Predicting RNA secondary structure

Computational Aspects of Molecular Structures
Lecture 7
Teresa Przytycka, PhD
RNA as a structural molecule, information transfer molecule, information decoding molecule

(B)

5' end

3' end

Acceptor arm

D-arm

Anticodon arm

Anticodon

5' cap

2'-O-methyl purine

7-methylguanosine in 5' to 5' triphosphate linkage

variable sequence and structure

mRNA

rRNA

tRNA

3' poly A tail

20-200
Five types of bases
Complementary nucleosides

Key:
- Dotted line: Hydrogen bond

A

T (DNA)

U (RNA)

G

C
RNA folding is hierarchical

- At the first level of organization is the primary structure, which is the sequence of nucleotides.
- The next level is secondary structure, the sum of the canonical (AU, CG, and GU) base-pairs.
- Tertiary structure is the three-dimensional arrangement of atoms.
- The quaternary structure is the interaction with other molecules, which are often either proteins or other RNA strands.
Motivation behind RNA secondary structure prediction

• Secondary structure contacts are generally stronger than tertiary structure contacts.

• The formation of secondary structure occurs on a faster timescale than tertiary structure.

• Therefore, RNA secondary structure can generally be predicted without knowledge of tertiary structure.
Stacking

- Base-pairs are usually coplanar
- are almost always stacked
- steams – continuous stacks
- 3D structure of a stack is a helix
Example of RNA secondary structure naming conventions

- unstructured single strand
- bulge loop
- steam
- multi-brunch loop
- interior loop
- hairpin loop
- hairpin
Basic properties RNA secondary structure

• Base pairs almost always exhibit a clear nested pattern: if \(i,j\) and \(i',j'\) where \(i < i'\) are indexes of two base pairs then non-nesting translates to one of the following conditions:
  1. \(i < j < i' < j'\)
  2. \(i < i' < j' < j\)

• Secondary structure – such maximal nested set of base pairs.

• Base pairs that do not follow the nested pattern are pseudo-knots.

\[
\text{AUCGAUAAU} \quad \text{AUCGAUAAU}
\]

\text{psedoknot)
Bacillus subtilis RNase P RNA

M - multi-loop
I - interior loop
B - bulge loop
H - hairpin loop

Circular representation of the secondary structure form the previous slide
Dot plot representation of the Rnase (*B. subtilis*) folding
Main approaches to RNA secondary structure prediction

- Energy minimization
  - dynamic programming approach
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure

- Comparative sequence analysis
  - use phylogenetic information/sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
Dot plot

Class work: Predict secondary structure for RNA “ACGUGCGU” assuming -1 for a standard pair of 0 for any non-standard pair.
Dynamic programming approach

- Solve problem for all sub problems of size 1 and 2 (the solution is zero in both cases)
- Iteratively, knowing the solution of all problems of size less than k compute the solution of all problems of size k.
The subproblems

• Input $X = x_1, x_2, x_3, x_4, x_5, x_6, \ldots, x_n$
• Subproblems of size 2:

\[ x_1, x_2, x_3, x_4, x_5, x_6, \ldots, x_n \]

• Subproblems of size 4:

\[ x_1, x_2, x_3, x_4, x_5, x_6, \ldots, x_n \]
Dynamic programming approach

Let $E(i,j)$ = minimum energy for sub-chain starting at $i$ and ending at $j$

$\alpha(r_i,r_j)$ = energy of pair $r_i$, $r_j$ ($r_j$ = base at position $j$)

a) $i,j$ is paired $E(i,j) = E(i+1,j-1) + \alpha(r_i,r_j)$
b) $i$ is unpaired $E(i,j) = E(i+1,j) + E(j,j)$ Since $E(j,j) = 0$ cases b and c are included in case d

c) $j$ is unpaired $E(i,j) = E(i,j-1) + E(i,i)$
d) bifurcation $E(i,j) = E(i,k)+E(k+1,j)$
RNA secondary structure algorithm

- **Given:** RNA sequence $x_1,x_2,x_3,x_4,x_5,x_6,...,x_L$
- **Initialization:**
  
  for $i = 1$ to $L$ do $E(i, i) = 0$
  
  for $i = 1$ to $L-1$ do $E(i, i+1) = 0$ (some versions of the algorithm assume that the base pair between $i$ and $i+1$ is possible. In this case this line is removed and the recursion starts with $n=1$. Zuker algorithm, puts 0 even on the next diagonal then $n$ starts with $n=3$)
- **Recursion:**
  
  for $n = 2$ to $L$  
  
  for $i = 1$ to $L-n$ do  
  
  $j = i + n$
  
  $E(i,j) = \min\{ E(i+1,j-1)+ \alpha(r_i,r_j) , \min_{i<=k<j} \{E(i,k)+E(k+1,j)}\}$
- **Cost:** $O(n^3)$
Let \( s(r_i, r_j) = -1 \) if \( r_i, r_j \) form a base pair and 0 otherwise (this variant is known as Nussinov algorithm).

