Advances in RNA Structure Prediction
The current state and (near) future of the ViennaRNA Package

Ronny Lorenz

University of Vienna, Theoretical Biochemistry Group (TBI)

Vienna, Austria, February 22, 2021
Modified Bases and Base Pairs
Modified Bases in RNA

Post-transcriptional RNA modifications (epitranscriptome):

- **Modomics** Database\(^1\) lists 172 different modified bases
- Commonly known modifications: \(I, \Psi, m^6A, m^1A, m^5C, \ldots\)
- Function and purpose of modifications still largely unknown
- **Structural effects of base modifications:**
  - correct folding of ncRNAs into functional structures (tRNA, rRNA, etc.)
  - regulation of protein binding sites (mRNAs, IncRNAs)
  - regulation of RNA-RNA binding sites (siRNA, miRNA)
  - Modifications may change pairing partner preference
  - Modifications may (de-)stabilize loop formation

Modifications in tRNA

- 93 known post-transcriptional modifications

- Modifications can be subtle from the RNA structure perspective
- Some are essential to induce structural domain rearrangements

---

Modifications in tRNA²

A

[Diagram of tRNA structures and modifications]

B

[Diagram of tRNA structures and modifications]

RNA Secondary Structure Prediction and Modified Bases

How to model modified bases in prediction algorithms?

Actual Requirements:
- Enhanced Nucleotide Alphabet
- Additional base pairing rules
- Corresponding energy parameters

Obstacles:
- 2D structure effects known only for a minority of modifications
- 3D effects either unknown or impossible to model
- Combinatorial explosion for energy parameters and pairing rules

Status quo:
- Some modifications prevent base pairing
- Stacking energies are available for $\Psi \bullet A$, $I \bullet U$, $I \bullet C$
- Some data available for (de-)stabilizing effects in literature
tRNA Secondary Structure Prediction

Example: human tRNA$^{\text{Phe}}$

- 17 out of 76 nucleotides are modified
tRNA Secondary Structure Prediction

Example: human tRNA$^\text{Phe}$

- 17 out of 76 nucleotides are modified
- Some modifications are known to block reverse transcriptase$^3$

tRNA Secondary Structure Prediction

Example: human tRNA$^{\text{Phe}}$

- 17 out of 76 nucleotides are modified
- Some modifications are known to block reverse transcriptase$^3$
- $\Psi \bullet A$ Nearest Neighbor stacking parameters are available$^4$

---


tRNA Secondary Structure Prediction

Example: human tRNA^{Phe}

- 17 out of 76 nucleotides are modified
- Some modifications are known to block reverse transcriptase$^3$
-Ψ•A Nearest Neighbor stacking parameters are available$^4$
- Dihydouridines (D) destabilize stacking$^5$

tRNA Secondary Structure Prediction

Pre-study on 606 sequences tRNAdb (RT-blocking modifications only)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Nucleotides</th>
<th>Performance w/o modification</th>
<th>Performance w/ modification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>total</td>
<td>cloverleaf</td>
<td>total modified RT blocking</td>
</tr>
<tr>
<td>Bacteria</td>
<td>139</td>
<td>10936</td>
<td>869</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>866</td>
<td>0.663</td>
<td>0.766</td>
</tr>
<tr>
<td></td>
<td></td>
<td>83/139</td>
<td>68/76</td>
<td></td>
</tr>
<tr>
<td>Archaea</td>
<td>76</td>
<td>5924</td>
<td>459</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>767</td>
<td>0.605</td>
<td>0.799</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44/76</td>
<td>68/76</td>
<td></td>
</tr>
<tr>
<td>Eukaryotes (nuclear)</td>
<td>242</td>
<td>18841</td>
<td>2982</td>
<td>574</td>
</tr>
<tr>
<td></td>
<td></td>
<td>604</td>
<td>0.604</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td></td>
<td>128/242</td>
<td>68/76</td>
<td></td>
</tr>
<tr>
<td>Eukaryotes (mitochondria)</td>
<td>111</td>
<td>7993</td>
<td>720</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>605</td>
<td>0.605</td>
<td>0.661</td>
</tr>
<tr>
<td></td>
<td></td>
<td>47/111</td>
<td>68/76</td>
<td></td>
</tr>
<tr>
<td>Eukaryotes (plastids)</td>
<td>38</td>
<td>2972</td>
<td>307</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>694</td>
<td>0.694</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22/38</td>
<td>72/96</td>
<td></td>
</tr>
<tr>
<td>Eukaryotes (total)</td>
<td>391</td>
<td>29806</td>
<td>4009</td>
<td>716</td>
</tr>
<tr>
<td></td>
<td></td>
<td>613</td>
<td>0.613</td>
<td>0.687</td>
</tr>
<tr>
<td></td>
<td></td>
<td>197/391</td>
<td>72/96</td>
<td></td>
</tr>
<tr>
<td>tRNAdb (total)</td>
<td>606</td>
<td>46666</td>
<td>5337</td>
<td>841</td>
</tr>
<tr>
<td></td>
<td></td>
<td>635</td>
<td>0.635</td>
<td>0.719</td>
</tr>
<tr>
<td></td>
<td></td>
<td>324/606</td>
<td>68/76</td>
<td></td>
</tr>
</tbody>
</table>

