RNA Structure Prediction with Constraints

Ronny Lorenz[1], Ivo L. Hofacker[123], Peter F. Stadler[413567]

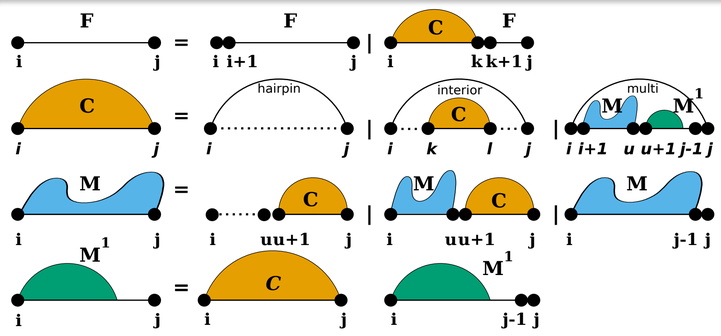
[1] Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria.
[2] Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währinger Straße 17 A-1090 Wien, Austria.
[3] Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg, Denmark.
[4] Bioinformatics Group, Dept. of Computer Science, and Interdisciplinary Center for Bioinformatics, University Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.
[5] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.
[6] Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany.
[7] Santa Fe Institute, 1399 Hyde Park Rd., NM87501 Santa Fe, USA.

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

1. Introduction

Although being successful in a wide variety of applications, pseudo-knot free **RNA secondary structure prediction is by no means perfect**. This not only applies to physics- or SCFG-based approaches, but for all popular methods to date. A most intuitive way to improve the quality of physics-based methods that use the standard *Nearest Neighbor (NN) energy model* is to incorporate experimental evidence to guide the structure prediction. For some data, this only requires constraints on the production rules of the RNA folding grammar (Fig. 1). More elaborate approaches depend on an extension of the grammar itself.

With the `ViennaRNA Package` we offer a generic yet systematic way to augment structure prediction. Our approach is most flexible by handing over the control for derivation and energy evaluation of the implemented RNA folding grammar to the user. It enables rapid tuning of the decomposition scheme without caring for our implementation of the recursions. Particular application scenarios such as to incorporate experimental RNA structure probing data or RNA-ligand binding can therefore be achieved almost in no time.



RNA secondary structure decomposition scheme.
Each decomposition can be subject to hard constraints that decides whether it is taken into account, or soft constraints that alter the energy evaluation by adding "bonus energies".

2. Structure Constraints

The recursive nature of the RNA folding grammar limits constraints to those that act independently on individual production rules. In each step two kinds of conceptually different constraints can be applied:
(i) *Hard constraints* that limit the candidate space by pruning particular derivation trees, and
(ii) *Soft constraints* acting on the evaluation level by adding "bonus energies".

We efficiently encode **hard constraints** as an upper-triangular Boolean matrix $\mathbb{X}^\tau$. Entries x_{ij}^τ determine whether base pair (i,j) may be part of a loop τ. Diagonal entries x_{ii}^τ store if a single nucleotide may be unpaired. A recursive structure counting algorithm then becomes

$$N_{ij} = \mathbb{X}_{ii} N_{i+1,j} + \sum_{k=i}^{j} \mathbb{X}_{ij} N_{i+1,j+1} N_{k+1,j}$$

To uniquely address all m nucleotide positions in decomposition $d \in D$ we use a Boolean function

$$f : \mathbb{N}^m \times D \to 0,1$$

Instead of restricting the candidate space, **soft constraints** constitute a bias in the ensemble of solutions. Again, we use an upper triangular matrix Δ where each entry δ_{ij} holds the auxiliary energy contribution of base pair (i,j) and diagonal entries δ_{ii} are used for contributions of a single nucleotide i. Note, that contributions b_i of both states, paired and unpaired, can be encoded as a single value δ_i, since the biased energy $E(\psi)$ of any structure ψ is

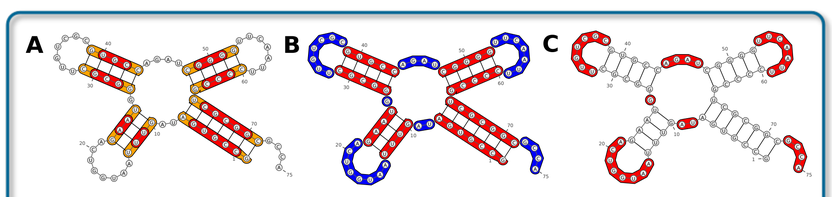
$$E(\psi) = E_0(\psi) + \sum_{i=1}^{n} b_i^p + \sum_{i \in \psi^u} (b_i^u - b_i^p) = E_0(\psi) + E' + \sum_{i \in \psi^u} \delta_i$$

