RNA Secondary Structure - Biological Functions & Prediction
Hidden Markov Models of Genomic & Protein Features

- Hidden Markov Model terminology
- Viterbi algorithm
- Examples
  - CpG Island HMM
  - TMHMM (transmembrane helices)
"Trellis" Diagram for Viterbi Algorithm

<table>
<thead>
<tr>
<th>Position in Sequence</th>
<th>1</th>
<th>…</th>
<th>i-2</th>
<th>i-1</th>
<th>i</th>
<th>i+1</th>
<th>i+2</th>
<th>…</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden States</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Full set of possible transitions from position i to i+1
Rabiner notation

\[ P_{gg} = 0.99999 \quad P_{ig} = 0.001 \]
\[ P_{ii} = 0.999 \]
\[ P_{gi} = 0.00001 \]

CpG Island HMM

\[ \pi \]

Emission Probabilities \( b_{j(k)} \)

Transition probabilities \( a_{ij} \)

Initiation probabilities \( \pi_j \)

Genome

CpG Island:

\[ \text{C: 0.3, G: 0.3, A: 0.2, T: 0.2} \]

Genome:

\[ \text{A: 0.2, C: 0.2, G: 0.3, T: 0.3} \]
More Viterbi Examples

What is the optimal parse of the sequence for the CpG island HMM defined previously?

• $(ACGT)_{10000}$

• $A_{1000}C_{80}T_{1000}C_{20}A_{1000}G_{60}T_{1000}$

Powers of 1.5:

\[
\begin{align*}
N &= 20 & 40 & 60 & 80 \\
(1.5)^N &= 3 \times 10^3 & 1 \times 10^7 & 3 \times 10^{10} & 1 \times 10^{14}
\end{align*}
\]
Real World HMMs
“Profile HMM” with insertions/deletions

Of course, can have insertion/deletion states for HMM models of DNA/RNA as well.

© Cold Spring Harbor Laboratory Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
TMHMM (v. 2.0)

Prediction of transmembrane helices in proteins

Help/Information (updated Sept 13, 2001)

One of the World Wide Web Prediction Servers from the Center for Biological Sequence Analysis

Correctly predicts ~97% of transmembrane helices according to authors


© Center for Biological Sequence Analysis. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Architecture of TMHMM

(a)

(c)

Courtesy of Biomedical Informatics Publishing Group. Used with permission.
Optimal Parse

TMHMM Output for Mouse Chloride Channel CLC6

- Posterior Probability

- Transmembrane
- Inside
- Outside

Page number: 10
RNA Secondary Structure

• Biological examples of RNA structure
• Predicting $2^\circ$ structure by covariation
• Predicting $2^\circ$ structure by energy minimization

Readings
NBT Primer on RNA folding, Z&B Ch. 11.9
RNA Secondary and Tertiary Structure

Example: tRNA

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
RNA Secondary Structure Notation

Parentheses notation

..((((.....))))......(((((..................))))))...

Arc ('rainbow') notation

What do these structures look like?

What is the difference between these two structures?
Ribosome at 7 Å with tRNAs

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/


Slide courtesy of Rachel Green
Can build useful structures out of RNA

The exit channel for the growing polypeptide

Slide courtesy of Rachel Green

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/

RNA/protein distribution on the 50S ribosome

linguini = protein

fettucini = RNA

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/. Source: Ban, Nenad, Poul Nissen, et al. "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution." Science 289, no. 5481 (2000): 905-20.
The ribosome is a ribozyme

Nearest proteins and distances to active site (Å)

Slide courtesy of Rachel Green
What are the practical applications of knowing the ribosome structure?

Antibiotics!
ncRNAs: Challenges for Computational Biology

- Prediction of ncRNA structure
- Identification of ncRNA genes
- Prediction of ncRNA functions
RNA 2° structure by covariation / compensatory changes

Seq1: A C G A A A G U
Seq2: U A G U A A U A
Seq3: A G G U G A C U
Seq4: C G G C A A U G
Seq5: G U G G G A A C

Diagram of RNA structure.
Mutual information statistic for pair of columns in a multiple alignment

\[ M_{ij} = \sum_{x,y} f_{x,y}^{(i,j)} \log_2 \frac{f_{x,y}^{(i,j)}}{f_x^{(i)} f_y^{(j)}} \]

- \( f_{x,y}^{(i,j)} \) = fraction of seqs w/ nt. \( x \) in col. \( i \), nt. \( y \) in col. \( j \)
- \( f_x^{(i)} \) = fraction of seqs w/ nt. \( x \) in col. \( i \)

Sum over \( x, y = A, C, G, U \)

\( M_{ij} \) is maximal (2 bits) if \( x \) and \( y \) individually appear at random (A,C,G,U equally likely), but perfectly covary (e.g., always complementary)

Could use other measure of dependence (e.g., chi-square statistic)
Inferring 2° structure from covariation
What is needed for accurate inference of RNA secondary structure by covariation?

• Secondary structure more highly conserved than primary sequence

• Sufficient divergence between homologs for many variations to have occurred, but not so much that can’t be aligned

• Sufficient number of homologs sequenced
Classes of Non-coding RNAs

- tRNAs
- rRNAs
- UTRs
- snRNAs
- snoRNAs
- prok. terminators
- RNaseP
- SRP RNA
- tmRNA
- miRNAs
- lncRNAs
- riboswitches
Energy Minimization Approach

\[ \Delta G_{\text{folding}} = G_{\text{unfolded}} - G_{\text{folded}} \]

There are typically many possible folded states
- assumption that minimum energy state(s) will be occupied

\[ \Delta G = \Delta H - T \Delta S \]

Enthalpy favors folding
Entropy favors unfolding

What environmental variables affect RNA folding?
How Do Energy Minimization Algorithms Work?