ViennaRNA’s constraints framework to the rescue!

- RT-blocking modifications $\rightarrow$ hard constraints
- $\Psi$●$A$ stacking energies $\rightarrow$ soft constraints
- Dihydouridine (D) destabilization $\rightarrow$ soft constraints
  - C3’-endo sugar conformation is destabilized in favor of C2’-endo
  - more flexibility
  - promotes destacking
  - destabilization of 1.5 kcal/mol (mono), up to 5.3 kcal/mol (oligo)

This set of constraints on average already yields much better results!
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (623 sequences)

Prediction Method

- Without unmodified bases
- All modified bases masked
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (623 sequences)

Sensitivity (TPR) vs. Positive Predictive Value (PPV)

Prediction Method:
- Gray: Without unmodified bases
- Orange: All modified bases masked
- Blue: RT blocking bases masked
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (623 sequences)

Sensitivity (TPR) vs. Positive Predictive Value (PPV)

Prediction Method
- Without unmodified bases
- All modified bases masked
- RT blocking bases masked
- RT blocking bases masked and energies for Ψ
<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Sensitivity (TPR)</th>
<th>Positive Predictive Value (PPV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without unmodified bases</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>All modified bases masked</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>RT blocking bases masked</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>RT blocking bases masked and energies for Ψ</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>RT blocking bases masked and energies for D</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

Performance on tRNAdb data set (623 sequences)
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (623 sequences)

Sensitivity (TPR) vs. Positive Predictive Value (PPV)

Prediction Method:
- Without unmodified bases
- All modified bases masked
- RT blocking bases masked
- RT blocking bases masked and energies for \( \Psi \)
- RT blocking bases masked and energies for D
- RT blocking bases masked and energies for \( \Psi \) and D
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (eucaryotes, 242 sequences)

Prediction Method
- without unmodified bases
- all modified bases masked
- RT blocking bases masked
- RT blocking bases masked and energies for \( \Psi \)
- RT blocking bases masked and energies for \( D \)
- RT blocking bases masked and energies for \( \Psi \) and \( D \)
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (bacteria, 139 sequences)

Sensitivity (TPR) vs. Positive Predictive Value (PPV) for different prediction methods:
- Without unmodified bases
- All modified bases masked
- RT blocking bases masked
- RT blocking bases masked and energies for Ψ
- RT blocking bases masked and energies for D
- RT blocking bases masked and energies for Ψ and D
Performance on tRNAdb data set (archaea, 76 sequences)

**Prediction Method**
- without unmodified bases
- all modified bases masked
- RT blocking bases masked
- RT blocking bases masked and energies for Ψ
- RT blocking bases masked and energies for D
- RT blocking bases masked and energies for Ψ and D
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (eucaryotes_mito, 111 sequences)

Prediction Method
- without unmodified bases
- all modified bases masked
- RT blocking bases masked
- RT blocking bases masked and energies for $\Psi$
- RT blocking bases masked and energies for $D$
- RT blocking bases masked and energies for $\Psi$ and $D$
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (eucaryotes_plastids, 38 sequences)

Prediction Method
- without unmodified bases
- all modified bases masked
- RT blocking bases masked
- RT blocking bases masked and energies for Ψ
- RT blocking bases masked and energies for D
- RT blocking bases masked and energies for Ψ and D

Sensitivity (TPR)
Positive Predictive Value (PPV)
tRNA Secondary Structure Prediction

Performance on tRNAdb data set (virus, 17 sequences)

