RNA folding & ncRNA discovery

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB
Contents

- Non-coding RNAs and their functions
- RNA structures
- RNA folding
  - Nussinov algorithm
  - Energy minimization methods
- microRNA target identification
ncRNAs have important and diverse functional and regulatory roles that impact gene transcription, translation, localization, replication, and degradation

- Protein synthesis (rRNA and tRNA)
- RNA processing (snoRNA)
- Gene regulation
  - RNA interference (RNAi)
  - Andrew Fire and Craig Mello (2006 Nobel prize)
- DNA-like function
  - Virus
- RNA world
**Non-coding RNAs**

- A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein; small RNA (sRNA) is often used for bacterial ncRNAs.
- tRNA (transfer RNA), rRNA (ribosomal RNA), snoRNA (small RNA molecules that guide chemical modifications of other RNAs)
- microRNAs (miRNA, μRNA, single-stranded RNA molecules of 21-23 nucleotides in length, regulate gene expression)
- siRNAs (short interfering RNA or silencing RNA, double-stranded, 20-25 nucleotides in length, involved in the RNA interference (RNAi) pathway, where it interferes with the expression of a specific gene.)
- piRNAs (expressed in animal cells, forms RNA-protein complexes through interactions with Piwi proteins, which have been linked to transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells)
- long ncRNAs (non-protein coding transcripts longer than 200 nucleotides)
Riboswitch

- What’s riboswitch
- Riboswitch mechanism

Structures are more conserved

- Structure information is important for alignment (and therefore gene finding)
Features of RNA

- RNA typically produced as a single stranded molecule (unlike DNA)
- Strand folds upon itself to form base pairs & secondary structures
- Structure conservation is important

- RNA sequence analysis is different from DNA sequence
Canonical base pairing

Watson-Crick base pairing
Non-Watson-Crick base pairing G/U (Wobble)
tRNA structure
RNA secondary structure

- Stem
- Single-Stranded
- Interior Loop
- Bulge Loop
- Hairpin loop
- Pseudoknot
- Junction (Multiloop)
Complex folds

Pseudoknot

Kissing Hairpins

Hair-bulge interaction
Pseudoknots

\[ i < i' < j < j' \]

\[ i < j < i' < j' \]
RNA secondary structure representation

- 2D
- Circle plot
- Dot plot
- Mountain
- Parentheses
- Tree model

(((...)))..((.....))
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**
  - using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted

- **Simultaneous folding and alignment (structural alignment)**
Assumptions in energy minimization approaches

- Most likely structure similar to energetically most stable structure
- Energy associated with any position is only influenced by local sequence and structure
- Neglect pseudoknots
Base-pair maximization

- Find structure with the most base pairs
  - Only consider A-U and G-C and do not distinguish them

- Nussinov algorithm (1970s)
  - Too simple to be accurate, but stepping-stone for later algorithms
Problem definition
- Given sequence $X = x_1x_2 \ldots x_L$, compute a structure that has maximum (weighted) number of base pairings

How can we solve this problem?
- Remember: RNA folds back to itself!
- $S(i,j)$ is the maximum score when $x_i \ldots x_j$ folds optimally
- $S(1,L)$?
- $S(i,i)$?
“Grow” from substructures

\[ S(i, j) = \max \begin{cases} 
  S(i + 1, j - 1) + w(i, j) \\
  S(i + 1, j) \\
  S(i, j - 1) \\
  \max_{i < k < j} S(i, k) + S(k + 1, j)
\end{cases} \]

\( w(i, j) = 1 \) if \( i, j \) are complementary (i.e., GC, CG, AU or UA); 0 otherwise
Dynamic programming

- Compute $S(i,j)$ recursively (dynamic programming)
  - Compares a sequence against itself in a dynamic programming matrix

- Three steps
Initialization

Example:

GGGAAAUCC

\[
S(i, i) = 0 \quad \forall \quad 1 \leq i \leq L \\
S(i, i - 1) = 0 \quad \forall \quad 2 \leq i \leq L
\]

→ the main diagonal

→ the diagonal below

\(L\): the length of input sequence
Recursion

Fill up the table (DP matrix) -- diagonal by diagonal

\[ S(i, j) = \max \begin{cases} 
  S(i + 1, j - 1) + w(i, j) & (1) \\
  S(i + 1, j) & (2) \\
  S(i, j - 1) & (3) \\
  \max_{i < k < j} S(i, k) + S(k + 1, j) & (4) 
\end{cases} \]

\[ w(i, j) = \begin{cases} 
  1 & \text{i, j are complementary} \\
  0 & \text{otherwise} 
\end{cases} \]
What are the other “optimal” structures?
An exercise

- Input: AUGACAU
- Fill up the table
- Trace back

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

- Give the optimal structure
- What’s the size of the hairpin loop
Energy minimization methods

- Nussinov algorithm (base pair maximization) is too simple to be accurate
- Energy minimization algorithm predicts secondary structure by minimizing the free energy ($\Delta G$)
- $\Delta G$ calculated as sum of individual contributions of:
  - loops
  - stacking
Free energy computation