Input: GGAAAUCC

\[
\begin{array}{cccccccc}
G & 0 & 0 &   &   &   &   &   &   \\
G & 0 & 0 &   &   &   &   &   &   \\
A &   &   & 0 & 0 &   &   &   &   \\
A &   &   & 0 & 0 &   &   &   &   \\
A &   &   & 0 & 0 &   &   &   &   \\
A &   &   &   &   &   &   &   &   \\
U &   &   &   &   &   & 0 & 0 &   \\
C &   &   &   &   &   & 0 & 0 &   \\
C &   &   &   &   &   &   &   & 0 \\
\end{array}
\]

\( E(i, j) \) = lowest energy conformation for subchain from \( i \) to \( j \)

Here we should have min energy for AAAUC
Example-continued

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>U</th>
<th>C</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \text{GGA} \]
\[
\min \left\{ E(G) + \alpha(GA), \min \left\{ E(G) + E(GA), E(GG) + E(A) \right\} \right\} = 0
\]

\[ \text{AAU} \]
\[
\min \left\{ E(A) + \alpha(AU), \min \left\{ E(A) + E(AU), E(AA) + E(U) \right\} \right\} = -1
\]
Example-continued

GGAA
\[
\min \{ E(GA) + \alpha(GA) \}
\min \{ E(G) + E(GAA), 
E(GA) + E(AA), 
E(GGA) + E(A) \}
\] = 0

AAAU
\[
\min \{ E(AA) + \alpha(AU) \}
\min \{ E(A) + E(AAU), 
E(AA) + E(AU), 
E(AAA) + E(U) \}
\] = -1

AAUC
\[
\min \{ E(AU) + \alpha(AC) \}
\min \{ E(A) + E(AUC), 
E(AA) + E(UC), 
E(AAU) + E(C) \}
\] = -1
Example-continued

Optimal solution

GAAAUC

\[
\begin{align*}
\text{min} \ {E(AAAU)} &+ \alpha(GC) \\
\text{min} \ \{E(G)+E(AAAUC), E(GA)+E(AAUC), E(GAA)+E(AUC), E(GAAA)+E(UC), E(GAAAA)+E(C)\} \\
\end{align*}
\]

\[= -2\]

Stacking is shown as a diagonal

Secondary structure - hairpin
Add base pair (i,j) and start tracing from cell i+1,j-1.

There is no base pair at (i,j) brunch and go to cells (i,k) and (k+1,j) and continue recovering base pairs from there.
RNA secondary structure algorithm

- **Given:** RNA sequence \(x_1,x_2,x_3,x_4,x_5,x_6,\ldots,x_L\)
- **Initialization:**
  
  for \(i = 1\) to \(L\) do \(E(i, i) = 0\)
  
  for \(i = 1\) to \(L-1\) do \(E(i, i+1) = 0\)
  
  Recursion:
  
  for \(n = 2\) to \(L\) # *iteration over length*
  
  for \(i = 1\) to \(L-n\) do # *iteration over subsequences of length \(n\)*
  
  \[ j = i + n \]
  
  \[ E(i,j) = \min\{ E(i+1,j-1) + \alpha(r_i,r_j) , \]
  
  \[ \min_{i\leq k < j} \{E(i,k)+E(k+1,j)\} \}
  
  if \(E(i,j) < E(i+1,j-1) + \alpha(r_i,r_j)\)
  
  \[ \text{trace}_\text{back}(i,j) = \]
  
  value \(k\) minimizing \(E(i,k)+E(k+1,j)\)
- **Cost:** \(O(n^3)\)
More realistic energy function

Loops have destabilizing effect structure (d) should have lower energy that (b).

Destabilizing contribution of loops should depend on the loop length (k).

Stacking has additional stabilizing contribution $\eta$.