- **Prediction Method**
  - without unmodified bases
  - all modified bases masked
  - RT blocking bases masked
  - RT blocking bases masked and energies for Ψ
  - RT blocking bases masked and energies for D
  - RT blocking bases masked and energies for Ψ and D

---

<table>
<thead>
<tr>
<th>Sensitivity (TPR)</th>
<th>Positive Predictive Value (PPV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>
Modified bases may heavily influence structure space

Takeaway Message:
- tRNAs require various modifications to adopt functional form
- Some can already be modeled through constraints
- Additional parameters do not necessarily increase performance
- Constraints become complex for more modifications and contexts
- Unrealistic to include full parameters with many modified bases
- No unique base annotation (tRNAdb, RNAmod, MODOMICS)

Outlook:
- Gather more data on structural effects of modified bases
- Rule and energy parameter set for pairs with modified bases
- Define fallback-rules for missing data
- Full integration of modified bases in ViennaRNA Package

1 open PostDoc Position in the RNAdeco SFB Project

---

6 Jühling et al., “tRNAdb 2009: compilation of tRNA sequences and tRNA genes.”, 2009, NAR 37, D159–D162
Modifications in tRNA

- Frequency of modifications in tRNAdb
Modifications in tRNA

- Frequency of modifications in tRNAdb
- Which modifications can be found where?
Modifications in tRNA

- Frequency of modifications in tRNAdb
- Which modifications can be found where?
- Which modifications might induce structural rearrangements?
Energy Parameters for Modified Bases

Where to get more NN parameters from?

- Typically obtained from UV-melting experiments
- More parameters to come from HRM fluorescence melting\(^9\)
- In-silico parameter estimation using Rosetta–RECESS\(^10\)

<table>
<thead>
<tr>
<th>NN</th>
<th>RECCES</th>
<th>Expt.(^11)</th>
<th>Diff.</th>
<th>NN</th>
<th>RECCES</th>
<th>Expt.(^12)</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'AI 3'UC</td>
<td>-1.16 ± 0.09</td>
<td>-1.57 ± 0.44</td>
<td>0.41</td>
<td>5'AI 3'UU</td>
<td>-0.04 ± 0.10</td>
<td>-0.41 ± 0.47</td>
<td>0.37</td>
</tr>
<tr>
<td>5'AC 3'UI</td>
<td>-0.74 ± 0.13</td>
<td>-1.02 ± 0.40</td>
<td>0.28</td>
<td>5'UU 3'AI</td>
<td>-0.80 ± 0.08</td>
<td>0.43 ± 0.44</td>
<td>1.23</td>
</tr>
<tr>
<td>5'UI 3'AC</td>
<td>-0.82 ± 0.07</td>
<td>-0.96 ± 0.40</td>
<td>0.14</td>
<td>5'UI 3'AU</td>
<td>-0.02 ± 0.11</td>
<td>0.37 ± 0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>5'UC 3'AI</td>
<td>-1.07 ± 0.09</td>
<td>-1.18 ± 0.44</td>
<td>0.11</td>
<td>5'AU 3'UI</td>
<td>-0.72 ± 0.06</td>
<td>-0.50 ± 0.44</td>
<td>0.22</td>
</tr>
<tr>
<td>5'GI 3'CC</td>
<td>-1.83 ± 0.10</td>
<td>-2.62 ± 0.40</td>
<td>0.79</td>
<td>5'GI 3'CUC</td>
<td>-1.09 ± 0.06</td>
<td>-1.34 ± 0.33</td>
<td>0.25</td>
</tr>
<tr>
<td>5'GC 3'CI</td>
<td>-1.96 ± 0.09</td>
<td>-1.89 ± 0.31</td>
<td>0.07</td>
<td>5'GI 3'CUC</td>
<td>-1.76 ± 0.11</td>
<td>-1.03 ± 0.30</td>
<td>0.73</td>
</tr>
<tr>
<td>5'CI 3'GC</td>
<td>-1.78 ± 0.16</td>
<td>-1.86 ± 0.31</td>
<td>0.08</td>
<td>5'CI 3'GU</td>
<td>-0.87 ± 0.14</td>
<td>-0.77 ± 0.39</td>
<td>0.10</td>
</tr>
<tr>
<td>5'CC 3'GI</td>
<td>-2.21 ± 0.05</td>
<td>-2.23 ± 0.40</td>
<td>0.02</td>
<td>5'CU 3'GI</td>
<td>-1.94 ± 0.13</td>
<td>-1.22 ± 0.37</td>
<td>0.72</td>
</tr>
<tr>
<td>5'II 3'CC</td>
<td>-1.03 ± 0.12</td>
<td>-</td>
<td>-</td>
<td>5'II 3'UU</td>
<td>0.07 ± 0.11</td>
<td>2.66 ± 0.88</td>
<td>2.59</td>
</tr>
<tr>
<td>5'IC 3'CI</td>
<td>-0.95 ± 0.13</td>
<td>-</td>
<td>-</td>
<td>5'II 3'UU</td>
<td>-0.09 ± 0.09</td>
<td>3.58 ± 1.09</td>
<td>3.67</td>
</tr>
<tr>
<td>5'CI 3'IC</td>
<td>-0.71 ± 0.17</td>
<td>-</td>
<td>-</td>
<td>5'UI 3'IU</td>
<td>0.52 ± 0.14</td>
<td>2.23 ± 0.91</td>
<td>1.71</td>
</tr>
</tbody>
</table>


