Computational methods for RNA secondary structure determination

Michael Zuker

What is RNA structure?

A 3D atomic resolution structure for *S. cerevisiae* Phe$^{TTC}$ tRNA.
RNA primary structure is the sequence of ribonucleotides (bases), and can be written as $5' - r_1r_2...r_n - 3'$, where $n$ is the number of bases and each $r_i$ is either A (adenine), C (cytosine), G (guanine) or U (uracil). Of course, bases are sometimes modified, either naturally or in vitro.
What is RNA secondary structure?

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. The recent work of Leontis and Westhof describe 18 distinct ways in which any two bases can pair.

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 tetraloops.

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”.

IMA - Oct29-07
What is RNA tertiary structure?

Base triples and pseudoknots are best excluded from the definition of secondary structure. They comprise some of the interactions that characterize tertiary structure. Angles between pairs of adjacent helices are part of tertiary structure. Adding (sufficient) tertiary structure to secondary structure can lead to the modeling of atomic resolution structures.

A helix is a collection of two or more adjacent base pairs.
A 3D-like secondary structure for *S. cerevisiae*

This image preserves some features of the 3D structure.

1TRA yeast Phe tRNA (S. cerevisiae)
The “cloverleaf” secondary structure for *S. cerevisiae*

This traditional secondary structure representation for tRNA does not convey 3D structure information.
Representations of RNA secondary structures.

The traditional representation of secondary structure. Example is a group II intron from Microscilla sp. PRE1. Source: Steve Zimmerly (www.fp.ucalgary.ca/group2introns/species.htm).

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

Colors depend on the probability of base pairs or the probability that a single-stranded base is single-stranded.
Color annotation of secondary structure.

<table>
<thead>
<tr>
<th>Probability Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9990 &lt; Prob. &lt;= 1.0000</td>
</tr>
<tr>
<td>0.9900 &lt; Prob. &lt;= 0.9990</td>
</tr>
<tr>
<td>0.9000 &lt; Prob. &lt;= 0.9900</td>
</tr>
<tr>
<td>0.6500 &lt; Prob. &lt;= 0.9000</td>
</tr>
<tr>
<td>0.3500 &lt; Prob. &lt;= 0.6500</td>
</tr>
<tr>
<td>0.1000 &lt; Prob. &lt;= 0.3500</td>
</tr>
<tr>
<td>0.0100 &lt; Prob. &lt;= 0.1000</td>
</tr>
<tr>
<td>0.0000 &lt; Prob. &lt;= 0.0100</td>
</tr>
</tbody>
</table>

The eight colors used to annotate probabilities. They emphasize the very high and very low probabilities.
Two plots in one.

- Bases are drawn along the circumference of a circle.
- Base pairs are circular arcs the intersect the circle at right angles.
- Black lines (edges) within the circle comprise a “tree representation” of the secondary structure. Every base pair and multi-branched 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.
Side by side comparison - 1

Microscilla sp. PRE1 (M_sp_I1)
Pseudoknot example

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 === Example of a simple pseudoknot.  
U-C-C-G-A-G-G-G
U C-C-C - 5′
C--U--C/

Stem 1: $C_1 \cdot G_{15}$, $C_2 \cdot G_{14}$, $C_3 \cdot G_{13}$,

is incompatible with

Stem 2: $U_8 \cdot A_{23}$, $C_9 \cdot G_{22}$, $C_{10} \cdot G_{21}$, $G_{11} \cdot C_{20}$, $A_{12} \cdot U_{19}$.

Is this possible? – Yes.
3D model of simple pseudoknot. Coordinates by F. Major
Same model - orientation shows coaxial stacking
In the circle plot, intersecting BP arcs indicate a pseudoknot
Can you find the pseudoknot?
Now you can. P3 and P7 create a pseudoknot.
Now you can see them in the original plot.