$\alpha(i,j)$ energy of a base pair

Nearest neighbor energy function takes into account neighboring elements but non long range effects
More realistic energy function requires slightly more involved recurrence

\[ E(i,j) = \min \{ E(i+1,j), E(i,j-1), \]
\[ \min \{ E(i,k) + E(k+1,j), \]
\[ L(i,j) \} \]

where

\[ L(i,j) = \begin{cases} 
\alpha(r_i, r_j) + \xi (j-i-1) & \text{if } L(i,j) \text{ is a hairpin loop;} \\
\alpha(r_i, r_j) + \eta + E(i+1, j-1) & \text{if hairpin} \\
\min_k \{ \alpha(r_i, r_j) + \beta(k) + E(i+k+1, j-1) \} & \text{if } i\text{-bulge} \\
\min_k \{ \alpha(r_i, r_j) + \beta(k) + E(i+1, j-k-1) \} & \text{if } j\text{-bulge} \\
\min_{k_1, k_2} \{ \alpha(r_i, r_j) + \gamma(k_1 + k_2) + E(i+k_1+1, j-k_2-1) \} & \text{if internal loop} 
\end{cases} \]

Extra “\( \min \)” gives \( O(n^4) \) algorithm
One step further...

- A popular RNA secondary structure prediction algorithm MFOLD (Zuker) uses tables for loop free energies measured experimentally and interpolated where not measured.

Furthermore it is known that the energy depends on the structure in each hairpin loop. Thus MFOLD uses a tables of trieloops and tetraloops (loops of size 3 and 4).

**Figure 1:** The `loop.dg` or `loop.TC` contains size based free energy increments for hairpin, bulge and interior loops up to size 30. Entries with ‘.’ are undefined.

<table>
<thead>
<tr>
<th>SIZE</th>
<th>INTERNAL</th>
<th>BULGE</th>
<th>HAIRPIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>3.8</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>3.2</td>
<td>5.6</td>
</tr>
<tr>
<td>4</td>
<td>1.7</td>
<td>3.6</td>
<td>5.5</td>
</tr>
<tr>
<td>5</td>
<td>1.8</td>
<td>4.0</td>
<td>5.6</td>
</tr>
<tr>
<td>6</td>
<td>2.0</td>
<td>4.4</td>
<td>5.3</td>
</tr>
<tr>
<td>7</td>
<td>2.2</td>
<td>4.6</td>
<td>5.8</td>
</tr>
<tr>
<td>8</td>
<td>2.3</td>
<td>4.7</td>
<td>5.4</td>
</tr>
<tr>
<td>30</td>
<td>3.7</td>
<td>6.1</td>
<td>7.7</td>
</tr>
</tbody>
</table>
Including restricted pseudoknots types into RNA secondary structure

Rivas and Eddy JMB, 1999, 2053-2068.

Recall Nussinov:

\[
    \max_{i,j} = \max_k \left[ \right] = \max_{i,j,k} \left[ \right]
\]

Rives ad Eddy idea:

\[
    \max_{i,k,l,j} \left[ \right] = \max_{i,j,k} \left[ \right] = \max_{i,j,k} \left[ \right]
\]

Informally, pseudoknots are accepted if all base pairs can be embedded on the upper or lower half plane without crossings.
Example of a pseudoknot which can be resolved this way

![Diagram of a pseudoknot]

Complexity: $O(n^6)$ – ok. for single RNA; problematic for the whole data base
Further generalizations are possible on respectively higher cost
Quantities measure of pair-wise sequence covariation

Mutual information $M_{ij}$ between two aligned columns $i, j$

$$M_{ij} = \sum_{x_i x_j} f_{x_i x_j} \log_2 \left( \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}} \right)$$

Where

- $f_{x_i x_j}$ frequency of the pair (observed)
- $f_{x_i}$ frequency of nucleotide $x_i$ at position $i$

Observations:

- $0 \leq M_{ij} \leq 2$
- $i, j$ uncorrelated $M_{ij} = 0$
The need for suboptimal structure prediction

- The free energy in the “nearest neighbor” model is incomplete.
- Some known sequence effects on stability are non-nearest-neighbor. (The stabilities of model bulge loops and single non-canonical pairs show non-nearest-neighbor effects.)
- Some factors are not included in dynamic programming algorithms
- Not all RNA sequences are at equilibrium
- Some RNA sequences have more than one conformation.
Exhaustive suboptimal structure determination

• Wuchty, Fontana, Hofacker, Schuster; Biopolymers 1999: Modification of the dynamic programming algorithm so that it finds all suboptimal substructures within a given increment of energy form the optimal structure

• Number of secondary structures grows exponentially with increasing energy increment
Statistical Sampling

- Ding and Lawrence, 2003
- Efficient dynamic programming algorithm that samples suboptimal secondary structures from the Boltzmann ensembles of structures
- Method: Randomizing the trace back walk
- Application: can be used to compute probability of structural features
- Demonstrated that it is better to take a “centroid” as the predicted structure as opposed to the energy minimum structure.
- Software name: Sfold
Resources

• Vienna RNA secondary structure prediction web site:
  http://ww.tbi.univie.ac.at/~ivo/RNA/

