SHAPE directed RNA folding

Dominik Luntzer[1], Ronny Lorenz[1], Ivo L. Hofacker[167], Peter F. Stadler[123478], Michael T. Wolfinger[15]

[1]Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria [2]Bioinformatics Group of the Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany [3]Interdisciplinary Center for Bioinformatics of the University of Leipzig [4]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany [5]Center for Integrative Bioinformatics Vienna (CIBIV) & Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria [6]Bioinformatics and Computational Biology Research Group, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria [7]Center for RNA in Technology and Health, Univ. Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark [8]Fraunhofer Institute for Cell Therapy and Immunology, Perlickstrasse 1, D-04103 Leipzig, Germany [9]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA

Contact: ronny@tbi.univie.ac.at - http://www.tbi.univie.ac.at

1. Introduction

The spatial structure of RNA plays an important role in genome regulation because it critically influences the interaction with proteins, and other nucleic acids. Knowledge of RNA structure is therefore crucial for understanding various biological processes.

Chemical and enzymatic probing methods provide information concerning the flexibility and accessibility at nucleotide resolution. As these methods are becoming a frequently used technology to experimentally determine RNA structure, for instance in terms of nucleotide-wise flexibility of the RNA backbone (SHAPE), there is increasing demand for efficient and accurate computational methods that incorporate probing data into secondary structure prediction. Existing implementations such algorithms, e.g. provided by the *ViennaRNA Package*, typically yield excellent prediction results for short sequences. However, accuracy decreases to between 40% and 70% for long RNA sequences due to imperfection of the thermodynamic parameters, and inherent limitations of the secondary structure model, such as tertiary interactions, pseudoknots, ligand binding, or kinetic traps. To alleviate the gap in available computational tools we have developed a framework for incorporating probing data into the structure prediction algorithms of the *ViennaRNA Package* by means of soft constraints in order to improve prediction quality.

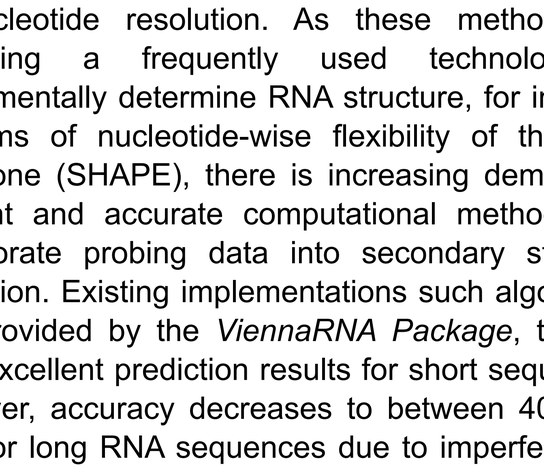

Fig. 1 RNA secondary structure of yeast tRNA-asp annotated with experimentally determined SHAPE reactivities.

2. Methods

Soft constraints guide the folding prediction by adding position-, or motif-specific pseudo-energy contributions to the free energies of certain loop motifs. This amounts to a distortion of the equilibrium ensemble of structures in favour of those that are consistent with experimental data. Mismatching motifs are penalized by positive contributions, while structure patterns where prediction and experiment agree with each other receive a “bonus” in form of a negative pseudo-energy. Current methods for guided secondary structure prediction by means of soft constraints mainly focus on the incorporation of SHAPE reactivity data. For that purpose, three algorithms are available that aim to transform normalized SHAPE reactivity data into meaningful pseudo-energy terms.

The first approach that applied SHAPE directed RNA folding uses the simple linear ansatz

$$\Delta G_{\mathrm{SHAPE}}(i) = m \ln(\mathrm{SHAPE\ reactivity}(i) + 1) + b$$

to convert SHAPE reactivity values to pseudo energies whenever a nucleotide i contributes to a stacked pair (*Deigan et al., 2009*). A positive slope m penalizes high reactivities in paired regions, while a negative intercept b results in a confirmatory *"bonus"* free energy for correctly predicted base pairs.