Tetrahymena thermophila LSU rRNA
GenBank# J01235
Eucarya, Protoctista, Ciliophora (IC1)
Jun 09, 1994

P3 and P7 create a pseudoknot
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 $r_i \cdot r_j$ 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.
Side by side comparison - 2

Structure dot plot for M_sp_I1_phylo

*Counts for each structure include overlap dots.
Prediction of RNA secondary structure

What are the common methods?

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

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

- A “golden rule” in biology: Structure is conserved more than sequence.

- This principle can be used to predict RNA secondary structure.

- It is used together with site directed mutagenesis to confirm the existence of specific base pairs.

- It can be used, for example, to design non-virulent strains of an RNA virus by interrupting significant secondary structure.
SSU rRNA: *Escherichia coli* versus *Deferribacter thermophilus*

**Escherichia coli**

(J01695)
1. Bacteria 2. Proteobacteria 3. gamma subdivision
4. Enterobacteriaceae and related symbionts
5. Enterobacteriaceae 6. *Escherichia*
   Nov 1999

**Deferribacter thermophilus**

(U75602)
1. Bacteria 2. Flexistipes group
3. Deferribacter
   February 2000

Symbols Used In This Diagram:

- G: C - Canonical base pair (A: U, G: C)
- G: U - G: U base pair
- G: A - G: A base pair
- U: U - Non-canonical base pair
Secondary structure comparison between two 16S rRNAs.

Escherichia coli

Compare a small domain in one with the corresponding domain in the other.

- BP is conserved – Both bases unchanged.
- BP is conserved – Both bases change (compensatory change)
- BP is conserved – One base changes (W-C ↔ wobble pair)
- BP not conserved – One base changes (W-C ↔ non-canonical pair)
Need an alignment of homologous sequences

Given $m$ RNA sequences, $R_1, R_2, \ldots, R_m$. The $i^{th}$ sequence has length $n_i$. After alignment, they all have a common length, $n$. They can be written as

\[
R_1 = r_1(1), \ r_1(2), \ r_1(3), \ \ldots, \ r_1(n), \\
R_2 = r_2(1), \ r_2(2), \ r_2(3), \ \ldots, \ r_2(n), \\
R_3 = r_3(1), \ r_3(2), \ r_3(3), \ \ldots, \ r_3(n), \\
\vdots \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \vdots \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \vdots \\
R_m = r_m(1), \ r_m(2), \ r_m(3), \ \ldots, \ r_m(n).
\]

$R_k(i)$ is the $i^{th}$ “base” in the $k^{th}$ sequence. It is A, C, G, U or “-”. The last symbol stands for an inserted gap.

Constructing a “correct” alignment is usually slow and difficult work. Many methods have been developed to automate this procedure.

“Correct” refers to computing an alignment that captures the evolution of these RNAs.
An easy example

The twenty 5S rRNAs shown below all have the same length. Alignment is simple. No gaps are introduced.

```
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
GCUUACGGCAUACCAACGUGCAGCAGCACCAUCGUCCGAGAAGCGUAGAGGUGCCGCUUUAGUACUUGGAUGGAGAAGCAGGUCGUGUAGGCCUUGCUGAAGCUU
```

Then what? How to find BPs conserved by compensatory mutations?
Mutual information content

- \( M(i, j) \) = “mutual information” between columns \( i \) and \( j \). It measures the “degree of correlation”.

- A large \( M(i, j) \) suggests that \( r_k(i) \cdot r_k(j) \) exits for all (or most) \( k \) between 1 and \( m \).

- Base pairs that are 100% conserved yield no mutual information.

