CSE 527
Lecture 21, 12/8/03
RNA Secondary Structure Prediction
Outline

- What is it
- How is it Represented
- Why is it important
- Examples
- Approaches
RNA Structure

• Primary Structure: Sequence

• Secondary Structure: Pairing

• Tertiary Structure: 3D shape
A tRNA 3d Structure
tRNA - Alt. Representations

Figure 1: a) The spatial structure of the phenylalanine tRNA form yeast

b) The secondary structure extracts the most important information about the structure, namely the pattern of base pairings.
A “Mountain” diagram

**Figure 3:** Mountain representation of the tRNA secondary structure shown in Figure 1. The three plateaus correspond to the three hairpin loops of the clover leaf structure.
Importance

• Ribozymes (RNA Enzymes)
• Retroviruses
• Effects on transcription, translation, splicing...
• Functional RNAs: rRNA, tRNA, snRNA, snoRNA, micro RNA, RNAi, .......
Secondary Structure of Several HIV strains
“5-finger” motif in RRE region of many HIV strains

**Figure 11:** Alignment of the RRE regions of 17 sequences based solely on the minimum free energy secondary structure. The mountain representation reveals the five-fingered motif, the Roman numerals correspond to the numbering of the hairpins in (Dayton et al., 1989). 5 out of 22 sequences showed a different pattern here. We find three different folding patterns each highlighted by one example. The first one (thick black line) corresponds to the consensus five-fingered motif that is presented in (Konings, 1992). The second one (light gray) is present among other in HIVLAI. It shown in (Huynen et al., 1995) that this structure has high structural versatility; one of its alternatives
RNA Pairing

• Watson-Crick Pairing
  • C - G ~ 3 kcal/mole
  • A - U ~ 2 kcal/mole
  • “Wobble Pair” G - U ~ 1 kcal/mole
• Non-canonical Pairs (esp. if modified)
Definitions

- Sequence $5' r1 r2 r3 ... rn 3'$ in \{A, C, G, T\}
- A **Secondary Structure** is a set of pairs $i \cdot j$ s.t.
  1. $i < j - 4$
  2. if $i \cdot j$ & $i' \cdot j'$ are two pairs with $i \leq i'$, then
     A. $i = i' \& j = j'$, or
     B. $j < i'$, or
     C. $i < i' < j' < j$

First pair precedes 2nd, or is nested within it. No “pseudoknots.”
Nested Precedes Pseudoknot
The corresponding secondary structure is:

\[
\begin{align*}
3' & - A \quad G \quad G \quad C & - U/ & U \\
U & - C \quad C \quad G \quad A & - G \quad G \quad G \\
U & \quad C \quad C \quad C & - 5' \\
C & - U & - C/
\end{align*}
\]

A Pseudoknot
Approaches to Structure Prediction

- Maximum Pairing
  + simple
  - too inaccurate

- Minimum Energy
  + Works on single sequences
    - Ignores pseudoknots
    - Only finds “optimal” fold

- Partition Function
  + Finds all folds
    - Ignores pseudoknots
Approaches, II

• Comparative sequence analysis
  + handles all pairings (incl. pseudoknots)
  - requires several (many?) aligned

• Stochastic Context-free Grammars
  Roughly combines min energy & comparative, but no pseudoknots

• Physical experiments (x-ray crystallography, NMR)
Nussinov: Max Pairing

- $B(i,j) = \#$ pairs in optimal pairing of $r_i \ldots r_j$
- $B(i,j) = 0$ for all $i, j$ with $i \geq j-4$; otherwise
- $B(i,j) = \max$ of:
  1. $B(i+1,j)$
  2. $B(i,j-1)$
  3. $B(i+1,j-1) + (\text{if } r_i \text{ pairs with } r_j \text{ then } 1 \text{ else } 0)$
  4. $\max \{ B(i,k)+B(k+1,j) \mid i < k < j \}$

Time: $O(n^3)$
Pair-based Energy Minimization

- $E(i,j) =$ energy of pairs in optimal pairing of $r_i \ldots r_j$
- $E(i,j) = \infty$ for all $i, j$ with $i \geq j-4$; otherwise
- $E(i,j) = \min$ of:
  1. $E(i+1,j)$
  2. $E(i,j-1)$
  3. $E(i+1,j-1) + e(r_i,r_j)$
  4. $\min \{ E(i,k) + E(k+1,j) | i < k < j \}$

Time: $O(n^3)$
Loop-based Energy Minimization

- Detailed experiments show that it’s more accurate to model based on loops, rather than just pairs

- Loop types
  - Stack
  - Hairpin loop
  - Bulge
  - Interior loop
Loop Examples
Zuker: Loop-based Energy, I

- $W(i,j) = \text{energy of optimal pairing of } r_i \ldots r_j$
- $V(i,j) = \text{as above, but forcing pair } i \cdot j$
- $W(i,j) = V(i,j) = \infty \text{ for all } i, j \text{ with } i \geq j - 4$
- $W(i,j) = \min(W(i+1,j), W(i,j-1), V(i+1,j-1), \min \{ E(i,k) + E(k+1,j) | i < k < j \})$
Zuker: Loop-based Energy, II

\[ V(i,j) = \min(\text{eh}(i,j), \text{es}(i,j)+V(i+1,j-1), VBI(i,j), VM(i,j)) \]

\[ VM(i,j) = \min \{ W(i,k)+W(k+1,j) \mid i < k < j \} \]

\[ VBI(i,j) = \min \{ \text{ebi}(i,j,i',j') + V(i',j') \mid \]
\[ i < i' < j' < j \& i'-i+j-j' > 2 \} \]

Time: \( O(n^4) \)

\( O(n^3) \) possible if \( \text{ebi}(.) \) is “nice”
Suboptimal Energy

- There are always alternate folds with near-optimal energies. Thermodynamics predicts that populations of identical molecules will exist in different folds; individual molecules even flicker among different folds.
- Zuker’s algorithm can be modified to find suboptimal folds.
- McCaskill gives a more elaborate dynamic programming algorithm calculating the “partition function,” which defines the probability distribution over all these states.
Two competing secondary structures for the Leptomonas collosoma spliced leader mRNA.
Example of suboptimal folding

Black dots: pairs in opt fold

Colored dots: pairs in folds 2-5% worse than optimal fold