RNA secondary structure

- Functions
- Representations
- Predictions

Many slides courtesy of M. Zuker, RPI Math
When RNA secondary structure matters

mRNA --> protein

Strong secondary structure can block translation.
especially sensitive is the...

**RBS**

ribosome binding site

species specific!

Registry of Standard Biological Parts
http://parts.igem.org/
microRNA (miR)

- Found in 3'UTR, introns, exons
- primary-microRNA
  - transcription, folding
  - dicer
- microRNA duplex
  - argonaute proteins
  - passenger strand degraded
- microRNA
  - blocks translation
  - deadenylation
  - endonuclease digestion

- nucleus
- cytoplasm
Ambiguous bases

How to remember the ambiguous codes? The NOT cases use the next (available) letter.

<table>
<thead>
<tr>
<th>IUPAC nucleotide code</th>
<th>Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>T (or U)</td>
<td>Thymine (or Uracil)</td>
</tr>
<tr>
<td>R</td>
<td>A or G</td>
</tr>
<tr>
<td>Y</td>
<td>G or C</td>
</tr>
<tr>
<td>S</td>
<td>G or C</td>
</tr>
<tr>
<td>W</td>
<td>A or T</td>
</tr>
<tr>
<td>K</td>
<td>G or T</td>
</tr>
<tr>
<td>M</td>
<td>A or C</td>
</tr>
<tr>
<td>B</td>
<td>C or G or T</td>
</tr>
<tr>
<td>D</td>
<td>A or G or T</td>
</tr>
<tr>
<td>H</td>
<td>A or C or T</td>
</tr>
<tr>
<td>V</td>
<td>A or C or G</td>
</tr>
<tr>
<td>N</td>
<td>any base</td>
</tr>
<tr>
<td>. or -</td>
<td>gap</td>
</tr>
</tbody>
</table>

NH₃ (amino) opposite to the glycosidic bond.

C=O (keto) opposite to the glycosidic bond.
RNA secondary structure is base pairing \( i \cdot j \)

RNA secondary structure is the collection (set) of base pairs that form in 3D. The hydrogen bonds of base pairs and the stacking of adjacent base pairs are responsible for most of the thermodynamic stability of an RNA. The most common base pairs are Watson-Crick (W-C): C·G, G·C, A·U and U·A. Base pairs between G and U, G·U and U·G, are called wobble pairs. Other pairs are called non-canonical.

A base pair between \( r_i \) and \( r_j \) is denoted as \( r_i \cdot r_j \) (\( i < j \)) or simply by \( i \cdot j \) when the context is clear. A secondary structure is a collection, \( S \), of base pairs that satisfy:

1. If \( i \cdot j \in S \), then \( j - i > 3 \). The number 3 is called the minimum hairpin loop size. This “rule” is broken by the existence of tetra-loops.

2. If \( i \cdot j, i' \cdot j' \in S \), then \( j = j' \) if \( i = i' \) and \( i = i' \) if \( j = j' \). This rule excludes base triples, and is violated in some structures.

3. If \( i \cdot j, i' \cdot j' \in S \), then either \( i < j < i' < j' \) (\( i \cdot j \) precedes \( i' \cdot j' \)) or \( i < i' < j' < j \) (\( i \cdot j \) includes \( i' \cdot j' \)). Violations of this rule also occur and create “pseudoknots”.
Not just Watson-Crick..

Recent work of Leontis and Westhof describe 12 distinct ways in which any two bases can pair.

- Some cannot occur.
- Some can be modeled but have not yet been observed in structures derived by X-ray crystallography or NMR methods.
- Many base pairs are structurally equivalent (isosteric).


Different ways of base-pairing allow RNA to adopt duplex structures beyond A and B helix.

If any base-pair is possible, how do we predict pairings?
Representations of RNA secondary structures.

Structure types within RNA sec struct

- E: exterior loop
- H: hairpin loop
- B: bulge loop
- I: interior loop
- M: multi-branch loop

2D plot
Circle Plot & Tree Diagram

Two plots in one.

- Bases are drawn along the circumference of a circle.
- Base pairs are circular arcs that intersect the circle at right angles.
- Black lines (edges) within the circle comprise a “tree representation” of the secondary structure. Every base pair and multibranched loop is a node. Nodes connecting consecutive base pairs can be collapsed into a single “helix” node.

Colors depend on the probability of base pairs, as in the standard plot.

note: No bp lines cross.
Converting from 2D plot sec struct to circle/tree.
Two base pairs (BPs), $r_i \cdot r_j$ and $r_{i'} \cdot r_{j'}$, can be called “incompatible” if $i < i' < j < j'$. That is, they violate the third rule for RNA secondary structure.

A pseudoknot is created by two incompatible helices (stems). That is, every BP in one is incompatible with every BP in the other.

```
A-C
3' - A-G-G-C-U / U
 U-C-C-G-A-G-G-G
 U  C-C-C - 5'
 C--U--C/
```

<= Example of a simple pseudoknot.

Circle plot=====>
3D model of simple pseudoknot. Coordinates by F. Major
Can you find the pseudoknot?

Tetrahymena thermophila LSU rRNA
GenBank# J01235
Eucarya, Protoctista, Ciliophora (IC1)
Jun 09, 1994
Now you can. P3 and P7 create a pseudoknot.
Now you can see them in the original plot.
Prediction of RNA secondary structure

• Comparative methods, phylogenetics
  - Considered the “gold standard”.
  - Labor intensive
  - Requires numerous homologous sequences that can be well aligned.