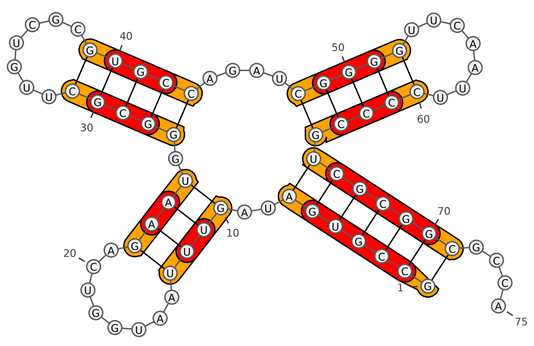

Fig. 2 RNA secondary structure of yeast tRNA-asp with indication where SHAPE reactivity derived pseudo-energies using the Deigan et al. approach are applied in the folding prediction. As a consequence of the method, pseudo-energies are applied twice for pairs inside a helix, and just once for terminal pairs.

A more consistent model considers nucleotide-wise experimental data in all loop energy evaluations (*Zarringhalam et al., 2012*). First, the observed SHAPE reactivity of nucleotide i is converted into the probability q_i that position i is unpaired by means of a non-linear map. Then pseudo-energies of the form

$$\Delta G_{\mathrm{SHAPE}}(x, i) = \beta |x_i - q_i|$$

are computed, where $x_i = 0$ if position i is considered unpaired and $x_i = 1$ if it is involved in a base pair. While the parameter β serves as scaling factor, the magnitude of discrepancy between prediction and experimental observation is represented by $|x_i - q_i|$.

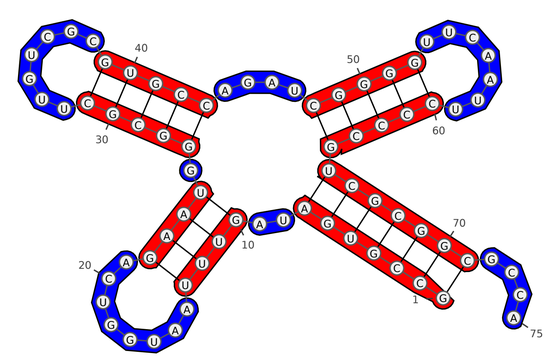

Fig. 3 RNA secondary structure of yeast tRNA-asp where structural parts that receive a bonus(malus) energy according to the Zarringhalam et al. method are highlighted in red (paired nucleotides) and blue (unpaired nucleotides).

A third, very distinct approach on incorporating SHAPE reactivity data to guide secondary structure prediction was suggested by *Washietl et al.* (2012). Here, the authors phrase the choice of the bonus energies as an optimization problem that aims to find a perturbation vector $\vec{\epsilon}$ of pseudo-energies that minimizes the discrepancy between the observed and predicted probabilities to see particular nucleotides unpaired, q_i and p_i, respectively. At the same time, the perturbation should be as small as possible.

$$F(\vec{\epsilon}) = \sum_{\mu} \frac{\epsilon_{\mu}^2}{\tau^2} + \sum_{i=1}^{n} \frac{(p_i(\vec{\epsilon}) - q_i)^2}{\sigma^2} \rightarrow \min$$

The tradeoff between the two goals is naturally defined by the relative uncertainties inherent in the SHAPE measurements and the energy model.

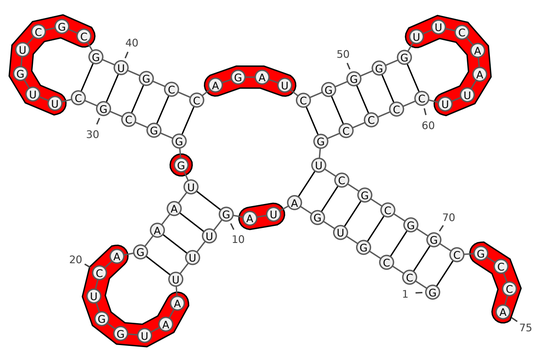

Fig. 4 RNA secondary structure of yeast tRNA-asp with highlighted unpaired nucleotides that receive a pertubation pseudo-energy according the method of Washietl et al.