Consider Simple Model: Base Pair Maximization

**Scoring System:**

+1 for base pair (C:G, A:U)
0 for anything else

Maximizing score equivalent to minimizing folding free energy for a model which assigns same enthalpy to all allowed base pairs (and ignores details such as base stacking, loops, entropy)

Nussinov algorithm: recursive maximization of base pairing
Recursive Maximization of Base Pairing

Given an RNA sequence of length N

Define $S(i,j)$ to be the score of the best structure for the subsequence $(i, j)$

Notice that $S(i,j)$ can be defined recursively in terms of optimal scores of smaller subsequences of the interval $(i,j)$

There are four possible ways that the score of the optimal structure on $(i,j)$ can relate to scores of optimal structures of nested subsequences:

1. $i,j$ pair
2. $i$ unpaired
3. $j$ unpaired
4. bifurcation

Base Pair Maximization Algorithm

\[ S(i,j) = \text{score of the optimal structure for the subsequence (i, j)} \]

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

1) Initialize an \(N \times N\) matrix \(S\) with \(S(i,i) = S(i,i-1) = 0\)

2) Fill in \(S(i,j)\) matrix recursively from the diagonal up and to the right (keep track of which choice was made at each step)

3) Trace back from \(S(1,N)\) (upper right corner of matrix) to diagonal to determine optimal structure
Dynamic Programming for Base Pair Maximization

Recursive definition of the best score for a sub-sequence \(i,j\) looks at four possibilities:

1. \(i,j\) pair
2. \(i\) unpaired
3. \(j\) unpaired
4. Bifurcation

Dynamic programming algorithm for all sub-sequences \(i,j\), from smallest to largest:

Initialization:

Courtesy of Macmillan Publishers Limited. Used with permission.
Base Pair Maximization Algorithm Issues

• What is computational complexity of algorithm? (for sequence of length $N$)

Answer: Memory - $O(N^2)$  Time - $O(N^3)$

• Can it handle pseudoknots?

Answer: No. Pseudoknots invalidate recursion for $S(i,j)$
Viral Pseudoknots and “Kissing loops”

Baranov et al. Virology 2005

RNA Energetics I

Free energy contributions to helix formation come from:

• base pairing:

\[
\begin{array}{ccc}
G & > & A \\
\uparrow & & \uparrow \\
C & > & U \\
\end{array}
\]

• base stacking:

\[
\begin{array}{ccc}
G_p & A \\
\leftrightarrow & \\
C_p & U \\
\end{array}
\]

Base stacking contributes more to free energy than base pairing.

© American Chemical Society. All rights reserved. This content is excluded from our Creative Commons license. For more information, see [http://ocw.mit.edu/help/faq-fair-use/](http://ocw.mit.edu/help/faq-fair-use/).

RNA Energetics I

Free energy contributions from:

- **base pairing:**
  \[
  \begin{array}{ccc}
  G & > & A \\
  C & > & U
  \end{array}
  \]

- **base stacking:**
  \[
  \begin{array}{ccc}
  G_p & A \\
  C_p & U
  \end{array}
  \]

are combined in Doug Turner’s Energy Rules:

Matrix for each X,Y stacking on each possibly base pair or free end

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>-1.30</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>-2.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-2.10</td>
<td></td>
<td>-1.00</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>-0.90</td>
<td></td>
<td>-1.30</td>
<td></td>
</tr>
</tbody>
</table>
RNA Energetics II

Other Contributions to Folding Free Energy

• Hairpin loop destabilizing energies
  - a function of loop length

• Interior and bulge loop destabilizing energies
  - a function of loop length

• Terminal mismatch and base pair energies
A more complex dynamic programming algorithm is used - similar in spirit to the Nussinov base pair maximization algorithm

Gives:

- minimum energy fold
- suboptimal folds (e.g., five lowest ΔG folds)
- probabilities of particular base pairs
- full partition function

Accuracy: ~70% of base pairs correct
Links & References

The Mfold web server:
http://mfold.rna.albany.edu/?q=mfold/rna-folding-form

The Vienna RNAfold package (free for download)
http://www.tbi.univie.ac.at/~ivo/RNA/

RNA folding references:


Vienna package by Ivo Hofacker
RNA Secondary Structure Prediction by Energy Minimization Summary

- Assumes folding energy decomposable into independent contributions of small units of structure
- Algorithms are guaranteed to find minimal free energy structure defined by the model
- In practice, algorithms predict ~70% of bp correct
- Errors result from
  - imprecision of the model/parameters
  - differences between *in vitro* and *in vivo* conditions
  - *in vivo* structure may not always have minimum free energy
Sample Mfold Output (Human U5 snRNA)

Minimum free energy structure

dG = -34.6 kcal/mol

Energy dot plot
Energy dot plot for a lysine riboswitch
Lysine interacts with the junctional core of the riboswitch and is specifically recognized through shape-complementarity within the elongated binding pocket and through several direct and K+‐mediated hydrogen bonds to its charged ends.

Controls expression of enzymes involved in biosynthesis and transport of lysine
