RNA Secondary Structure Prediction

02-710 Computational Genomics
Seyoung Kim
Outline

• RNA folding
• Dynamic programming for RNA secondary structure prediction
• Covariance model for RNA structure prediction
RNA Basics

- RNA bases A, C, G, U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U
  "wobble" pairing
  - Bases can only pair with one other base.

DNA → RNA → Protein
RNA Basics

- transfer RNA (tRNA)
- messenger RNA (mRNA)
- ribosomal RNA (rRNA)
- small interfering RNA (siRNA)
- micro RNA (miRNA)
- small nucleolar RNA (snoRNA)

http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
Pseudoknots

• Pseudoknots: a nucleic acid secondary structure containing at least two stem-loop structures which half of one stem is intercalated between the two halves of another stem.
Sequence Alignment as a method to determine structure

- Bases pair in order to form backbones and determine the secondary structure
- Aligning bases based on their ability to pair with each other gives an algorithmic approach to determining the optimal structure
Base Pair Maximization – Dynamic Programming Algorithm

Simple Example: Maximizing Base Pairing

\[ S(i,j) = \max \begin{cases} 
S(i + 1, j - 1) + 1 & \text{[if } i, j \text{ base pair]} \\
S(i + 1, j) \\
S(i, j - 1) \\
\max_{i < k < j} S(i, k) + S(k + 1, j) 
\end{cases} \]

\( S(i,j) \) is the folding of the subsequence of the RNA strand from index \( i \) to index \( j \) which results in the highest number of base pairs.
Base Pair Maximization – Dynamic Programming Algorithm

Simple Example: Maximizing Base Pairing

\[
S(i,j) = \max \begin{cases} 
S(i+1,j-1) + 1 & \text{[if } i,j \text{ base pair]} \\
S(i+1,j) \\
S(i,j-1) \\
\max_{i<k<j} S(i,k) + S(k+1,j) 
\end{cases}
\]

Base pair at \(i\) and \(j\)
Base Pair Maximization – Dynamic Programming Algorithm

Simple Example: Maximizing Base Pairing

\[ S(i,j) = \max \begin{cases} 
S(i + 1, j - 1) + 1 & \text{[if } i,j \text{ base pair]} \\
S(i + 1, j) & \\
S(i, j - 1) & \\
\max_{i < k < j} S(i, k) + S(k + 1, j) & 
\end{cases} \]

Unmatched at i
Base Pair Maximization – Dynamic Programming Algorithm

Simple Example: Maximizing Base Pairing

\[
S(i,j) = \max \begin{cases} 
S(i + 1, j - 1) + 1 & \text{[if } i, j \text{ base pair]} \\
S(i + 1, j) \\
S(i, j - 1) \\
\max_{i < k < j} S(i, k) + S(k + 1, j)
\end{cases}
\]

Umatched at \( j \)
Base Pair Maximization – Dynamic Programming Algorithm

Simple Example: Maximizing Base Pairing

\[ S(i, j) = \max \begin{cases} 
S(i + 1, j - 1) + 1 & \text{[if } i, j \text{ base pair]} \\
S(i + 1, j) & \\
S(i, j - 1) & \\
\max_{i < k < j} S(i, k) + S(k + 1, j) 
\end{cases} \]

Bifurcation
Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs

- Each score independent of overall structure

- Bifurcation adds extra dimension
Base Pair Maximization – Dynamic Programming Algorithm

- **Alignment Method**
  - Align RNA strand to itself
  - Score increases for feasible base pairs

- Each score independent of overall structure

- Bifurcation adds extra dimension

Initialize first two diagonal arrays to 0
Base Pair Maximization – Dynamic Programming Algorithm

- **Alignment Method**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Fill in squares sweeping diagonally
Base Pair Maximization – Dynamic Programming Algorithm

- **Alignment Method**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- **Each score independent of overall structure**
- **Bifurcation adds extra dimension**

*Bases cannot pair, similar to unmatched alignment*
Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bases can pair, similar to matched alignment

Images – Sean Eddy
Base Pair Maximization – Dynamic Programming Algorithm

- **Alignment Method**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Dynamic Programming – possible paths

```
S(i + 1, j - 1) + 1
```
Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Dynamic Programming – possible paths

Images – Sean Eddy
Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

**Dynamic Programming – possible paths**
Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bifurcation – add values for all k

\[ k = 0 : \text{Bifurcation max in this case} \]

\[ S(i,k) + S(k + 1, j) \]
Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Reminder:
For all $k$

$$S(i,k) + S(k+1, j)$$
Base Pair Maximization – Dynamic Programming Algorithm

- **Alignment Method**
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- **Each score independent of overall structure**
- **Bifurcation adds extra dimension**

**Reminder:** For all $k$

$$S(i, k) + S(k + 1, j)$$

Images – Sean Eddy
**Base Pair Maximization - Drawbacks**

- Base pair maximization will not necessarily lead to the most stable structure
  - May create structure with many interior loops or hairpins which are energetically unfavorable
- Comparable to aligning sequences with scattered matches – not biologically reasonable
Energy Minimization

- Thermodynamic Stability
  - Estimated using experimental techniques
  - Theory: Most Stable is the Most likely
- No Pseudoknots due to algorithm limitations
- Uses Dynamic Programming alignment technique
- Attempts to maximize the score taking into account thermodynamics
- MFOLD and ViennaRNA
Energy Minimization Results

- Linear RNA strand folded back on itself to create secondary structure
- Circularized representation uses this requirement
  - Arcs represent base pairing

