COMP598: Advanced Computational Biology Methods and Research

Modeling RNA 3D structure

Jérôme Waldispühl
School of Computer Science, McGill
Slides from Neocles Leontis & Jes Frellsen
Motivations and challenges

• Secondary structure is a simplification of the three-dimensional structure.

• Function is achieved through the 3D structure.

• Experimental determination the RNA 3D structure is hard.

• Modeling the 3D structure is also hard!

• Before the prediction, a work has to be done on modeling and alignment of 3D structure.
Beyond the secondary structure

The hierarchy of the model is not as obvious as expected:

- Is the secondary structure with/without pseudo-knot unique?
- Is there other type of interacting motifs? (for instance base triple)
Beyond the secondary structure

The type of interactions is not restricted to Watson-Crick base pairs:

SECIS element

5S ribosomal RNA

Non-canonical interactions

NB: non-stacking interaction
Classification of non Watson-Crick base pair interactions

What are we seeing when looking at the 3D structure?

“Loops” are not loops!

 Sites for non Watson-Crick base pairs.
Classification of non Watson-Crick base pair interactions

Modeling the nucleotide side-chain with interacting edges
Classification of non Watson-Crick base pair interactions

Consequence: 3 edges available for base-pairing.
Classification of non Watson-Crick base pair interactions

Orientation of edge interaction is also important: The glycosidic bond orientation.

Cys (default):

Trans:
Classification of non Watson-Crick base pair interactions

12 edge-to-edge interacting motifs
Classification of interactions

But the puzzle is still far to be completed! 😞

Base interacting with all 3 edges
Classification of interactions

The interacting motif is extended to model base triple.
More Features…

Base-Sugar conformation.

Anti (default):

Syn (Purines only):
More features…

Local strand orientation:

- Anti-parallel (default)
- Parallel

Locally parallel strands:
New symbols

Indicates Base Stacking

Indicates Change in Strand Orientation

Indicates syn conformation for base
Example: 5S motif

H. Marismortui
Euryarchaeota
Example: 5S motif

Loop A

Loop B

Loop C

Loop D

Loop E

E. coli
Bacteria
More features (2)...

Superposition of tetra and penta GNRA loops:

Interaction of GNRA loops are also conserved:

23S *H. marismortui* 23S *T. thermophilus*
Q: Given a description of a “known” motif, how to identify this motif in target structures?

Use graph theory, the problem of identifying a known pattern in a target graph reduces to the following:

1. Searching for isomorphic occurrences of the pattern (subgraph isomorphism).

2. Finding similar occurrences of the pattern (identifying a maximum common subgraph).

But it’s NP-complete…
FR3D: Find RNA 3D

(Sarver et al., 2008)

Leontis + Zirbel groups

Find small RNA motifs (two to 20 nucleotides) in PDB files.
FR3D example: C-loop search

Output:

<table>
<thead>
<tr>
<th>Filename (PDB)</th>
<th>Discrepancy from query</th>
<th>Motif Nucleotides</th>
<th>Pairwise Interactions</th>
<th>Structural Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2AW4</td>
<td>0.000</td>
<td>U 2680 C 2681 C 2683 U 2684 A 2725 A 2727 s35 cWW tWH s35 cWS cWW s35</td>
<td>UCA-CU...AA-A</td>
<td></td>
</tr>
<tr>
<td>1s72</td>
<td>0.127</td>
<td>C 2717 C 2718 C 2720 U 2721 A 2761 G 2763 s35 cWW tWH s35 cWS cWW s35</td>
<td>CCA-CU...AC-G</td>
<td></td>
</tr>
<tr>
<td>1kog</td>
<td>0.136</td>
<td>C 96 C 97 C 99 U 100 A 74 G 76 s35 cWW tWH s35 cWS cWW s35</td>
<td>CCA-CU...AU-G</td>
<td></td>
</tr>
<tr>
<td>2p01</td>
<td>0.229</td>
<td>G 1319 C 1320 A 1322 U 1323 A 1331 C 1333 s35 cWW tWH s35 ncWS ncWW s35</td>
<td>GCA-AU...AG-C</td>
<td></td>
</tr>
<tr>
<td>2AW4</td>
<td>0.232</td>
<td>C 1319 C 1320 A 1322 C 1323 G 1331 G 1333 s35 cWW tWH s35 ncWS cWW s35</td>
<td>CCA-AC...GG-G</td>
<td></td>
</tr>
<tr>
<td>2AW4</td>
<td>0.244</td>
<td>G 864 C 865 C 867 U 868 A 909 C 912 s35 cWW tWH s35 ncWS cWW s35</td>
<td>GCA-CU...AAAC</td>
<td></td>
</tr>
<tr>
<td>1s72</td>
<td>0.256</td>
<td>G 1425 C 1426 C 1428 U 1429 A 1437 C 1439 s35 cWW tWH s35 cWS cWW s35</td>
<td>GCA-CU...AG-C</td>
<td></td>
</tr>
<tr>
<td>2p01</td>
<td>0.278</td>
<td>G 864 C 865 C 867 U 868 A 909 C 912 s35 cWW tWH s35 ncWS cWW s35</td>
<td>GCA-CU...AAAC</td>
<td></td>
</tr>
<tr>
<td>1s72</td>
<td>0.380</td>
<td>G 371 C 372 A 374 U 375 A 389 C 390 s35 cWW tWH s35 cWS cWW s35</td>
<td>GCA-AU...A-C</td>
<td></td>
</tr>
<tr>
<td>1s72</td>
<td>0.402</td>
<td>G 958 C 959 C 962 C 963 A 1005 C 1008</td>
<td>GCGACC...AAAC</td>
<td></td>
</tr>
<tr>
<td>2AVY</td>
<td>0.415</td>
<td>A 371 C 372 A 374 U 375 A 389 U 390</td>
<td>ACA-AU...A-U</td>
<td></td>
</tr>
</tbody>
</table>
What do we learn?