$\Delta G = -4.6 \text{ KCAL/MOL}$
Loop parameters
(from Mfold)

DESTABILIZING ENERGIES BY SIZE OF LOOP

<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>2.8</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>3.2</td>
<td>5.4</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>3.6</td>
<td>5.6</td>
</tr>
<tr>
<td>5</td>
<td>2.1</td>
<td>4.0</td>
<td>5.7</td>
</tr>
<tr>
<td>6</td>
<td>1.9</td>
<td>4.4</td>
<td>5.4</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2.6</td>
<td>5.1</td>
<td>6.7</td>
</tr>
<tr>
<td>13</td>
<td>2.7</td>
<td>5.2</td>
<td>6.8</td>
</tr>
<tr>
<td>14</td>
<td>2.8</td>
<td>5.3</td>
<td>6.9</td>
</tr>
<tr>
<td>15</td>
<td>2.8</td>
<td>5.4</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Unit: Kcal/mol
Stacking energy
(from Vienna package)

```plaintext
# stack_energies
/* CG    GC    GU    UG    AU    UA    @  */
|    -2.0 |  -2.9 |  -1.9 |  -1.2 |  -1.7 |  -1.8 |  0 |
|    -2.9 |  -3.4 |  -2.1 |  -1.4 |  -2.1 |  -2.3 |  0 |
|    -1.9 |  -2.1 |   1.5 |  -.4  |  -1.0 |  -1.1 |  0 |
|    -1.2 |  -1.4 |  -.4  |  -.2  |  -.5  |  -.8  |  0 |
|    -1.7 |  -.2  |  -1.0 |  -.5  |  -.9  |  -.9  |  0 |
|    -1.8 |  -2.3 |  -1.1 |  -.8  |  -.9  |  -1.1 |  0 |
|     0    |     0  |     0  |     0  |     0  |     0  |  0 |
```
Mfold versus Vienna package

- **Mfold**
  - [http://frontend.bioinfo.rpi.edu/zukerm/download/](http://frontend.bioinfo.rpi.edu/zukerm/download/)
  - [http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi](http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi)
  - *Suboptimal structures*
    - The correct structure is not necessarily structure with optimal free energy
    - Within a certain threshold of the calculated minimum energy

- **Vienna** -- calculate the probability of base pairings
  - [http://www.tbi.univie.ac.at/RNA/](http://www.tbi.univie.ac.at/RNA/)
Mfold energy dot plot

Lower Triangle Shows Optimal Energy
Optimal energy: -11.9
-11.9 ≤ energy ≤ -11.6

Upper Triangle Basepairs Plotted: 97
-11.6 ≤ energy ≤ -11.2
-11.2 ≤ energy ≤ -10.9
Mfold algorithm

The main mathematical technique is to compute two possible different energies for each subsequence $S_{ij}$ of a given RNA sequence, $W(i,j)$ and $V(i,j)$ (satisfying $1 \leq i < j \leq N$). Let $W(i,j)$ be the minimum free energy of all possible admissible structures formed from $S_{ij}$, while $V(i,j)$ be the minimum free energy of all possible admissible structures formed from $S_{ij}$, in which $i$ and $j$ pair with each other. If $i$ and $j$ can not pair, then $V(i,j) = \infty$. All $V_{ij}$ and $W_{ij}$ are computed recursively.

$$V(i,j) = \min \begin{cases} E(FH(i,j)) \\ \min_{i<k<m<j} E(FL(i,j;k,m)) + V(k,m) \\ \min_{i+1<k<j-2} W(i+1,k) + W(k+1,j-1) \end{cases}$$  \hspace{1cm} (1)

$$W(i,j) = \min \begin{cases} W(i+1,j) \\ W(i,j-1) \\ V(i,j) \\ \min_{i<k<j-1}(W(i,k) + W(k+1,j)) \end{cases}$$ \hspace{1cm} (4a, 4b, 1-3, 5)

The situations are: (1-3) – $i$ and $j$ pair with each other; (1) – hairpin; (2) – interior loop (include interior loop, buldge, i.e., k=i+1 or m=j-1, and stacking, k=i+1 and m=j-1); (3) – closed bifurcation (multiple loops); (4) – at least $i$ or $j$ is unpaired; (5) – open bifurcation ($i$ and $j$ are not paired with each other)
Inferring structure by comparative sequence analysis

- Need a multiple sequence alignment as input

- Requires sequences be similar enough (so that they can be initially aligned)

- Sequences should be dissimilar enough for covarying substitutions to be detected

“Given an accurate multiple alignment, a large number of sequences, and sufficient sequence diversity, comparative analysis alone is sufficient to produce accurate structure predictions” (Gutell RR et al. Curr Opin Struct Biol 2002, 12:301-310)
RNA variations

- Variations in RNA sequence maintain base-pairing patterns for secondary structures (conserved patterns of base-pairing)

- When a nucleotide in one base changes, the base it pairs to must also change to maintain the same structure

- Such variation is referred to as *covariation*.
If neglect covariation

- In usual alignment algorithms they are doubly penalized

...GA...UC...
...GA...UC...
...GA...UC...
...GC...GC...
...GA...UA...
Covariance measurements