3. Availability

All three methods outlined above have been implemented into the *ViennaRNA Package 2.2*. Additional functionalities are available through the API of the *ViennaRNA* Library and the command line interface of *RNAfold* and *RNAalifold*. The novel standalone tool *RNApvmin* dynamically estimates a vector of pseudo-energies according to the method of Washietl et al. (2012), which can be used to guide structure prediction with *RNAfold*. This setup makes it easy for users to incorporate alternative ways of computing bonus energies, or to use the software with other types of probing data. Guided structure prediction has also been included into the *ViennaRNA Websuite*, a Web server providing an interface to many tools of the ViennaRNA Package, available at *http://rna.tbi.univie.ac.at* .

4. Results

We applied all three methods to a benchmark set containing 24 triples of sequences, their known reference structures, and corresponding SHAPE data. In this set, reference structures were either derived from X-ray crystallography experiments, or predicted by comparative sequence analysis. The use of SHAPE data driven soft constraints leads to improved prediction results for many RNAs. However, for some of the RNAs within our benchmark data the additional pseudo-energy terms impair prediction results, possibly due to several factors. First, experimental data may be inaccurate, and second the underlying energy model excludes pseudoknotted structures, which are present in approximately half of the benchmarked RNAs. From our benchmark we conclude, that none of the three implemented methods consistently outperforms the other two in terms of prediction performance.

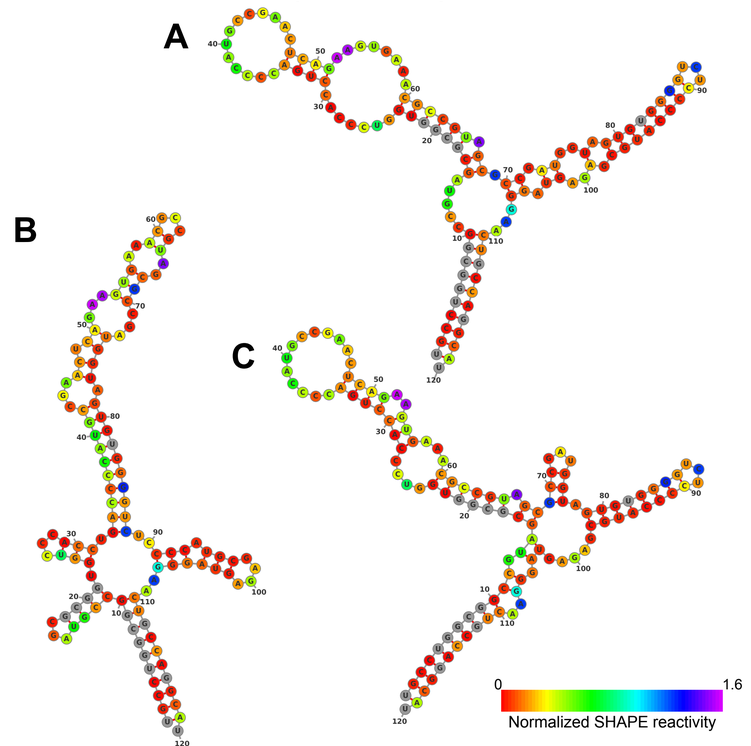

Fig. 5 Secondary structure prediction of E.coli 5S rRNA from our benchmark data set. **A** Structure reference, **B** prediction by *RNAfold* with default parameters, and **C** prediction by *RNAfold* with guiding pseudo-energies obtained from SHAPE reactivity data using *RNApvmin*. Grey nucleotides correspond to missing SHAPE reactivity data.

5. Acknowledgements

This work was partly funded by the Austrian Science Fund (FWF) project “RNA regulation of the transcriptome” (F43), Deutsche Forschungsgemeinschaft (DFG) STA 850/15-1 and the German ministry of science (0316165C as part of the e:Bio initiative).

[1] Luntzer, D., Lorenz, R., Hofacker, I.L., Stadler, P.F, Wolfinger, M.T. (2015), **Shape directed RNA folding.** *submitted*

[2] Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009). **Accurate SHAPE-directed RNA structure determination**. *PNAS, 106, 97–102.*

[3] Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H., and Clote, P. (2012). **Integrating chemical footprinting data into RNA secondary structure prediction**. *PLOS ONE, 7(10).*

[4] Washietl, S., Hofacker, I. L., Stadler, P. F., and Kellis, M. (2012). **RNA folding with soft constraints: reconciliation of probing data and thermodynamics secondary structure prediction**. *Nucleic Acids Research, 40(10), 4261–4272.*