- Positions of insertions/deletions
- Base-pair co-variations
- Base conservations
- Problem: Limited number of examples
Q: Given a structure, how to identify “unknown” motifs within it?

1. Identify all secondary structure elements of the RNA tertiary structure;
   
   Rationale: motifs as “often embedded within regular helical regions forming internal loops, but may also comprise hairpin or junction loops.”

2. Calculate a similarity measure for each pair of structural elements;
   
   Rationale: Computing the largest extensible common noncanonical subgraph.

3. Cluster the structural elements according to the similarity measure.
Two structural elements containing 16S K-turn motifs.

Djelloul M, Denise A RNA 2008;14:2489-2497
Dendrogram of hierarchical clustering of H.m 23S RNA produced with hclust.

Djelloul M, Denise A RNA 2008;14:2489-2497
Recurrent motifs found in ribosomal structures.

Djelloul M, Denise A RNA 2008;14:2489-2497
Crystal structures of four putative new motifs superimposed.

Djelloul M, Denise A RNA 2008;14:2489-2497
## Predicting RNA 3D structures

<table>
<thead>
<tr>
<th>Program</th>
<th>Input</th>
<th>Model</th>
<th>Simulation method</th>
<th>Description / Webpage</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Automatic prediction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iFoldRNA</td>
<td>Sequence</td>
<td>Coarse-grained three bead model</td>
<td>Replica exchange, molecular dynamics</td>
<td>Uses discrete molecular dynamics and force fields to simulate RNA folding dynamics. <a href="http://troll.med.unc.edu/ifoldrna/">http://troll.med.unc.edu/ifoldrna/</a></td>
<td>[132, 133]</td>
</tr>
<tr>
<td>FARNA (Rosetta)</td>
<td>Sequence, secondary structure</td>
<td>Coarse-grained one bead model</td>
<td>Fragment assembly, Monte Carlo</td>
<td>Uses 3-nt. fragment library, Monte Carlo simulations and a potential function to predict the structure. <a href="http://www.rosettacommons.org/manuals/archive/rosetta3.0_user_guide/index.html">http://www.rosettacommons.org/manuals/archive/rosetta3.0_user_guide/index.html</a></td>
<td>[125, 127]</td>
</tr>
<tr>
<td>NAST</td>
<td>Secondary structure, tertiary contacts</td>
<td>Coarse-grained one bead model</td>
<td>Molecular dynamics</td>
<td>Performs molecular dynamics simulations guided by a knowledge-based statistical potential function <a href="https://simtk.org/home/nast">https://simtk.org/home/nast</a></td>
<td>[131]</td>
</tr>
<tr>
<td><strong>Interactive manipulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Modeling and predicting RNA 3D structure: MC-Fold | MC-Sym pipeline
(F. Major group, UdM)

Cycle decomposition of the 3D structure using the Leontis-Westhof nomenclature.