- \( M(i, j) \) is the “relative entropy” between a pair of probability distributions. If \( f_{i,j}(B_1, B_2) \) is the observed frequency of the base pair, \( B_1 \cdot B_2 \), in columns \( i \) and \( j \), and if \( f_i(B) \) is the observed frequency of \( B \) in column \( i \), then

\[
M(i, j) = \sum_{B_1, B_2 \in \{A, C, G, U\}} f_{i,j}(B_1, B_2) \log_2 \frac{f_{i,j}(B_1, B_2)}{f_i(B_1)f_j(B_2)}.
\]

Comment: The sum of \( f_{i,j}(B_1, B_2) \) over all pairs and the sum of \( f_i(B) \) over all bases may be < 1, since gaps are ignored.
Mutual information plot for 20 Eukaryote 5S rRNAs

Mutual information levels:
- [1.40-2.00]
- [1.00-1.40)
- [0.70-1.00)
- [0.60-0.70)
Twenty sequences perfectly aligned

- Only (4+4+2+4=14) out of 39 (40 with non-canonical $U \cdot U$) base pairs are identified (35%)

- There is a fair amount of “noise”.

- 100% conserved base pairs not shown. They greatly add to the “noise”, but can be useful to “fill in” or extend stems (helices). Total of 86 base pairs in plot.
Comparative model for *Bombyx mori* 5S rRNA

BPs with $\text{MI} \geq 0.6$ are annotated in color. Its free energy can be used to compare with the minimum energy folding.
Alignment of 302 5S rRNAs from Eukaryotes (M. Szymanski)

| Species          | A_kurodai | G_amansi | P_tenera | C_caldari | S_obliquus | N_flexilis | K_flaccidu | C_scutata | Chlorella | C_reinha1 | T_visurgen | G_phoenic | P_poarum | S_vulgare | R_toruloid | S_salmonic | P_suaveol | C_tussilag | P_ostreat | C_radiatus | T_papilion | T_thermoph | P_polyceph | C_campylu | P_waltlO2 | P_waltlO1 | I_iguana | X_tropic3 | X_laevis1 | X_laevis3 | X_boreal2 | S_kowalev1 | A_pernyi | B_mori1 | D_melanog | H_gammarus | I_illeceb |
|------------------|-----------|----------|----------|-----------|------------|------------|------------|-----------|-----------|-----------|------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|

| Base           | A_kurodai | G_amansi | P_tenera | C_caldari | S_obliquus | N_flexilis | K_flaccidu | C_scutata | Chlorella | C_reinha1 | T_visurgen | G_phoenic | P_poarum | S_vulgare | R_toruloid | S_salmonic | P_suaveol | C_tussilag | P_ostreat | C_radiatus | T_papilion | T_thermoph | P_polyceph | C_campylu | P_waltlO2 | P_waltlO1 | I_iguana | X_tropic3 | X_laevis1 | X_laevis3 | X_boreal2 | S_kowalev1 | A_pernyi | B_mori1 | D_melanog | H_gammarus | I_illeceb |
|----------------|-----------|----------|----------|-----------|------------|------------|------------|-----------|-----------|-----------|------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
“Low noise” MI plot

MI plot: 316 aligned Eukaryotic 5S rRNAs

Mutual information levels:
- [1.40-2.00]
- [1.00-1.40)
- [0.70-1.00)
- [0.60-0.70)

- Most BPs are detected.
- The irregularities in stems is a consequence of embedded gaps in the alignment.
- Alignment BPs must be converted when a particular sequence is extracted and “degapped”.

Wow!
"Low noise" MI plot - *Bombix mori* numbering

Helices appear properly when numbering with respect to a single sequence.
Comparative model for *Bombyx mori* 5S rRNA - Update

BPs with \( \text{MI} \geq 0.5 \) are annotated in color. All base pairs are supported by comparative data.
Transfer RNA – tRNA

- A huge number are known.
- Secondary structure deduced from perhaps 12 sequences in 1969 (Michael Levitt)
- For this presentation, 654 aligned tRNAs were selected from Sprinzl’s database

Sample entry:

DA0380 TGC HALOBACTERIUM CUT. ARCHAE
-GGGCCATAGCTCAGT--GGT--AGAGTGCCCTCCTTTGCAAGGAGGAT-17more-GCCCTGGGTTTCGAATCCCAGTGGGTCCA---
  ==*==== *=== ===* ===== ===== ===== =========*==
  A stem Dstem D aC aC TPsiC TPsiC Astem
MI plot for 654 aligned tRNAs (Sprintzle, 1993)

MI (Mutual Information) levels:

- [1.40, 2.00]
- [1.00, 1.40)
- [0.70, 1.00)
- [0.50, 0.70)

IMA - Oct29-07
With 654 sequences, Secondary structure is very well determined using MI. The quality of the alignment is critical.
MI plot + conserved BPs for 654 aligned tRNAs (Sprintz'l, 1993)

MI (Mutual Information) levels:
Conserved BPs:

- [1.40, 2.00]
- [1.00, 1.40)
- [0.70, 1.00)
- [0.50, 0.70)
MI plot + conserved BPs for 654 aligned tRNAs (Sprintztl, 1993)

Plotting conserved BPs adds noise. Only one extra BP is discovered. Is it worth it?
MI plot & tRNA-TGC *Halobacterium cutirubrum*

**MI (Mutual Information) levels:**
- Conserved BPs:
  - [1.40, 2.00]
  - [1.00, 1.40)
  - [0.70, 1.00)
  - [0.50, 0.70)

Output of sir_graph (®)
mfold_util 4.0

---

*Halobacterium cutirubrum* tRNA-TGC

5'
G
G
G
C
C
C
AT
A
GCTC
AG
T
G
G
T
T
C
GAA
T
C
C
C
A
GT
G
G
G
T
C
C
A
5'
3'
Energy Minimization

- How is energy assigned? Answer: “nearest-neighbor” energy rules are used.

- A stem with \( n \) BPs is broken into \( n - 1 \) “BP stacks”. Energy \( \Delta G \) is assigned to the “BP stacks”, but takes into account hydrogen bonds and stacking. These energies are negative “favorable”.

- Mismatched BPs at the ends of stems also contribute to stability.

- The loops are destabilizing.
\( \delta G \) for BP stacks

NN (nearest neighbor) free energies for RNA at 37°.
Doug Turner’s group at the University of Rochester.

\[
\begin{align*}
\delta G \left( \begin{array}{c} 5' - CGAGTATTCGG - 3' \\
3' - GCTCATAAGCC - 5' \end{array} \right) &= \delta G \left( \begin{array}{c} 5' - CG - 3' \\
3' - GC - 5' \end{array} \right) + \\
\delta G \left( \begin{array}{c} 5' - GA - 3' \\
3' - CT - 5' \end{array} \right) + \delta G \left( \begin{array}{c} 5' - AG - 3' \\
3' - TC - 5' \end{array} \right) + \delta G \left( \begin{array}{c} 5' - GT - 3' \\
3' - CA - 5' \end{array} \right) + \\
\delta G \left( \begin{array}{c} 5' - TA - 3' \\
3' - AT - 5' \end{array} \right) + \delta G \left( \begin{array}{c} 5' - AT - 3' \\
3' - TA - 5' \end{array} \right) + \delta G \left( \begin{array}{c} 5' - TT - 3' \\
3' - AA - 5' \end{array} \right) + \\
\delta G \left( \begin{array}{c} 5' - CG - 3' \\
3' - GC - 5' \end{array} \right) + \delta G \left( \begin{array}{c} 5' - GG - 3' \\
3' - CC - 5' \end{array} \right)
\end{align*}
\]

- Don’t sum “scores” as in sequence alignment.
- Consider two BPs at a time.
- Consecutive \( \delta G \)'s are not independent.
Other stacking free energies
\( \delta G \) for mismatched pairs and dangling

In the example structure on the left:

- Stacking of the \( C_{13} \cdot A_{19} \) mismatch stabilizes the H-loop.
- Stacking of the \( C_{4} \cdot C_{27} \) mismatch and the \( U_{8} \cdot U_{24} \) mismatch stabilizes the I-loop.
- These negative (favorable) energies are added to the unfavorable (positive) energies of the H-loop or the I-loop. \textbf{They are really associated with the adjacent stem. The energy assignment to the loop is done for algorithmic reasons.}
- Stacking of single bases at the end of stems is also considered (not shown).
In the same example structure on the left:

- Both the H-loop and the I-loop have penalty energies that grow logarithmically with loop size (number of single-stranded bases).