• Free energy calculations
  - Works on single sequences.
  - Fast, cheap and easy to perform.
  - Unreliable in general.
  - Cannot predict pseudoknots. (Some time consuming exceptions exist.)

• Dot plot
  - Easy.
  - Can be done on a single sequence.
  - Cannot find non-canonical base pairs.

• Mutual information (MI)
  - Requires a deep multiple sequence alignment
  - Can find non-canonical base-pairs.
Comparative modeling assume conserved structure between homologs

- A “golden rule” in biology: Structure is conserved more than sequence.
- This principle can be used to predict RNA secondary structure.

SSU rRNA: *Escherichia coli* versus *Deferrribacter thermophilus*
RNA structure by energy minimization

Assumes energy is the sum of base pairs.

$$\Delta G(S) = \sum_{r_i, r_j \in S} e(i, j).$$

where, $e(i, j) = 0$ if $j - i < 4$

Forward summation of energy matrix $E$:

$$E(i, j) = \min \left\{ E(i+1, j), E(i, j-1), e(i, j) + E(i+1, j-1), \min_{k=i+1}^{j-1} (E(i, k) + E(k+1, j)) \right\}$$

- Add to loop
- Start a helix, or add a base pair
- Join helices in multi-loop
- Too close, can't pair
1. energy matrix is initialized starting from diagonal:

\[ E(i, j) = \min \left\{ E(i+1, j), E(i, j-1), e(i, j) + E(i+1, j-1), \min_{k=i+1}^{j-1} (E(i, k) + E(k+1, j)) \right\} \]

2. base-pairing is found by tracing back from (1, n)

Stack A becomes the answer: a list of base pairs
Stack B is a list of unfinished segments

Find k such that \( E_{i,j} = E_{i,k} + E_{k+1,j} \)
Push \((i,k)\) and \((k+1,j)\) onto Stack B.
Stop with error if no such k exists.
RNA structure by Dot Plot

The most abstract representation of secondary structure.

- Bases are not drawn.
- Base pairs are dots (filled in circles, squares, diamonds or other shapes).
  Row is 5' base & Column is 3' base. A dot in row i and column j represents the base pair \( r_i : r_j \)
  in the RNA whose sequence is \( r_1 r_2 \ldots r_n \).
- Helices and hairpin loops are easy to detect.
- Bulge and interior loops are a bit harder to detect.
- Multi-branch loop detection is not easy.
RNA structure by Dot Plot
1) Run BLAST search to get homologs.
2) Prune sequences to remove redundancy.*
3) Prune columns to remove uninformative data. (conserved positions tell you nothing)
4) Calculate \textit{mutual information} \( (M_{i,j}) \) for all pairs of positions \((i,j)\).
**Mutual information, in general**

A measure of the surprisingness of a pair of events.

\[
M = \sum_{a,b \in \{\text{events}\}} f(a,b) \log_2 \left( \frac{f(a,b)}{f(a)f(b)} \right)
\]

frequency of a pair of events observed together = \(N(a,b \text{ events})/N(\text{total events})\)

in bits, because we used \(\log_2\)

sum over all event pair types

expected frequency of a,b events together is the product of the frequencies of the events separately
Mutual information for base-pairs

A measure of the surprisingness of the evolution of two positions in the sequence.

\[
M(i,j) = \sum \frac{f_{i,j}(B_1,B_2)}{f_i(B_1)f_j(B_2)} \log_2 \left( \frac{f_{i,j}(B_1,B_2)}{f_i(B_1)f_j(B_2)} \right)
\]

\(B_1, B_2 \in \{A,C,G,T\}\)

sum over all base-pair types

expected frequency of base \(B_1, B_2\)

Exercise: Calculate \(M(i,j) = \square\)
Can you find pairs of positions with high MI?

W-C

non-canonical
Mutual information matrix for 20 aligned sequences

very noisy.

Take home message: You need lots of sequence to do mutual information analysis. And this is for RNA where the signal is strong. Try protein. You'll need thousands of sequences....
Same RNA, $M_{i,j}$ for 302 aligned sequences

Wow! cleaner
Helices now appear as straight lines of dots, after re-numbering to remove gaps in the MSA.
Prediction of RNA secondary structure

• Comparative methods, phylogenetics
  – Considered the “gold standard”.
  – Labor intensive
  – Requires numerous homologous sequences that can be well aligned.

• Free energy calculations
  – Works on single sequences.
  – Fast, cheap and easy to perform.
  – Unreliable in general.
  – Cannot predict pseudoknots. (Some time consuming exceptions exist.)

• Dot plot
  - Easy.
  - Can be done on a single sequence.
  - Cannot find non-canonical base pairs.

• Mutual information (MI)
  - Requires a deep multiple sequence alignment
  - Can find non-canonical base-pairs.
1. What are the IUPAC codes?
2. What are the different types of RNA structure?
3. What is mutual information? How is it calculated?
4. What are the sources of energy for RNA structure?
5. What algorithm is used to calculate the energy over all RNA structures?
6. What is expressed in a dot plot?
7. What is expressed in a circle plot?
8. What is a pseudoknot?
9. What is a non-canonical basepair?
10. Can you convert a dotplot into a graph?
11. Can you convert a graph into a circle plot?
12. Can you see a pseudoknot in a graph, circle, dotplot?