Images – David Mount
Energy Minimization Results

- All loops must have at least 3 bases in them
  Equivalent to having 3 base pairs between all arcs

**Exception:** Location where the beginning and end of RNA come together in circularized representation

Images – David Mount
Trouble with Pseudoknots

- Pseudoknots cause a breakdown in the Dynamic Programming Algorithm.
- In order to form a pseudoknot, checks must be made to ensure base is not already paired – this breaks down the recurrence relations.
Energy Minimization Drawbacks

• Compute only one optimal structure
• Usual drawbacks of purely mathematical approaches
  – Similar difficulties in other algorithms
    • Protein structure
    • Exon finding
Alternative Algorithms - Covariation

• Incorporates Similarity-based method
  – Evolution maintains sequences that are important
  – Change in sequence coincides to maintain structure through base pairs (Covariance)
    • Cross-species structure conservation example – tRNA

• Manual and automated approaches have been used to identify covarying base pairs

• Models for structure based on results
  – Ordered Tree Model
  – Stochastic Context Free Grammar
Alternative Algorithms - Covariation

Expect areas of base pairing in tRNA to be covarying between various species.
Base pairing creates same stable tRNA structure in organisms
Alternative Algorithms - Covariation

Mutation in one base yields pairing impossible and breaks down structure
Alternative Algorithms - Covariation

Covariation ensures ability to base pair is maintained and RNA structure is conserved.
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
- Traverse root to leaves, from left to right
- Pseudoknots still not represented
- Tree does not permit varying sequences
  - Mismatches
  - Insertions & Deletions

Images – Eddy et al.
Covariance Model

• HMM which permits flexible alignment to an RNA structure –
  – emission and transition probabilities
• Model trees based on finite number of states
  – Match states – sequence conforms to the model:
    • MATP – State in which bases are paired in the model and sequence
    • MATL & MATR – State in which either right or left bulges in the sequence and the model
  – Deletion – State in which there is deletion in the sequence when compared to the model
  – Insertion – State in which there is an insertion relative to model
• Transitions have probabilities
  – Varying probability – Enter insertion, remain in current state, etc
  – Bifurcation – no probability, describes path
Alignment to CM Algorithm

- Calculate the probability score of aligning RNA to CM
- Three dimensional matrix – $O(n^3)$
  - Align sequence to given subtrees in CM
  - For each subsequence calculate all possible states
- Subtrees evolve from Bifurcations
  - For simplicity Left singlet is default

Images – Eddy et al.
Alignment to CM Algorithm

For each calculation take into account the
- Transition (T) to next state
- Emission probability (P) in the state as determined by training data

\[ S_{i,j,y}(y = MATP) = \max_{y_{\text{next}}} [S_{i+1,j-1,y_{\text{next}}} + \log T(y_{\text{next}} | y) + \log P(x_i, x_j | y)] \]
Alignment to CM Algorithm

Images – Eddy et al.

• For each calculation take into account the
  • Transition (T) to next state
  • Emission probability (P) in the state as determined by training data

\[
S_{i,j,y}(y = MATR, INSR) = \max_{y_{next}} \left[ S_{i,j-1,y_{next}} + \log T(y_{next} \mid y) + \log P(x_j \mid y) \right]
\]
Alignment to CM Algorithm

For each calculation take into account the

• Transition (T) to next state
• Emission probability (P) in the state as determined by training data

\[
S_{i,j,w}(y = MATR, INSR) = \max_{y_{next}} [S_{i,j-1,w_{next}} + \log T(y_{next} \mid y) + \log P(x_j \mid y)]
\]
Alignment to CM Algorithm

For each calculation take into account the
- Transition (T) to next state
- Emission probability (P) in the state as determined by training data

Deletion – does not have an emission probability (P) associated with it

\[ S_{i,j,y}(y = DEL) = \max_{y_{next}} [S_{i,j,y_{next}} + \log \mathcal{T}(y_{next} | y)] \]
Alignment to CM Algorithm

- For each calculation take into account the
  - Transition (T) to next state
  - Emission probability (P) in the state as determined by training data

Images – Eddy et al.

\[ \text{Si,j,y}(y = BIFURC) = \max_{i-1<=mid<=j} \left[ S_{i,mid,y,left} + S_{mid+1,j,y,right} \right] \]
Model Training

unaligned sequences

random alignment

multiple alignment

alignment

(EM)

parameter reestimation

covariance model

model construction (structure prediction)
Covariance Model (CM) Training Algorithm

- \( S(i,j) = \text{Score at indices } i \text{ and } j \text{ in RNA when aligned to the Covariance Model} \)

\[
S(i,j) = \max \left\{ \begin{array}{l}
S(i+1,j-1) + M(i,j) \\
S(i+1,j) \\
S(i,j-1) \\
\max_{i<k<j} S(i,k) + S(k+1,j)
\end{array} \right.
\]

- \( M_{i,j} = \sum_{x_i,x_j} f_{x_i,x_j} \log_2 \frac{f_{x_i,x_j}}{f_{x_i}f_{x_j}} \)

  Frequency of seeing the symbols (A, C, G, T) together in locations \( i \) and \( j \) depending on symbol.

  Independent frequency of seeing the symbols (A, C, G, T) in locations \( i \) or \( j \) depending on symbol.

- Frequencies obtained by aligning model to "training data" – consists of sample sequences
  - Reflect values which optimize alignment of sequences to model
Mutual information for RNA Secondary Structure Prediction
Covariance Model Drawbacks

• Needs to be well trained
• Not suitable for searches of large RNA
  – Structural complexity of large RNA cannot be modeled
  – Runtime
  – Memory requirements
References