- \( \delta G \approx 1.75RT \ln(l) \), for loop size \( l \).

- In addition, there is an I-loop asymmetry penalty. The asymmetry of the I-loop in the example is \( 1 = |5 - 4| \). This is the difference between the number of single-stranded bases on each side of the loop.
The energy dot plot (EDP)

- The minimum free energy (mfe) of a folding, $\Delta G_{\text{mfe}}$, can be computed.
- Let $\delta G \geq 0$ be a free energy increment.
- All secondary structures with free energies between $\Delta G_{\text{mfe}}$ and $\Delta G_{\text{mfe}} + \delta G$ are superimposed in a single plot.
- Different colors are used for different values of $\delta G$.
- For $\delta G_1 < \delta G_2$, with colors $c_1$ and $c_2$, respectively, $c_1$ is “on top” and obscures $c_2$.

The probability dot plot (PDP)

- All base pairs with probabilities above some cutoff are plotted as “dots”.
- The area of the dots is (usually) proportional to the probability.
- The dots are colored to indicate a probability range.
The folding problem is ill-conditioned in general

As bad as it gets.

Three conjugate\footnote{Two foldings are called conjugate if they share no base pairs in common.} foldings of a short sequence. From left to right, $\Delta G = -4.9$ kcal/mol (mfe folding), $\Delta G = -4.7$ and $\Delta G = -4.6$!
Other views of the “three conjugate” foldings example

Clockwise from above left:


2. EDP (energy dot plot). Superposition of all possible foldings within 0.3 kcal/mol from mfe.

3. PDP (probability dot plot) A plot of all base pairs with probabilities ≥ 0.01. The maximum probability is 0.44 (poor).
Heresy: Statistics on one Datum

For a single secondary structure, predicted or otherwise:

- Not all base pairs are created equal.
- The EDP and PDP can be used to distinguish.
- In a secondary structure plot:
  1. color base pairs according to the probability that they form.
  2. color single-stranded bases according to the probability that they are not paired.
- Higher probability base pairs are more likely to be correct.
The consensus structure.

The consensus structure contains all base pairs whose probabilities are $> 50\%$, and no others. (Ye Ding & Chip Lawrence call this the “centroid”.)

Quiz:

- Prove that the consensus structure is valid (no base triples). [20 points]
- Prove that the consensus structure contains no pseudoknots. [40 points]
Good versus Bad RNAs

- A good RNA has an “uncluttered” EDP or PDP. Base pairs in a mfe folding are more likely to be correct. High probability base pairs are “usually” correct, but a good RNA has a lot more high probability base pairs than a bad RNA.

- A bad RNA has a “cluttered” EDP or PDP. A mfe folding has fewer correct base pairs. Even though high probability base pairs are likely to be correct, there are relatively few of them.
Example: A good RNA versus a bad RNA.

E. californium 5S rRNA

D. discoidium 5S rRNA

“Uncluttered” PDP versus a “cluttered” one.
Consensus structures for the same pair of 5S rRNAs

Not much predicted with confidence *versus* a lot predicted with high confidence. In this case, the second folding is a very good prediction.
The numbers

Structure dot plot for E_califor

Full Overlap 29 1 dG = ? 38° 2 Energy is Undef.

*Counts for each structure include overlap dots.