<table>
<thead>
<tr>
<th>#</th>
<th>Class</th>
<th>Base pairs</th>
<th>LSU index</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>LS-P-LS-P</td>
<td>(W/W,W/W)</td>
<td>02562, 02583, 02570, 02571</td>
<td>Watson-Crick tandem</td>
</tr>
<tr>
<td>(2)</td>
<td>LS-P-LS-P</td>
<td>(H/S)</td>
<td>02696, 02697, 02698, 02699</td>
<td>GNRA loop</td>
</tr>
<tr>
<td>(3)</td>
<td>LS-P-LS-P</td>
<td>(H/S,H/H)</td>
<td>01532, 01533, 01658, 01659</td>
<td>Non Watson-Crick tandem</td>
</tr>
<tr>
<td>(4)</td>
<td>LS-P-LS-P</td>
<td>(H/H,W/H)</td>
<td>0977, 0979, 09103, 09104</td>
<td>Non Watson-Crick tandem</td>
</tr>
<tr>
<td>(5)</td>
<td>LP-LS-P-S</td>
<td>(S/H,H/S)</td>
<td>01971, 01972, 01973, 02009</td>
<td>Non Watson-Crick tandem</td>
</tr>
<tr>
<td>(6)</td>
<td>LS-P-L-S</td>
<td>(H/S)</td>
<td>01097, 01098, 01258, 01259</td>
<td>GNRA interior loop</td>
</tr>
<tr>
<td>(7)</td>
<td>LS-P-L-S</td>
<td>(H/S)</td>
<td>01392, 01393, 01394, 01395</td>
<td>Double-stacked bulge</td>
</tr>
<tr>
<td>(8)</td>
<td>LS-P-S-P</td>
<td>(W/H,W/W)</td>
<td>02118, 02276, 02277, 02470</td>
<td>Non Watson-Crick tandem</td>
</tr>
<tr>
<td>(9)</td>
<td>P-S-P-LS</td>
<td>(W/H,H/S)</td>
<td>00481, 00485, 00486, 00482</td>
<td>Non Watson-Crick tandem</td>
</tr>
<tr>
<td>(10)</td>
<td>LS-P-P-P</td>
<td>(S/H,W/S,W/W)</td>
<td>01231, 02498, 02522, 02523</td>
<td>Base triple</td>
</tr>
</tbody>
</table>
MC-Fold workflow
MC-Sym workflow

> SRL
GGGUGCUCAGUACGAGAGGAAACGCCACCC
((((((.((((..)))))))))))))

NCM Fusion
Beyond conserved 3D motifs

The 3D structure can be modeled by enumeration of the degree of freedom of the polynucleotide.

Each nucleotide in an RNA molecule can be represented by the base type and 7 dihedrals angles.
A continuous probabilistic model of local RNA 3D structure (Jes Frellsen et al.)

Modeling and estimating the angle distributions.

- Each variable is multi-modal
  - Can be described by a mixture of simple distributions
  - Von Mises distribution
- The angles co-vary both within nucleotides and between consecutive nucleotides
  - We model this by a sequential model
- Each variable lies on a circle
  - Requires directional statistics
A continuous probabilistic model of local RNA 3D structure

• An DBN with 3 random variables per angle:
  • Discrete input variable indicating angle type (7 states)
  • Hidden variable with 20 states
  • Output variable representation the angle value and the CPDs given the hidden state is modelled by Von Mises distributions

• Structure of an IOHMM with continuous output (except bookkeeping)
• Does not impose a groping of the angles
• Parameters are estimated by stochastic EM from experimental data
A continuous probabilistic model of local RNA 3D structure

- The model captures the distribution of the individual angles
- The model captures the pairwise dependencies between the angles
A continuous probabilistic model of local RNA 3D structure

Generation of decoy with a simple simulated annealing scheme:

1. Sample a whole structure, $S$, without clashes
2. Make new structure, $S'$, by resampling four consecutive angles in $S$ (randomly picked)
3. Evaluate $S'$
   a. If it has clashed it is rejected
   b. If it has a better energy than $S$ then $S'$ is set to be the new $S$
   c. If it has a worse energy then with probability, $p$, $S'$ is set to be the new $S$
      (otherwise it is rejected)
   d. Go to step 2

In the scheme we used
- $p = e^{(E-E')/T}$, where $T$ decreases with time
- a simple "energy function" that promotes structure with the same Watson-Crick base pair as are found in the target structure
A continuous probabilistic model of local RNA 3D structure: Results

<table>
<thead>
<tr>
<th>Target Structure</th>
<th>Length (Bases)</th>
<th>Decoys &lt; 4Å</th>
<th>Decoys &lt; 3Å</th>
<th>Lowest RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ZIH</td>
<td>12</td>
<td>58.8%</td>
<td>21.3%</td>
<td>1.55Å</td>
</tr>
<tr>
<td>1RNG</td>
<td>12</td>
<td>55.1%</td>
<td>3.5%</td>
<td>2.48Å</td>
</tr>
<tr>
<td>1XWP</td>
<td>13</td>
<td>28.3%</td>
<td>5.8%</td>
<td>2.03Å</td>
</tr>
<tr>
<td>1I4B</td>
<td>13</td>
<td>34.6%</td>
<td>0.1%</td>
<td>2.91Å</td>
</tr>
<tr>
<td>1PJY</td>
<td>22</td>
<td>10.0%</td>
<td>1.9%</td>
<td>1.89Å</td>
</tr>
</tbody>
</table>

Results computed from 1500 decoys

![1ZIH Target structure](image1.png)

![Best decoy](image2.png)