Full control over all decomposition steps is achieved through a callback mechanism with

$$f : \mathbb{N}^m \times D \to \mathbb{R}$$

3. Structure Probing Data

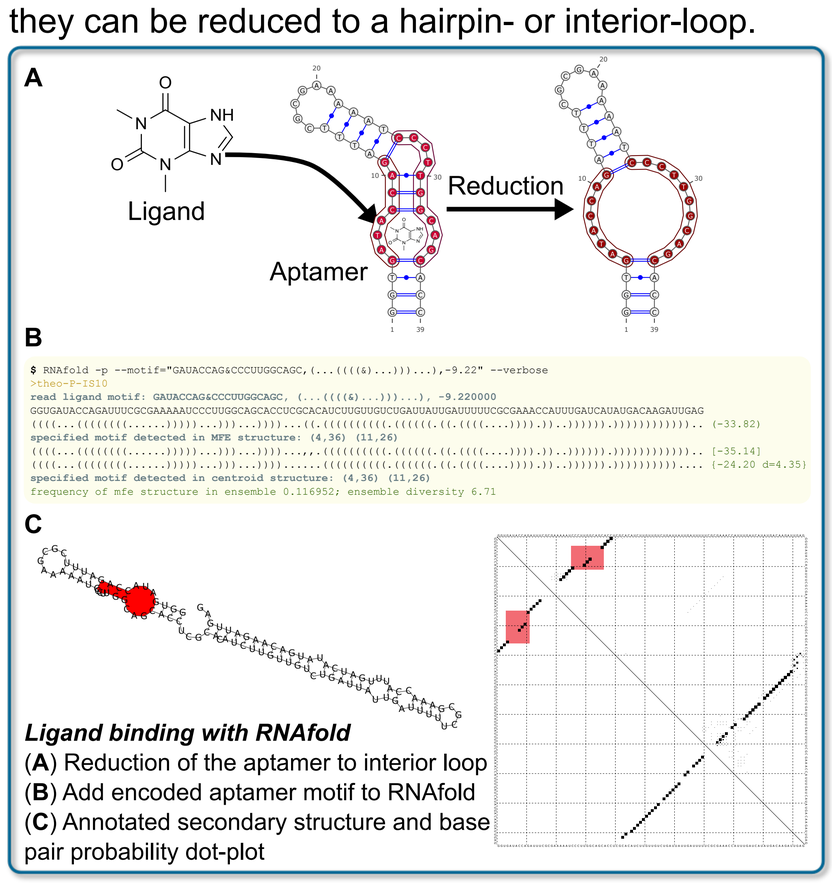
Chemical and enzymatic probing of RNA structures reveals at nucleotide resolution if a nucleotide is more likely to be paired, or unpaired. Especially the coupling of probing methods with high-throughput *Next Generation Sequencing (NGS)* generates massive amounts of data suitable to **guide structure prediction**. In most cases, such data is simply converted into pseudo-energies and added to certain structure configurations. More recent methods tend to use statistical background models to first convert probing data into configuration likelihoods and only then into energy terms.



Typical configurations SHAPE reactivity data is applied to.
(**A**) Deigan et al. (2009) method, stacked nucleotides
(**B**) Zarringhalam et al. (2012) method, paired and unpaired nucleotides
(**C**) Washietl et al. (2012) method, unpaired nucleotides

4. Self-enclosed Loops and Ligands

Our generic soft constraint implementation enables almost effortless modeling of ligands that bind to self-enclosed loops of the RNA, i.e. loops that appear as a single decomposition in the RNA folding grammar. More complex aptamers can be handled if they can be reduced to a hairpin- or interior-loop.