Structure dot plot for D_discoid

Full Overlap 12 1 dG = ? 38° 2 Energy is Undef.

*Counts for each structure include overlap dots.
Entropy: Quantifying “good” and “bad”

If \( P = \{ p_i \}_{i=1}^{\infty} \) is a countable probability distribution (may be finite), then

\[
\sum_{i=1}^{\infty} p_i = 1.
\]

The (Shannon) entropy of \( P \) is defined by

\[
\mathcal{S}(P) = -\sum_{i=1}^{\infty} p_i \log(p_i).
\]

(Treat 0 \( \log(0) \) as 0.) For infinite distributions, the entropy might be \( +\infty \). We do not consider infinite or continuous distributions here. The base for the logarithm is not specified and may be chosen arbitrarily to scale the entropy. For example, base 2 gives the entropy in “bits”.

When there are \( n \) “outcomes”, \( 1 \leq i \leq n \) and the maximum entropy distribution sets \( p_i = 1/n \), so that \( \mathcal{S}(P) = -\sum_{i=1}^{n} p_i \log(p_i) = \log(n) \).

If \( \mathcal{F}(R) \) is the set of all foldings of \( r_1 r_2 \ldots r_n \), then the maximum entropy loss to select a single folding is given by making all foldings equally likely, and so:

\[
\mathcal{S}(\mathcal{F}(R)) \sim Cn - 1.5 \ln(n) \text{ as } n \to \infty, \text{ for some constant } C. \text{ What is } C?\]
The expected number of foldings

Suppose that an RNA sequence is totally random, with equal probabilities for A, C, G and U. Then the probability that two different bases can pair is $\frac{3}{8}$. This represents $1/16$ for A·U, U·A, C·G, G·C, G·U and U·G. In general, call this probability $p$.

Many years ago, I showed that the expected number of foldings, $T_n$, on a random RNA sequence of length $n$ is given by

$$T_n \sim \frac{(1 + 4\sqrt{p})^{\frac{1}{4}}}{2\sqrt{\pi p}^{\frac{3}{4}}} n^{-\frac{3}{2}} \left( \frac{1 + \sqrt{1 + 4\sqrt{p}}}{2} \right)^{2n + 2},$$

as $n \to \infty$.

This implies a maximum entropy of $2\log \left( \frac{1 + \sqrt{1 + 4\sqrt{p}}}{2} \right)$ per base. This reduces to $2\log(1 + \sqrt{1 + 4\sqrt{p}}) - 2$ bits/base.

For $p = 3/8$, the maximum entropy is $1.029$ per base. For $p = 1/4$, it is $0.900$ bits per base.
The entropy of the Boltzmann distribution of all foldings.

The partition function for $\mathcal{F}(R)$ is defined by

$$Z = \sum_{F \in \mathcal{F}(R)} \exp \left( - \frac{\Delta G(F)}{RT} \right)$$

The probability of folding $F$ is

$$e^{-\frac{\Delta G(F)}{RT}}$$

The log probability is

$$-\frac{\Delta G(F)}{RT} - \ln Z$$

so

$$\mathcal{S}(\mathcal{F}(R)) = \frac{\Delta G(F) - \Delta G_{\text{ens}}}{RT}.$$ 

$\Delta G_{\text{ens}}$ is computed directly by standard algorithms and $\overline{\Delta G(F)}$ can be estimated by averaging the free energies of a stochastic sample of foldings. Replacing $RT$ with $RT \ln(2)$ expresses the entropy in bits.

I have chosen entropy per base to be the overall measure of “well-definedness” for RNA folding.
Entropy of random RNA

For each of 302 Eukaryote 5S rRNAs, 100 random sequences were generated with the same length and dinucleotide distribution. For each of these, the entropy of the Boltzmann distribution was computed with a sample of 100 foldings. GC content in the random sequences was 55.57 ± 3.79%.

The entropy was 0.226 ± 0.007 bits/base.

