Outline

• Nucleic Acids Basics
• The Problem of Predicting Nucleic Acid Structure
• Thermodynamics and Phylogeny Comparison
• A commonly used program for predicting RNA secondary structure---mFOLD
• Predicting RNA Tertiary Structures by Mc-Sym

Biological Functions of Nucleic Acids

• DNA → transcription → mRNA → translation → Protein

<table>
<thead>
<tr>
<th>N</th>
<th>A Base</th>
<th>Ribose Sugar</th>
<th>Phosphate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide</td>
<td>A Base + A Ribose Sugar + A Phosphate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• A Base Can be One of the Five Rings (next):
Nucleic Acid Bases

- Pyrimidines
  ![Pyrimidines](image)
- Purines
  ![Purines](image)

- Pyrimidines and Purines Can Base-Pair (Watson-Crick Pairs)
  ![Base-Pairing](image)

Nucleic Acids As Heteropolymers

- Nucleosides, Nucleotides
  ![Nucleosides](image)
- Single Stranded DNA
  ![Single Stranded DNA](image)

- A single stranded RNA will have OH groups at the 2' positions.
- Note the directionality of DNA or RNA.

Structure Overview of Nucleic Acids

- Unlike three dimensional structures of proteins, DNA molecules assume simple double helical structures independent on their sequences. There are three kinds of double helices that have been observed in DNA: type A, type B, and type Z, which differ in their geometries. The double helical structure is essential to the coding functional of DNA. Watson (biologist) and Crick (physicist) first discovered double helix structure in 1953 by X-ray crystallography.

- RNA, on the other hand, can have as diverse structures as proteins, as well as simple double helix of type A. The ability of being both informational and diverse in structure suggests that RNA was the prebiotic molecule that could function in both replication and catalysis (The RNA World Hypothesis). In fact, some virus encode their genetic materials by RNA (retrovirus).
Three Dimensional Structures of Double Helices

A-DNA

B-DNA

Z-DNA

Forces That Stabilize Nucleic Acid Double Helix

- There are two major forces that contribute to stability of helix formation
  - Hydrogen bonding in base-pairing
  - Hydrophobic interactions in base stacking

Types of DNA Double Helix

- Type A: major conformation of RNA, minor conformation of DNA;
- Type B: major conformation of DNA;
- Type Z: minor conformation of DNA;
- Type A: narrow, tight; Type B: wide, less tight; Type Z: left-handed, least tight.
Secondary Structures of Nucleic Acids

• DNA is primarily in duplex form.
• RNA is normally single stranded which can have a diverse form of secondary structures other than duplex.

More Secondary Structures

Predicting RNA Secondary Structures

• By Thermodynamics Method
  • Minimize Gibbs Free Energy
• By Phylogenetic Comparison Method
  • Compare RNA Sequences of Identical Function From Different Organisms
• By Combination of the Above Two Methods
  • In principle, this could be the most powerful method
Thermodynamics

- Gibbs Free Energy, \( G \)
  - Describes the energetics of biomolecules in aqueous solution. The change in free energy, \( \Delta G \), for a chemical process, such as nucleic acid folding, can be used to determine the direction of the process:
    - \( \Delta G = 0 \): equilibrium
    - \( \Delta G > 0 \): unfavorable process
    - \( \Delta G < 0 \): favorable process
  - Thus the natural tendency for biomolecules in solution is to minimize free energy of the entire system (biomolecules + solvent).

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

- \( \Delta H \) is enthalpy, \( \Delta S \) is entropy, and \( T \) is the temperature in Kelvin.

The change of order of the system.

Molecular interactions, such as hydrogen bonds, van der Waals and electrostatic interactions can contribute to the \( \Delta H \) term. \( \Delta S \) describes the change of order of the system.

Thus, both molecular interactions as well as the order of the system determine the direction of a chemical process.

- For any nucleic acid solution, it is extremely difficult to calculate the free energy from first principle
- Biophysical methods can be used to measure free energy changes

The Equilibrium Partition Function

- For a population of structures, \( S \), a partition function \( Q \) and the probability for a particular folding, \( s \), can be calculated:
  
  \[
  Q = \sum_{s} e^{\frac{\Delta G}{k_{B}T}}
  \]

- The heat capacity for the RNA can be obtained:
  
  \[
  G = \langle RT \ln Q \rangle \quad \text{and} \quad C_{v} = \frac{\partial^{2} G}{\partial T^{2}}
  \]

Heat capacity can be measured experimentally.