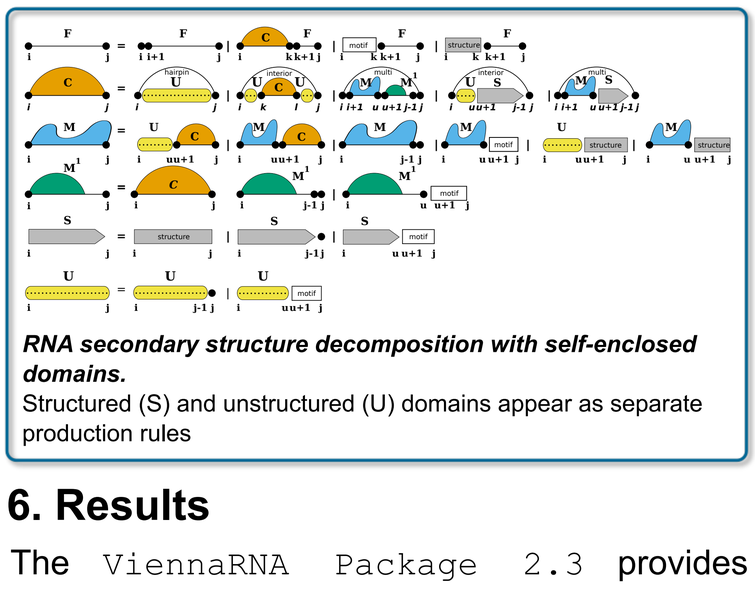


Ligand binding with RNAfold
(**A**) Reduction of the aptamer to interior loop
(**B**) Add encoded aptamer motif to RNAfold
(**C**) Annotated secondary structure and base pair probability dot-plot

5. Self-enclosed Domains

Still, our constraints framework alone cannot treat arbitrary sequence intervals as a binding site, e.g. for protein binding. **Interval constraints** for exterior- and multi-branch loops require a modification of the RNA folding grammar itself such that the entire interval appears as a single decomposition rule.

For that purpose, we introduce additional production rules that account for self-enclosed structured and unstructured domains. This enables us to distinguish between intervals that exhibit unusual intramolecular base pairing, such as G-Quadruplexes, and base pair free intervals that interact with external factors such as single strand binding proteins.



RNA secondary structure decomposition with self-enclosed domains.
Structured (S) and unstructured (U) domains appear as separate production rules

6. Results

The `ViennaRNA Package 2.3` provides ready-to-use programs that easily incorporate structure constraints and RNA-ligand binding into secondary structure prediction. We provide convenient interfaces to process input of *SHAPE* reactivity data, simple structure constraints, and motif-based RNA-ligand binding. Our `C`-library and the corresponding `Perl` and `Python` interfaces grant full access to the constraints and domain extension features including the callback mechanism that modifies the RNA folding grammar. Hence, new approaches to include structure probing data or RNA-ligand interaction can be treated as *add-ons* rather than requiring a re-implementation of all recursions.

```
F i 0 k    [TYPE] [ORIENTATION] # Force nucleotides i...i+k-1 to be paired
F i j k    [TYPE] # Force helix of size k starting with (i,j) to be formed
P i 0 k    [TYPE] # Prohibit nucleotides i...i+k-1 to be paired
P i j k    [TYPE] # Prohibit pairs (i,j),...,(i+k-1,j-k+1)
P i-j k-l  [TYPE] # Prohibit pairing between two ranges
C i 0 k    [TYPE] # Nucleotides i,...,i+k-1 must appear in context TYPE
C i j k           # Remove pairs conflicting with (i,j),...,(i+k-1,j-k+1)
E i 0 k e         # Add pseudo-energy e to nucleotides i...i+k-1
E i j k e         # Add pseudo-energy e to pairs (i,j),...,(i+k-1,j-k+1)
UD m e     [LOOP] # Add ligand binding to unpaired motif m with binding
                  # energy e in particular loop type(s)

                  # [LOOP]        = { E, H, I, M, A }
                  # [TYPE]        = [LOOP] + { i, m }
                  # [ORIENTATION] = { U, D }
```

Command file syntax
Hard and soft constraints, as well as ligand binding motifs can be specified in a simple command text file. Programs such as `RNAfold` then parse these commands and modify the prediction algorithms accordingly.

7. Acknowledgments

This work was partly funded by the Austrian Science Fund FWF project "RNA regulation of the transcriptome" (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] Lorenz, R., Hofacker, I.L., Stadler, P.F. (2016), RNA folding with hard and soft constraints. Algorithms for Molecular Biology 11:1, 1-13.