurrence from the throughput sequencing and thus by a mutation rate. ShapeMapper program(weeks et al [8]), after being adapted to Single End reads input, is subsequently called to calculate reactivity scores.

2.2 Modeling tools

We use SHAPE reactivities within an iterative modeling strategy based on a combination of predictive methods:

1. Free-Energy minimization. `RNAfold` predicts the most thermodynamically stable structure compatible with an RNA sequence [4]. It performs an exact optimization, using dynamic programming, of the free-energy within the Turner energy model (Zuker-Stiegler algorithm [10]).

2. Partition function analysis. `RNAfold` also allows analyzing the folding landscape at the thermodynamic equilibrium. An efficient dynamic-programming scheme (MCcaskill algorithm [6]) produces the base-pairing probability matrix, at the Boltzmann equilibrium. Those probabilities are used to provide a support for predicting base-pairs. Our current method is based on finding the more reliable pairwises according to their probabilities, then use them as structural constraints to recalculate the energy values.

3. Comparative analysis. `RNAalifold` computes the consensus structure for a set of – previously aligned – homologous sequences. It optimizes a credibility score (Hofacker algorithm [1]), which primarily depends on compensatory mutations.

2.3 Main workflow

The present workflow relies on a conservative integrated approach where sets of structures are retrieved from different methods in order to detect similar substructures.
Methods 1, 2 are called to build a first set of structures.

The method 3 is used to reveal the most credible base pairs according to the internal scoring scheme of `RNAalifold`. These base-pairs are then considered as constraints within a new run of RNAfold, thus a second ensemble of structures is built.

Substructures that are found in the both datasets are kept and the pairing probabilities that are higher than a certain threshold are introduced as structural constraints for the next run. This process is performed iteratively until no additional bases is predicted. Contrasting with typical methods, which include SHAPE data in the iterative modeling strategy, we chose to keep such a data for a final validation step. In this last step, the compatibility between predicted single/double stranded regions and SHAPE induced accessibilities is asessed. The resulting substructures are eminently supported by the SHAPE reactivities, as shown in Figure 1.

3 Preliminary results and discussion

Our preliminary suggest the existence of conserved and stable hairpin loops in the UTR regions of the Ebola genome are structured into hairpin-loop substructures stemming from the exterior region. Moreover, we found little evidence for complex tree-like structures that are the landmark of structural ncRNAs. One such example is shown in Figure 1. The method proposed features the most conserved stem-loop substructures. The first stem-loop of the 5'-UTR is highly reminiscent of typical structures found in viral genomes. By analogy, we hypothesize that the role of this substructure is to protect the RNA from being degraded by nucleases. The SHAPE data is generally consistent with our predictions, and indeed labels as accessible most single-stranded regions in predicted hairpins. Conversely, helices are labeled as inaccessible.

We plan to further improve our modeling strategy by including Boltzmann sampling and structure-based clustering instead of our current voting mechanisms. We also wish to include additional information, such as SHAPE experimental data using different reagents, possibly including probing on different homologous sequences to establish a SHAPE-informed structural consensus.

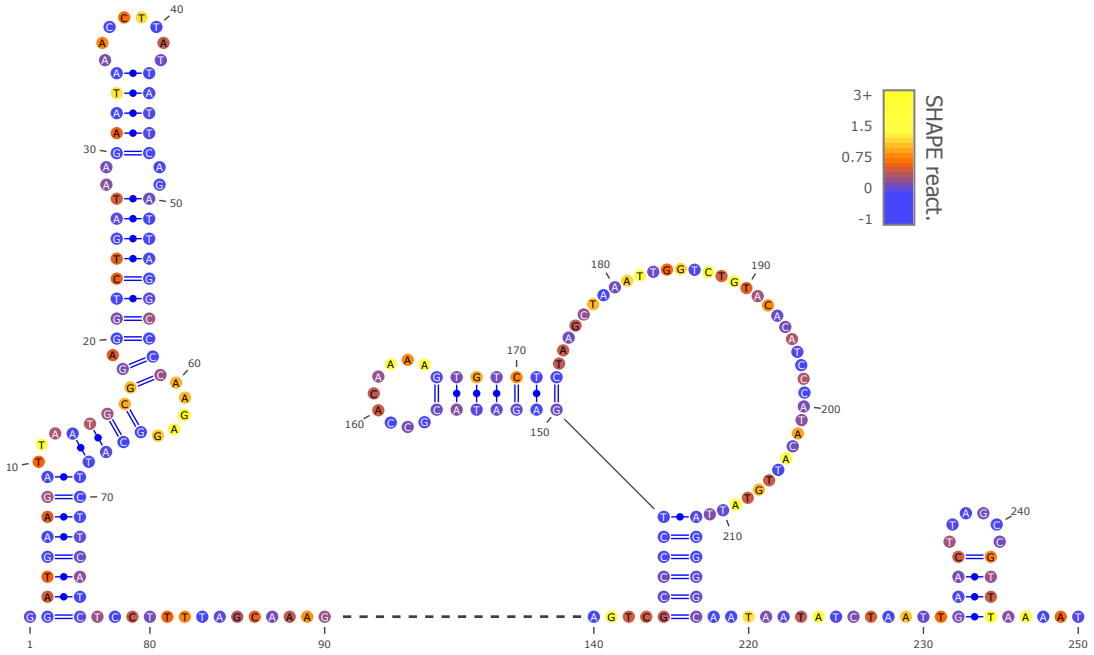

Figure 1: Secondary structure inferred for the 5'UTR of the VP24 gene colored by SHAPE reactivity.

Acknowledgment

This work is supported by FRM Fondation de la Recherche Medicale. The authors would like to thank Steven Busan (Department of Chemistry, University of North Carolina), Alice Heliou (AMIB, INRIA saclay), Vladimir Reinharz (McGill University, Montreal) and Jules Desforges (Faculté de biologie et médecine, Lausanne) for their valuable feedback and support.

Bibliography

[1] S H Bernhart, I L Hofacker, S Will, A R Gruber, and P F Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

[2] S M Crary, J S Towner, J E Honig, T R Shoemaker, and S T Nichol. Analysis of the role of predicted RNA secondary structures in Ebola virus replication. *Virology*, 306(2):210–218, 2003.

[3] Fethullah Karabiber, Jennifer L McGinnis, Oleg V Favorov, and Kevin M Weeks. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA (New York, N.Y.)*, 19(1):63–73, 2013.

[4] Ronny Lorenz, Stephan H Bernhart, Christian zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. {ViennaRNA} Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[5] Julius B. Lucks. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic acids research*, 42(21), 2014.

[6] J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[7] Masfique Mehedi, Thomas Hoenen, Shelly Robertson, Stacy Ricklefs, Michael A. Dolan, Travis Taylor, Darryl Falzarano, Hideki Ebihara, Stephen F. Porcella, and Heinz Feldmann. Ebola Virus RNA Editing Depends on the Primary Editing Site Sequence and an Upstream Secondary Structure. *PLoS Pathogens*, 9(10), 2013.

[8] Matthew J Smola, Greggory M Rice, Steven Busan, Nathan A Siegfried, and Kevin M Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature protocols*, 10(11):1643–1669, 2015.

[9] Kevin A Wilkinson, Edward J Merino, and Kevin 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.

[10] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.