Comparative Analysis of RNA structure
Stages of a comparative analysis

- Initial definition of the basic secondary structure
  - A few sequences
  - May use thermodynamic prediction or probing data
  - Resolution: helices

- Refinement of 2° structure & identification of tertiary contacts
  - >50 sequences
  - Identify base-base interactions and higher-order covariations
  - Resolution: base-pairs

- Tertiary modeling
  - Local structures
  - Assembled global structure
  - Resolution: angles and distances
The problem: For every short string of bases (sequence), which complementary sequence (if any) is the correct pairing partner?

For every 4 base sequence, there will be a perfect complement every 50 bases on average!
Initial definition of 2° structure

Dot plot of all 5bp complementarities in the *B. subtilis* RNase P RNA
Initial definition of 2° structure

The solution: Compare more than one homologous sequence for variation consistent (or inconsistent) with each possible pairing.

This is the same concept as a genetic analysis of mutation and second-site intragenic reversion.
Initial definition of 2° structure

Comparative evidence for secondary structure is based on covariation in sequence alignments

P4 and P6 in bacterial RNase P RNAs

<table>
<thead>
<tr>
<th></th>
<th>82–85</th>
<th>276–279</th>
<th>66–74</th>
<th>353–360</th>
</tr>
</thead>
<tbody>
<tr>
<td>eubacterial consensus</td>
<td>GGGC....GCC..</td>
<td>AAGuCCGGG...CCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermotoga maritima</td>
<td>GAAGC....GCUC..</td>
<td>AAGuCCGGA...UCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deinococcus radiodurans</td>
<td>GGGC....GCC..</td>
<td>AAGuCCGGG...CCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptomyces bikiniensis</td>
<td>GAAGC....GCUC..</td>
<td>ACGuCCGGG...CCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhodospirillum rubrum</td>
<td>GAAC....GUUU..</td>
<td>AAGuCCGGG...CCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agrobacterium tumefaciens</td>
<td>AAAU....GUUU..</td>
<td>AAGuCCGGG...CCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcaligenes eutrophus</td>
<td>GGGC....GCC..</td>
<td>AAGuCCGGA...UCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thiobacillus ferrooxidans</td>
<td>GGGC....GCC..</td>
<td>AAGuCCGGG...CCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g/d protobacteria</td>
<td>GGGC....GCC..</td>
<td>AAGuCCGGG...CCC...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacillus brevis</td>
<td>GGGC....GCC..</td>
<td>AAGuCCAGG...CCU...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Bacillus spp.</td>
<td>GGGC....GCC..</td>
<td>AAGuCCCAUG...CAUGG...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Initial definition of 2° structure

CSA of secondary structure is based on the observation that secondary (and higher-order) structure is conserved despite sequence variation

P4 and P6 in bacterial RNase P RNA

\[
\begin{array}{cccccccccc}
\text{A} & \text{U} & \text{A} & \text{U} & \text{G} & \text{C} & \text{G} & \text{C} & \text{A} & \text{U} \\
\text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} \\
\text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} \\
\text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} \\
\text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{C} \\
\end{array}
\]
Comparative analysis is an iterative process: newly-identified structure allows the alignment to be refined so that new structure can be identified.

Increasingly disparate sequences are added when it becomes possible, or when the number of sequences allows, subsets can be analyzed separately to examine structure unique to them.
Initial definition of 2° structure

Initial secondary structure of the *E. coli* RNase P RNA
As the number of sequences increases, the basic secondary structure can be refined to high resolution using statistical methods to identify and quantitate sequence covariation.

\[ H = -\sum_b f_b \ln f_b \]

- \( H \) is a measure of the variability of a sequence position \( b \) in set (A G C U) or (A=A A•G A•C A=U G•A G•G G=C ... U•U)
- \( f_b \) = frequency of each base or base pair

\[ M(x,y) = H(x) + H(y) - H(x,y) \]
- \( x \) and \( y \) are pairs of positions in an alignment
- \( M(x,y) \leq \min[H(x) H(y)] \)
- \( X^2 = 2(M \times n) \)

\( M(x,y) \) is maximized when both positions are highly variable and also perfectly correlated.
Refinement of 2° & ID of tertiary contacts

Example $M(x,y)$ analysis of a basepair

- Position 234 of bacterial RNase P against all other positions in the alignment

Position 234 of bacterial RNase P RNA basepairs with position 247
Refinement of 2° & ID of tertiary contacts

Example $M(x,y)$ analysis of a helix

C  G
G  A

312  A – U  317
311  G – U  318
310  U – G  319
309  G – C  320
308  A – U  321
307  C – G  322
306  C – G  323
305  G – C  324

P18 in E. coli
RNase P RNA

312
311
310
309
308
307
306
305

317
318
319
320
321
322
323
324
Refinement of 2° & ID of tertiary contacts

2-dimensional ‘Dotplot’ of $M(x,y)$ of bacterial RNase P RNA
Refinement of $2^\circ$ & ID of tertiary contacts

Base triples

These are identified in the same way as secondary basepairs, except that all three bases covary with each other.
Refinement of 2° & ID of tertiary contacts

Base triples

W/C basepair

bases 94 & 104

G 0.0519 0.0000 0.0000 0.0000 0.0519 0.0000
A 0.0825 0.0000 0.0000 0.0825 0.0000 0.0000
U 0.2784 0.0000 0.2784 0.0000 0.0000 0.0000
C 0.5773 0.5773 0.0000 0.0000 0.0000 0.0000
(gap) 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

bases 94 & 316

G 0.3505 0.0515 0.2577 0.0412 0.0000 0.0000
A 0.5082 0.4845 0.0206 0.0412 0.0000 0.0000
U 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
C 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
(gap) 0.0412 0.0412 0.0000 0.0000 0.0000 0.0000

bases 104 & 316

G 0.3505 0.0000 0.0412 0.2577 0.0515 0.0000
A 0.6082 0.0619 0.0412 0.0206 0.4845 0.0000
U 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
C 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
(gap) 0.0412 0.0412 0.0000 0.0000 0.0000 0.0000

G•A or A•G

U•G or A•C
Refinement of 2° & ID of tertiary contacts

A high-resolution CSA secondary structure is comparable in accuracy to a clean primary sequence or a high-resolution X-ray diffraction tertiary structure.

