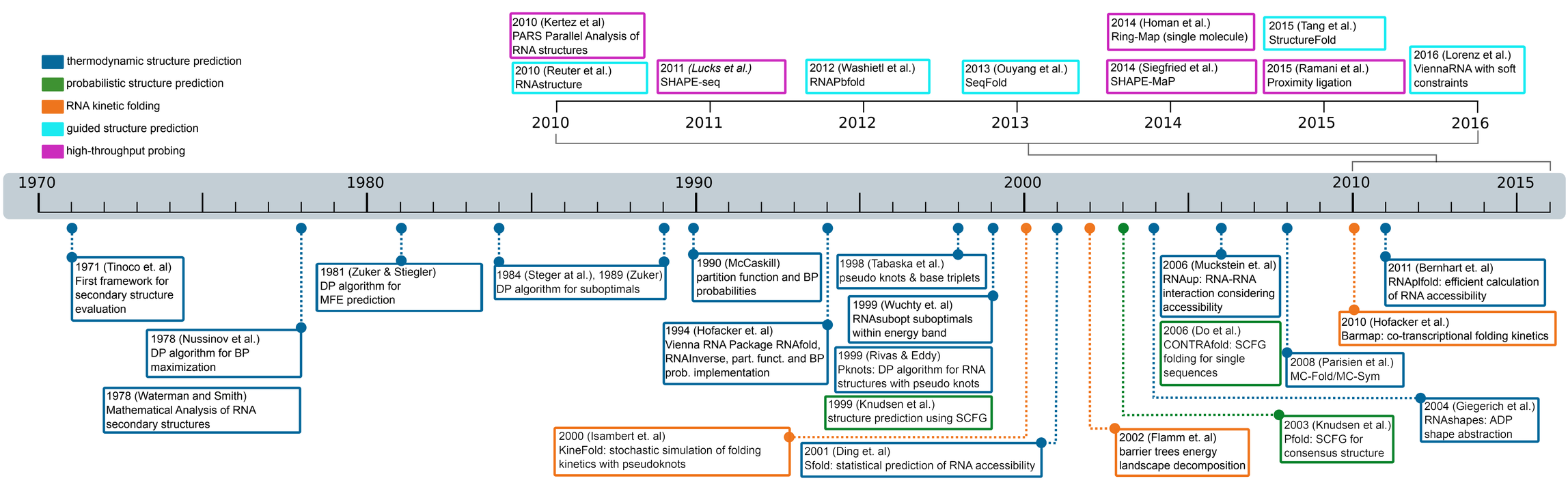
Probing assisted RNA folding

Michael T. Wolfinger[13], Ronny Lorenz[1], Andrea Tanzer[1], Ivo L. Hofacker[12]

[1]Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria
[2]Bioinformatics and Computational Biology Research Group, University of Vienna, Währingerstraße 17, 1090 Wien, Austria
[3]Center for Anatomy and Cell Biology, Medical Unversity of Vienna, Währingerstraße 13, 1090 Wien, Austria

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



1. Introduction

RNA function is determined by RNA structure and therefore knowledge of the spatial structure of RNA is an asset for understanding various biological processes such as RNA regulation. Chemical and enzymatic probing methods such as SHAPE allow for fine-grained assessment of RNA structure at nucleotide resolution. The advent of high-throughput structural probing such as SHAPE-seq or PARS has spurred the development of computational techniques that incorporate such experimental data as auxiliary information.

Popular RNA folding algorithms, as implemented e.g. in 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.

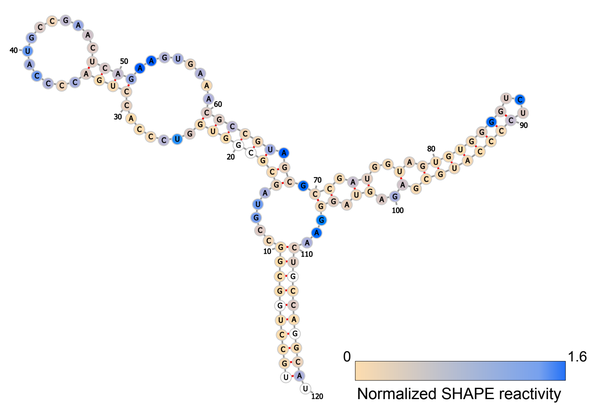


Fig. 1 RNA secondary structure of E.coli 5S rRNA annotated with experimentally determined SHAPE reactivities.

2. Soft Constraints

To alleviate the gap in available prediction tools we have developed a framework for incorporating probing data into the structure prediction algorithms of the *ViennaRNA Package* by means of *Soft Constraints* that guide the folding prediction by adding position-, or motif-specific pseudo-energy contributions to the free energies of certain loop motifs.

We have recently implemented previously published methods to incorporate SHAPE probing data into the ViennaRNA Package [1], two of which include an ad-hoc conversion of SHAPE reactivities into pseudo free energies. Later approaches first convert reactivities into probabilities of being (un)paired and compute pseudo energies from these likelihoods.

3. Probabilistic RNA folding

The conversion of RNA structure probing data into pairing probabilities is not trivial. In fact, reactivity values measured for different structural contexts, e.g. paired and unpaired bases, are similar and can therefore not be well separated. Thus, there is no simple way to infer whether a given nucleotide is paired just based on raw readout.