The entropies of 100 randomized versions of the RNase P RNA from *Desulfovibrio desulfuricans* (length 360) gave similar results, as did foldings of the first 150 bases of these sequences.
Good and bad revisited

- Good refers to low entropy RNAs, for which many base pairs are predicted with high probability.
- Bad refers to high entropy RNAs, for which few base pairs are predicted with high probability.
- Problems: As is usual in biology, there are exceptions to each “rule”.
  1. In “good” or “bad” RNAs (especially for “good”), one expects that base pairs with high probability should be correct. Some high probability base pairs are not correct and some correct base pairs have very low probabilities.
  2. In homologous RNAs that are similar enough to be readily aligned, some high probability base pairs in one may be low probability base pairs in the other, and vice versa.
<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>( \text{min} \mathcal{S} )</th>
<th>( \bar{\mathcal{S}} )</th>
<th>( \text{max} \mathcal{S} )</th>
<th>( r ) GC content</th>
<th>( r ) ( \delta G_{\text{mfe}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5S rRNA Eubacteria</td>
<td>439</td>
<td>0.123</td>
<td>0.223</td>
<td>0.311</td>
<td>-0.308</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Thermus sp1</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Planctomyces limnophilus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5S rRNA Archaea</td>
<td>57</td>
<td>0.063</td>
<td>0.165</td>
<td>0.246</td>
<td>-0.386</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Sulfolobus faci2</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Halobacterium saccharovorum</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5S rRNA Eukaryota</td>
<td>302</td>
<td>0.106</td>
<td>0.182</td>
<td>0.264</td>
<td>-0.119</td>
<td>0.543</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Lineus geniculatus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Spirogyra sp</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group I introns</td>
<td>81</td>
<td>0.161</td>
<td>0.212</td>
<td>0.307</td>
<td>-0.431</td>
<td>0.362</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Staurastrum sp. Chl (M753) SSU rRNA</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Zea mays Chl tRNA Leu</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group II introns</td>
<td>51</td>
<td>0.177</td>
<td>0.217</td>
<td>0.258</td>
<td>-0.770</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Trichodesmium Chl erythraeum (sp I1)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Sinorhizobium meliloti (I1)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNase P Eubacteria</td>
<td>179</td>
<td>0.161</td>
<td>0.205</td>
<td>0.263</td>
<td>-0.573</td>
<td>0.762</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Desulfovibrio desulfuricans</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Campylobacter jejuni</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNase P Archaea</td>
<td>36</td>
<td>0.162</td>
<td>0.213</td>
<td>0.278</td>
<td>-0.745</td>
<td>0.693</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Aeropyrum pernix</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Methanococcus maripaludus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tRNA</td>
<td>803</td>
<td>0.090</td>
<td>0.201</td>
<td>0.308</td>
<td>-0.207</td>
<td>0.443</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Chl) Euglena gracilis CAT</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Minimum, average and maximum entropies, in bits/base, for different types of RNA. \( r \) is the correlation of entropy with GC content or minimum free energy.
Best versus worst. 5S rRNA from Archaea. I

**Sulfolobus faci2**

Probability Dotplot for Sulfaci2

0.0010 <= Probability <= 1.0000

0.9990 < Prob. <= 1.0000
0.9900 < Prob. <= 0.9990
0.9000 < Prob. <= 0.9900
0.6500 < Prob. <= 0.9000
0.3500 < Prob. <= 0.6500
0.1000 < Prob. <= 0.3500
0.0100 < Prob. <= 0.1000
0.0000 < Prob. <= 0.0100

Upper Triangle Base Pairs Plotted: 54

**Halobacterium saccharovorum**

Probability Dotplot for Halosacc

0.0010 <= Probability <= 0.9532

0.9990 < Prob. <= 1.0000
0.9900 < Prob. <= 0.9990
0.9000 < Prob. <= 0.9900
0.6500 < Prob. <= 0.9000
0.3500 < Prob. <= 0.6500
0.1000 < Prob. <= 0.3500
0.0100 < Prob. <= 0.1000
0.0000 < Prob. <= 0.0100