- Basepairs can be tested individually, with $p$-values attached
- Highly variable regions can be studied in small, closely-related subgroups
- Only the most conserved (invariant) sequences cannot be examined
- The possibility of remaining helices can be excluded
Refinement of 2° & ID of tertiary contacts

Bacterial RNase P RNA

- All but 3 basepairs (consisting of invariant bases) in this structure are supported by sequence covariation
- Additional secondary structure cannot be accommodated
Refinement of 2° & ID of tertiary contacts

Comparative analysis is the gold standard for determination of RNA secondary structures, but...

**Strengths:**
- objective, quantitative
- automatable & visualizable
- basepair resolution
- can distinguish thermodynamically equivalent possibilities
- only biologically-relevant structures identified
- identifies any base-base interaction with alternative versions

**Weaknesses:**
- phylogenetic affects add complexity
- seq sample affects
- alignment basically a manual process
- Mxy best for final stages of secondary analysis & analysis of tertiaries
- no specific information from invariant sequences
- no specific information from idiosyncratic sequences
- difficult to incorporate biochemical data
Initial definition of 2° structure

Comparative sequence analysis is the undisputed gold standard for the determination of RNA secondary structure

- Some RNA secondary structures determined by CSA:
  - tRNA
  - ssu rRNA
  - lsu rRNA
  - 5S rRNA
  - 5.8S rRNA
  - RNase P RNA
  - SRP RNA
  - Group I introns
  - Group II introns
  - splicing snRNAs
  - guide snoRNAs
  - 6S RNA
  - tmRNA
  - RNase MRP RNA
  - telomerase RNA
  - hammerhead ribozyme
  - hairpin ribozyme
  - delta virus ribozymes
  - T-box RNA
  - Riboswitches
  - Txn terminators
  - Attenuators
  - Rep origin regulators
  - antisense RNAs
  - Editing guide RNAs
  - SELEX aptamers!
  - &c, &c ...

Tertiary modeling

Independent substructures
  - hairpin elements
  - large insertions/deletions

Helical stacks
  - covariation in length
  - steric constraint/adjacency

Secondary structure motifs
  - replaceable alternatives
  - structure classes

Phylogenetic minimum core
  - essential vs stabilization vs trivial
  - inside vs outside

Assembly of global models
Tertiary modeling

Independent substructures

These usually seem to stack onto a conserved helix for stabilization
Tertiary modeling

Helical stacking

Helices in pseudoknots usually stack
Tertiary modeling

Independent substructures

These also usually seem to play a role in stabilization
Tertiary modeling

Helical stacking

Stacking is a major force in directing the folding of RNA
Tertiary modeling

Another GNRA:receptor interaction

GAAA tetraloop

Also notice the lone basepair
Tertiary modeling

Local structure models
Tertiary modeling

Base triples

This is a common GNRA:minor groove tertiary interaction
Tertiary modeling

Phylogenetic minimal core structure

The phylogenetic minimum core contains all of the essential sequence and structure. Other elements generally contribute (importantly or not) to stability, or are trivial elements that are tolerated as long as they do not interfere with the rest of the molecule.
Tertiary modeling

Conserved inside, variable outside

Variable sequences and helices of conserved length are concentrated in the heart of the RNA. Variable sequences and structures are peripheral.
Tertiary modeling

But just because you know the structure of all the parts doesn’t mean you know the structure of the whole!
Tertiary modeling

The need for constraints

• The number of possible foldings of a polymer of any length are astronomical

• Constraints are bits of information that limit the possibilities, i.e. they constrain the model.

• Some constraints:
  • The secondary structure!
  • Phylogenetic variation
  • Tertiary interactions
  • Helical stacking data
  • Distance measurements & crosslinks
  • Known overall shape
  • etc, etc....
Tertiary modeling

Interactive computer modeling : Massire & Westhof

Constraint #1 : the detailed secondary structures of type A and B RNase P RNAs
Tertiary modeling

Interactive computer modeling: Massire & Westhof

Constraints #2 & 3: All known tertiary interactions & stacking partners (including some suggested by the preliminary models, and some reasonable guesses)
Tertiary modeling

Interactive computer modeling: Massire & Westhof

Pieced all together to satisfy crosslinking data, biochemical data, and aesthetics, and...Viola!
Tertiary modeling

Models vs crystal structure

Comparative model

Crystal structure
Stages of a comparative analysis

- Initial definition of the basic secondary structure
  - A few sequences
  - May use thermodynamic prediction or probing data
  - Resolution: helices

- Refinement of 2° structure & identification of tertiary contacts
  - >50 sequences
  - Identify base-base interactions and higher-order covariations
  - Resolution: base-pairs

- Tertiary modeling
  - Local structures
  - Assembled global structure
  - Resolution: angles and distances
Comparative Analysis of RNA structure