Using probing data for a reference RNA with known secondary structure, however, allows one to derive distributions of the measured reactivity values for different structural contexts. These distributions can then be fitted to a probability density model to compute for each nucleotide i the conditional probability $P(r_{\mathcal{S}}(i)|\pi_i)$ to observe a reactivity $r_{\mathcal{S}}(i)$ given its structural context π_i . Eddy [3] already suggested to convert these conditional probabilities into a pseudo energy

$$\Delta G_{\mathcal{S}}(\pi_i, i) = -RT \log P(r_{\mathcal{S}}(i)|\pi_i)$$

and apply it to each derivation of the MFE algorithm, where i is added to a growing substructure. As a consequence, the soft constrained MFE structure maximizes the probability to observe to observe the probing data.

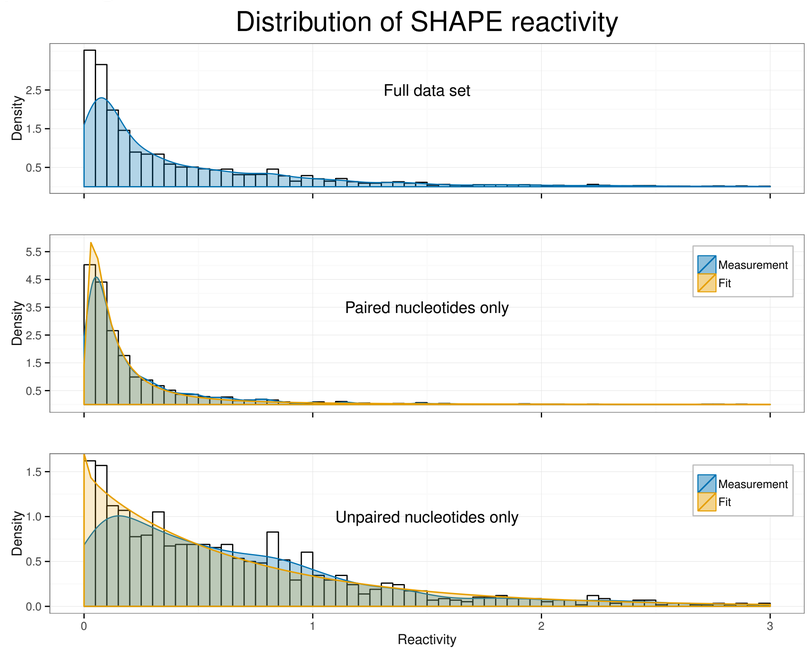


Fig. 2 SHAPE reactivity distributions for E.coli 23S rRNA.

Using Bayes' rule, the posterior probability of a structure context π_i given its reactivity $r_{\mathcal{S}}(i)$ is

$$p(\pi_i|r_{\mathcal{S}}(i)) = \frac{P(r_{\mathcal{S}}(i)|\pi_i) \cdot p(\pi_i)}{p(r_{\mathcal{S}}(i))}.$$

Still, the probabilities $p(\pi_i)$ and $p(r_{\mathcal{S}}(i))$ are unknown a priori and can only be estimated from training data. An ad-hoc implementation of this idea is provided by the RME program [4].

4. A self-consistent method

To overcome the dependency on training data and thus abandon ad-hoc assumptions inherent in previous methods, the reactivity distribution of each distinguished structure context must be inferred from the data itself. Therefore, the observed reactivities, i.e. the mixture of distributions, needs to be deconvoluted. This, however, is by far not an easy task.

Nevertheless, under the assumption that the RNA's structure ensemble is dominated by a single conformation we can use computed equilibrium probabilities $p(\pi_j)$ and parameterized model distributions Γ^j to obtain

$$p(r_S(i)) = \sum_j p(\pi_j) \cdot \Gamma^j,$$

i.e. the probability to observe $r_{\mathcal{S}}(i)$.

Moreover, the likelihood to observe the measured pattern of probing data is

$$L(r_S) = \prod_i p(r_S(i)).$$

Now, the aim is to find a parameterization for the distributions Γ^j that maximizes $L(r_S)$.

We propose to iteratively use the posterior probabilties $p(\pi_i|r_{\mathcal{S}}(i))$ as soft constraint to update the probabilities $p(\pi_j)$ for the next round. This strategy is then applied until convergence.

5. Outlook

In this self-consistent framework it is even possible to optimize for a combination of individual probing techniques k, such as Pb(II), SHAPE, DMS, PARS, etc.:

$$\Delta G_\Psi = \sum_k \Delta G_\Psi^k$$

In a PARS experiment, for example, rather than computing log-odds of nuclease S1 and V1 treatment, their intensities can be independently converted into pseudo energy contributions.

6. Acknowledgements

This work was partly funded by the Austrian Science Fund FWF project "RNA regulation of the transcripto-me" (F43) and the Austrian/ French project "RNA-Lands" (FWF-I-1804-N28 and ANR-14-CE34-0011).

[1] Lorenz, R., Luntzer, D., Hofacker, I.L., Stadler, P.F, Wolfinger, M.T. (2016), **SHAPE directed RNA folding**. *Bioinformatics 32, 145-14*.
[2] Lorenz, R., Wolfinger, M.T., Tanzer, A. and Hofacker, I.L. (2016), **Predicting RNA Structures from Sequence and probing Data.** *Methods.*
[3] Eddy, S.R. (2014), **Computational analysis of conserved RNA secondary structure in transcriptomes and genomes**. *Annu Rev Biophys 43, 433-456.*
[4] Wu, Y. et al. (2014), **Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data**. *Nucleic Acid Res 43, 7247-7259.*