• Mfold
  http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html

AAGACUUCGGUCUGGGCCGACAUUC
Covariance method

- In a correct multiple alignment RNAs, conserved base pairs are often revealed by the presence of frequent correlated compensatory mutations.

```
GCCUUCGCGGC
GACUUCGCGGUG
GGCUUCGCGGCC
```

Two boxed positions are co-varying to maintain Watson-Crick complementary. This covariation implies a base pair which may be then extended in both directions.
Figure 5.12. Covariation found in tRNA sequences reveals base interactions in tRNA secondary and tertiary structures. (A) Alignment of tRNA sequences showing regions of interacting base pairs. (+) Transition; (−) transversions; (?) ambiguous nucleotide. (B) Diagram of tRNA structure illustrating base–base interactions revealed by a covariation analysis. Adapted from the Web site of R. Gutell at http://www.rna.icmb.utexas.edu.
A

Phe: *Agmenellum quadruplicatum*

*Phe: Spinacea oleracea*

**RDGD**

85.3% Similarity

**Phe: S. cerevisiae**

**Ala: Thp. tenax**

**RDGD**

53.4% Similarity

B

Accepter stem ---- D stem ---- TΨC stem
Examples

\[
\begin{array}{c|c|c}
\text{A} & \text{U} & \text{A} \\
\text{A} & \text{U} & \text{U} \\
\text{C} & \text{G} & \text{C} \\
\text{G} & \text{U} & \text{G} \\
\end{array}
\]

\[
\begin{align*}
F_{Ai} &= .5 \\
F_{Ci} &= .25 \\
F_{Gi} &= .25 \\
F_{Uj} &= .5 \\
F_{Cj} &= .25 \\
F_{Gj} &= .25 \\
\end{align*}
\]

\[
\begin{align*}
F_{AU} &= .5 \\
F_{CG} &= .25 \\
F_{GC} &= .25 \\
\end{align*}
\]

\[
M_{ij} = \sum_{x_ix_j} f_{x_ix_j} \log_2 \left( \frac{f_{x_ix_j}}{f_{x_i} f_{x_j}} \right) =
\]

\[
.5 \log_2 \left( \frac{.5}{(.5* .5)} \right) + 2 \times .25 \log_2 \left( \frac{.25}{(.25* .25)} \right) =
\]

\[
.5 \times 1 + .5 \times 2 = 1.5
\]

\[
\begin{array}{c|c|c}
\text{A} & \text{U} & \text{A} \\
\text{A} & \text{U} & \text{U} \\
\text{A} & \text{U} & \text{U} \\
\end{array}
\]

\[
M_{ij} = 4 \times .25 \log 4 = 2
\]

\[
\begin{array}{c|c|c}
\text{U} & \text{A} & \text{G} \\
\text{A} & \text{C} & \text{G} \\
\end{array}
\]

\[
M_{ij} = 1 \log 1 = 0
\]
Example of prediction based on covariance

Secondary Structure of Vertebrate Telomerase RNA

Telomerase is a ribonucleoprotein enzyme that maintains telomere length by adding telomeric sequence onto chromosome ends.

Method:
To determine the secondary structure of vertebrate telomerase RNA, 32 new telomerase RNA genes were cloned and sequenced: 18 mammals, 2 birds, 1 reptile, 7 amphibians, and 4 fishes.
Next step: alignment

8 conserved regions found (here shown the first one CR1)
Next step: covariation analysis

• Conserved regions only
• Standard pairs – positive evidence
• Non-canonical base pairs G/U; G/A; C/A also considered – neutral
• Other pairs – negative evidence
Figure 2. Proposed Secondary Structure of Vertebrate Telomerase RNAs
Differences between the structures can be examined ...
Figure 5. Comparison of Ciliate and Vertebrate Telomerase RNA Structures

The outline of the minimum-consensus structures of ciliate and vertebrate telomerase RNAs are shown. Template regions are indicated with black filled boxes. Vertebrate-specific structural elements are shaded. The diagram on the left illustrates a possible evolutionary course from the ancestral telomerase RNA to ciliate and vertebrate RNAs.
Revolution in RNA Secondary Structure Prediction

David H. Mathews

RNA structure formation is hierarchical and, therefore, secondary structure, the sum of canonical base-pairs, can generally be predicted without knowledge of the three-dimensional structure. Secondary structure prediction algorithms evolved from predicting a single, lowest free energy structure to their current state where statistics can be determined from the thermodynamic ensemble. This article reviews the free energy minimization technique and the salient revolutions in the dynamic programming algorithm methods for secondary structure prediction. Emphasis is placed on highlighting the recently developed method, which statistically samples structures from the complete Boltzmann ensemble.

Keywords: RNA secondary structure prediction; free energy; partition function; nearest neighbor parameters; dynamic programming algorithm