COMP 598: Variations on RNA secondary structure prediction

Jérôme Waldispühl
School of Computer Science
McGill University
Circular RNAs (circRNAs)

- Large ratio of transcripts
- Mostly located in cytoplasm

(Salzman et al., 2012)
(putative) Functions

• Sponge up Micro RNA to reduce silencing effect
• Binding proteins to regulate RBP & RNA interactions
• Micro RNA transport
• Bind and regulate messenger RNA

Viroids
• Smallest infectious pathogens known
• Mostly found in plants
• Do not encode proteins
• Affect host through RNA silencing

(Jakub Friedl, 2007)
circRNA structure prediction

\[ M(i, j) = \min \begin{cases} M(i, j - 1) \\ \min_{i \leq k < j} \{ M(i, k - 1) + M(k + 1, j - 1) + \delta(k, j) \} \end{cases} \]

It does not consider circular base pairs \((i, j)\) such that \(i > j\)...

Solution 1 (Zuker, 1984):
• Duplicate input sequence
• Find \(\min_{i < j}\{ M(i, j) + M(j, i + n) \} \)

Solution 2 (Hofacker & Stadler, 2006):
Origin of covariations

(Rivas et al., 2017)
Significant RNA base pairs (R-scape)
Locally stable structures (RNAplfold)

• Consider only base pair (i,j) such that j-i < \( \eta \)
• Average probability over all overlapping window
• RNAplfold (Bernhart et al., 2006)
Stable structures in mRNAs (SPARCS)

Zhang et al., 2013
Non-coding RNA detection (RNAz)

(Preprocessed) multiple sequence alignment

Extract single sequences and remove all gaps

Use RNAfold for minimum free energy ($E_f$) calculation of each sequence

Use RNAfold for consensus free energy calculation of the alignment ($E_{consensus}$)

Calculate the structure conservation index

$SCI = \frac{E_{consensus}}{E_f}$

Calculate normalized Shannon entropy of the alignment

$H = -\frac{1}{I} \sum_{z} p_z \log_2 p_z$

Calculate average z-score

$z = \frac{1}{n} \sum_{z \in A} \frac{E_{consensus} - E_f}{\sigma}$

Was the input a structural alignment?

SVM trained on LocARNAITE alignments

SVM trained on CluetaIV alignments

Estimate mean ($\mu_z$) and standard deviation ($\sigma$) by an explicit sequence sampling procedure

Classification probability

(gruber et al., 2010)
RNA folding dynamics

From slides from P. Schuster & I. Hofacker
Thermodynamic vs. Kinetic Folding

Equilibrium properties can be calculated efficiently

But what about dynamics?

• On what time scale is equilibrium reached?
• How fast/slow is re-folding between dissimilar structures?
• What structures are populated initially?

\[
\begin{align*}
\text{GUCCG} & \quad \iff \\
\text{GUCGG} & \quad \text{GUCC} \\
-7.9 \text{kcal/mol} & \quad \iff \quad -8.0 \text{kcal/mol}
\end{align*}
\]
RNA switches toggle between active and inactive states by changing conformation. They are commonly used to control mRNA translations. They can be triggered by:

- binding of proteins or small ligands
- chemical modification, e.g. tRNA
- temperature dependent switches
- timed mRNA switches, e.g. HOK
Examples of RNA switches

A Ribozyme with two functions (Schultes & Bartel 2000)

Chemical modification triggers the cloverleaf fold of a tRNA (Helm & Giegé 1999)
Predicting dynamics of RNA folding

Folding dynamics described by a Markov process with master equation:

\[
\frac{dp_x}{dt} = \sum_{y \in X} r_{xy} p_y(t), \quad \text{with} \quad r_{xx} = -\sum_{y \neq x} r_{yx}
\]

- Integration of the master equation (toy models only).
- Qualitative analysis of the energy landscape to identify possible traps (local minima). It enables to design coarse grained versions of the Markov process.

Need to model the rate \( r_{xy} \). For small moves Metropolis rule is sufficient.
Elementary move set

Add or remove base pairs
Class 1: Shift inside internal loops or bulges
Elementary move set

Class 2: Shift involving free ends
Kinetic folding algorithm

Simulate folding kinetics by a Monte-Carlo type algorithm:

Generate all neighbors using the move-set
- Basepair Insertion
- Basepair Deletion

Assign rates to each move, e.g.

\[ P_i = \min \left\{ 1, \exp \left( -\frac{\Delta E}{kT} \right) \right\} \]

Advance clock \( 1/\sum_i P_i \).
select a move with probability proportional to its rate
Barrier tree
Calculating barrier trees

The flooding algorithm:

Read conformations in energy sorted order.
For each confirmation $x$ we have three cases:

(a) $x$ is a local minimum if it has no neighbors we’ve already seen

(b) $x$ belongs to basin $B(s)$, if all known neighbors belong to $B(s)$

(c) if $x$ has neighbors in several basins $B(s_1) \ldots B(s_k)$ then it’s a saddle point that merges these basins. Basins $B(s_1), \ldots, B(s_k)$ are then united and are assigned to the deepest of local minimum.
Energy landscape of a toy sequence
Energy landscape of a designed RNA
Refolding path

- The two component structure is kinetically preferred, because both hairpins act as nucleation centers.
- For the full length chain 75% of trajectories reach the two component structure first.
Refolding path

The folding path from $S_1$ to $S_0$
Further...

- Implemented in Vienna RNA package as kinfold
- Coarse grained version for larger RNAs
- Used to model co-transcriptional folding