- Mutual information (desirable for large datasets)
  - Most common measurement
  - Used in CM (Covariance Model) for structure prediction

- Covariance score (better for small datasets)
Mutual information

\[ MI_{ij} = \sum_{x_i y_j} f_{x_i y_j} \log_2 \frac{f_{x_i y_j}}{f_{x_i} f_{x_j}} \]

- \( f_{x_i} \): frequency of a base in column \( i \)
- \( f_{x_i y_j} \): joint (pairwise) frequency of a base pair between columns \( i \) and \( j \)
- Information ranges from 0 and \( ? \) bits
- If \( i \) and \( j \) are uncorrelated (independent), mutual information is 0
Mutual information plot
Structure prediction using MI

- $S(i, j) = \text{Score at indices } i \text{ and } j; M(i, j) \text{ is the mutual information between } i \text{ and } j$
- The goal is to maximize the total mutual information of input RNA
- The recursion is just like the one in Nussinov Algorithm, just to replace $w(i, j)$ (1 or 0) with the mutual information $M(i, j)$

$$S(i, j) = \max \begin{cases} 
S(i + 1, j - 1) + M(i, j) \\
S(i + 1, j) \\
S(i, j - 1) \\
\max_{i < k < j} S(i, k) + S(k + 1, j) 
\end{cases}$$
Covariance-like score

- RNAalifold
- Desirable for small datasets
- Combination of covariance score and thermodynamics energy
Covariance-like score calculation

The score between two columns $i$ and $j$ of an input multiple alignment is computed as following:

$$C_{ij} = \frac{1}{\binom{N}{2}} \sum_{\alpha < \beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^{\alpha} \Pi_{ij}^{\beta} = \sum_{XY, X'Y'} f_{ij}(XY) D_{XY, X'Y'} f_{ij}(X'Y')$$

$$d_{ij}^{\alpha,\beta} = 2 - \delta(a_i^\alpha, a_i^\beta) - \delta(a_j^\alpha, a_j^\beta)$$

$N$ is the number of sequences in the alignment; $\alpha$ and $\beta$ are two sequences; $B=$\{GC, CG, AU, UA, GU, UG\} is the set of allowed base pairs; $\Pi$ is a pairing matrix with $\Pi_{ij}=1$ if $i$ and $j$ can form a base pair (i.e., $(i,j) \in B$), otherwise 0; $\delta(a_i^\alpha, b_i^\beta)$ is 1 if $a_i^\alpha = a_i^\beta$, otherwise 0; $D$ is $16 \times 16$ matrix with entries $D_{XY, X'Y'} = d_H(XY, X'Y')$ if both $XY \in B$ and $X'Y' \in B$ and $D_{XY, X'Y'} = 0$, otherwise. $d_H(XY, X'Y')$ is again the Hamming distance of $XY$ and $X'Y'$. 
Covariance model

- A formal covariance model, CM, devised by Eddy and Durbin
  - A probabilistic model
  - $\approx$ A Stochastic Context-Free Grammer
  - Generalized HMM model

- A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus

- Provides very accurate results

- Very slow and unsuitable for searching large genomes
CM training algorithm

1. Unaligned sequence

2. Multiple alignment

3. Covariance model

4. EM

5. Parameter re-estimation

6. Modeling construction
Binary tree representation of RNA secondary structure

- Representation of RNA structure using Binary tree
- Nodes represent
  - Base pair if two bases are shown
  - Loop if base and “gap” (dash) are shown
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions

Images – Eddy et al.
Overall CM architecture

MATP emits pairs of bases: modeling of base pairing

BIF allows multiple helices (bifurcation)
Covariance model drawbacks

- Needs to be well trained (large datasets)
- Not suitable for searches of large RNA
  - Structural complexity of large RNA cannot be modeled
  - Runtime
  - Memory requirements
ncRNA gene finding

- *De novo* ncRNA gene finding
  - Folding energy
  - Number of sub-optimal RNA structures

- Homology ncRNA gene searching
  - Sequence-based
  - Structure-based
  - Sequence and structure-based
Rfam & Infernal

- Rfam 9.1 contains 1379 families (December 2008)
- Rfam 10.0 contains 1446 families (January 2010)
- Rfam is a collection of **multiple sequence alignments and covariance models** covering many common non-coding RNA families
- Infernal searches Rfam covariance models (CMs) in genomes or other DNA sequence databases for homologs to known structural RNA families

http://rfam.janelia.org/
An example of Rfam families

- TPP (a riboswitch; THI element)
  - RF00059
  - is a riboswitch that directly binds to TPP (active form of VB, thiamin pyrophosphate) to regulate gene expression through a variety of mechanisms in archaea, bacteria and eukaryotes
Simultaneous structure prediction and alignment of ncRNAs

The grammar emits two correlated sequences, x and y
References

- A computational pipeline for high throughput discovery of cis-regulatory noncoding RNAs in Bacteria, PLoS CB 3(7):e126
Understanding the transcriptome through RNA structure

- 'RNA structurome’
- Genome-wide measurements of RNA structure by high-throughput sequencing