Upper Triangle Base Pairs Plotted: 382
Best versus worst. 5S rRNA from Archaea. II

*Sulfolobus faci2*

*Halobacterium saccharovorum*

IMA - Oct29-07
**Best versus worst. 5S rRNA from Archaea. III**

_Sulfolobus faci2_

Structure dot plot for Sulfaci2 phylo/consen

*Counts for each structure include overlap dots.

_Halobacterium saccharovorum_

Structure dot plot for Halosacc phylo/consen

*Counts for each structure include overlap dots.
Best versus worst. 5S rRNA from Eubacteria. I

Thermus sp1

Planctomyces limnophilus

Probability Dotplot for Thermu_sp1

0.0010 <= Probability <= 1.0000

0.9990 < Prob. <= 0.9990

0.9000 < Prob. <= 0.9000

0.6500 < Prob. <= 0.6500

0.3500 < Prob. <= 0.3500

0.1000 < Prob. <= 0.1000

0.0100 < Prob. <= 0.0100

0.0000 < Prob. <= 0.0000

Upper Triangle Base Pairs Plotted: 110

Probability Dotplot for Pmyc_limno

0.0010 <= Probability <= 0.9996

0.9990 < Prob. <= 1.0000

0.9900 < Prob. <= 0.9990

0.9000 < Prob. <= 0.9900

0.6500 < Prob. <= 0.9000

0.3500 < Prob. <= 0.6500

0.1000 < Prob. <= 0.3500

0.0100 < Prob. <= 0.1000

0.0000 < Prob. <= 0.0100

Upper Triangle Base Pairs Plotted: 254
Best versus worst. 5S rRNA from Eubacteria. II

Thermus sp1

Planctomyces limnophilus

Output of sir_graph (®)
mfold_util 4.0
Best versus worst. 5S rRNA from Eubacteria. III

Thermus sp1

Planctomyces limnophilus

*Counts for each structure include overlap dots.

IMA - Oct29-07
Problems I

The tandem base pairs in black have probabilities of roughly 1/2 of 1 percent. They occur in the phylogenetic models for 5S rRNA and seem correct in terms of alignment and covariation. This motif strongly contradicts the energy rules. What is happening here?

The table below shows good conservation evidence for the base pair \( \text{A}_{27} \cdot \text{U}_{52} \). MI = 0.85 bits.

<table>
<thead>
<tr>
<th>Number</th>
<th>.</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>246</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>16</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>246</td>
</tr>
</tbody>
</table>

Lowest entropy

\( L. \text{genicul} \)
Problems II

The U3 RNAs from *L. collosoma* and *T. brucei* are problematic. Their entropy is high, but the problem is that they contradict each other.
Rfam - Database of aligned RNAs with consensus structure

The Rfam database, at
www.sanger.ac.uk/Software/Rfam/,
and at rfam.janelia.org/
is a significant and growing web resource.

- Currently, 607 families. Each family has a seed file as well as a full alignment file.
- The purpose of the database is to use families to search genomic data for similar sequences.
- Many entries contain little or no mutual information to support consensus structures annotated in the alignment files.

A total of 14654 sequences from 607 Rfam seed files were processed.
HDV_ribozyme - A lowest entropy example

The M28267.1/87-775 entry has the lowest entropy out of 15 seed sequences. Not just the lowest, but low, at 0.08 bits/base. Part of the Rfam consensus structure has very low probability.
Some remarks on Rfam

The use of single sequence entropy calculations on Rfam RNAs might identify families for which the alignment needs to be improved, or families that might better be split into two (or more) groups.

Low entropy foldings might be useful in generating an initial folding for a family that could then be refined as needed using the other members together with a multiple sequence alignment.