\(^10\) Chou et al., “Blind tests of RNA nearest-neighbor energy prediction”, 2016, PNAS July 26, 113 (30) 8430-8435


\(^12\) D. J. Wright, J. L. Rice, D. M. Yanker, and B. M. Znosko, *Biochemistry*, vol. 46, no. 15, 2007
RNA-RNA Interactions
RNA-RNA interactions

ViennaRNA Package 2.5.0alpha2 contains RNAmultifold\textsuperscript{13}

- Interaction of $N$ RNAs with $n = n_1 + n_2 + \ldots + n_N$
- Single or all permutations of a given complex
- All connected complexes up to $N$ constituents
- Implements MFE, partition function, equilibrium concentrations
- $O(n^3)$ base pair probabilities ($O(n^2 N)$ overhead)

RNA-RNA interactions

Example: Splicosomal snRNA complex formation
- simplified model: no mRNA, proteins, modifications, etc.
- subsequent increase in concentration of U6, U4, U5 and U2

<table>
<thead>
<tr>
<th>U6</th>
<th>U4</th>
<th>U5</th>
<th>U2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0337</td>
<td>4.4931</td>
<td>2.3147</td>
<td>0.0014</td>
</tr>
<tr>
<td>11.9961</td>
<td>6.9295</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Importance of binary interactions: $\Delta G_{A|B} = RT \ln Q - RT \ln Q_{A|B} \geq 0$
RNA-RNA interactions

Conclusion, Outlook, and Takeaway Message:

- **RNAmultifold** available in ViennaRNA Package 2.5.0a2
- Very fast NUPACK alternative
- Same model and parameters as for single sequences
- Merge process into mainline ViennaRNA in progress
- Suboptimal structure prediction still requires attention
- Re-use of DP matrices for different permutations in the future
- New benchmark against NUPACK 4 required\(^\text{14}\)

RNA Structure Probing, Pseudoknots, and Structure Motifs
PaRNAssus: Deciphering Complex RNA Structure by Probing and Predictions

- Joint project between FWF (Austria) and ANR (France)
- Exp. probing at different conditions (ions, temperature, agents)
- Separate/Deconvolute (differential) probing signal
- Detection of higher-order structure motifs from probing signals
- Novel heuristics for PK and non-canonical structure prediction
- Implementation of selected already available PK grammars
- Refactoring of RNAPKplex for constraints support almost done
- Connect probing data and folding kinetics simulations
PaRNAssus: Deciphering Complex RNA Structure by Probing and Predictions

- Joint project between FWF (Austria) and ANR (France)
- Exp. probing at different conditions (ions, temperature, agents)
- Separate/Deconvolute (differential) probing signal
- Detection of higher-order structure motifs from probing signals
- Novel heuristics for PK and non-canonical structure prediction
- Implementation of selected already available PK grammars
- Refactoring of RNAPKplex for constraints support almost done
- Connect probing data and folding kinetics simulations

2 open PhD Positions in Structural RNA Bioinformatics
Acknowledgements

- Christoph Flamm
- Ivo L. Hofacker
- Yann Ponty
- Bruno Sargueil
- Thomas Spicher
- Peter F. Stadler
- Yuliia Varenyk

Thank You for your attention!

2 open PhD Positions in Structural RNA Bioinformatics
1 open PostDoc Position in the RNAdeco SFB Project