Energy Minimization Method (mFOLD)

- An RNA sequence is called \( R = \{ r_{1}, r_{2}, r_{3}, \ldots, r_{n} \} \), where \( r \) is the \( i^{th} \) ribonucleotide and it belongs to a set of \( \{ A, U, G, C \} \)
- A secondary structure of \( R \) is a set \( S \) of base pairs, \( i, j \), which satisfies
  - \( 1 < i < j < n \)
  - \( j - i > 4 \)
  - If \( i, j \) and \( i', j' \) are two base pairs, \( \text{max}(j - i), \text{max}(j' - i') > 4 \)

\[
E_{ij} = \begin{cases} 
0 & \text{if } i \neq j \text{ and } i, j \text{ are not paired} \\
\varepsilon & \text{if } i = j \\
-\varepsilon & \text{if } i = j' \text{ or } j = j' \\
\varepsilon & \text{if } i' = j \text{ or } j' = i' \text{ (includes pseudoknots which is }-1/2) \\
-\varepsilon & \text{if } i' = j' \text{ (includes pseudoknots which is }-1/2) \\
\varepsilon/2 & \text{if } i = j' \text{ or } j = i' \\
-\varepsilon/2 & \text{if } i' = j \text{ or } j' = i \\
\end{cases}
\]

- The objective is to minimize \( E(S) \).
Free Energy Parameters

- Extensive database of free energies for the following RNA units has been obtained (so called “Tinoco Rules” and “Turner Rules”):
  - Single Strand Stacking energy
  - Canonical (AU/GC) and non-canonical (GU) basepairs in duplexes
- Still lacking accurate free energy parameters for:
  - Loops
  - Mismatches (AA, CA etc)
- Using these energy parameters, the current version of mFOLD can predict ~73% phylogenetically deduced secondary structures.

Dynamic Programming (mFOLD)

- An Example of \( W(i,j) \)
  - A matrix \( W(i,j) \) is computed that is dependent on the experimentally measured basepair energy \( e(i,j) \)
  - Recursion begins with \( i=1, j=n \)
  1. If \( W(i+1,j)=W(i,j) \), then \( i \) is not paired. Set \( i=i+1 \) and start the recursion again.
  2. If \( W(i,j-1)=W(i,j) \), then \( j \) is not paired. Set \( j=j-1 \) and start the recursion again.
  3. If \( W(i,j)=W(i,k)+W(k+1,j) \), the fragment \( k+1, j \) gets put on a stack and the fragment \( i...k \) is analyzed by setting \( j=k \) and going back to the recursion beginning.
  4. If \( W(i,j)=e(i,j)+W(i+1,j-1) \), a basepair is identified and is added to the list by setting \( i=i+1 \) and \( j=j-1 \)

Suboptimal Folding (mFOLD)

- For any sequence of \( N \) nucleotides, the expected number of structures is greater than \( 1.8^N \)
- A sequence of 100 nucleotides has \( 3 \times 10^{25} \) foldings. If a computer can calculate 1000 strs./s, it would take \( 10^{13} \) years!
- mFOLD generates suboptimal foldings whose free energy fall within a certain range of values. Many of these structures are different in trivial ways. These suboptimal foldings can still be useful for designing experiments.
Running mFOLD

- [http://bioinfo.math.rpi.edu/~mfold/rna/form1.cgi](http://bioinfo.math.rpi.edu/~mfold/rna/form1.cgi)

- **Constraints can be entered**
  1. Force bases \(i+i,...,i+k-1\) to be double stranded by entering:
     \[ F_{i\ldots i+k-1} \] on 1 line in the constraint box.
  2. Force consecutive base pairs \(i+j,...,i+k-1,j-k+1\) by entering:
     \[ F_{i+j} \] on 1 line in the constraint box.
  3. Force bases \(i+i,...,i+k-1\) to be single stranded by entering:
     \[ P_{i\ldots i+k-1} \] on 1 line in the constraint box.
  4. Prohibit the consecutive base pairs
     \(i+j,...,i+k-1,j-k+1\) by entering:
     \[ P_{i+j} \] on 1 line in the constraint box.
  5. Prohibit bases \(i\ldots j\) from pairing with bases \(k\ldots l\) by entering:
     \[ P_{ij} \] on 1 line in the constraint box.

---

**Fold**

5'-CUUGGAUGGGUGACCACCUGGG-3'

No constraint

\[ F_{i\ldots i+21} \] entered

---

**Secondary Structure Prediction**

**for Aligned RNA Sequences**

- Both energy as well as RNA sequence covariation can be combined to predict RNA secondary structures

- To quantify sequence covariation, let \(f(x)\) be the frequency of base X at aligned position \(i\) and \(f(x|y)\) be the frequency of finding X in \(i\) and Y in \(j\), the mutual information score is (Chiu & Kolodziejczak and Gutef & Wolfe)

\[
M_{ij} = \prod_x \frac{f_i(x|y) \log \frac{f_i(x|y)}{f_i(x)f_i(y)}}{f_i(x)f_i(y)}
\]

if for instance only GC and GU pairs at positions \(i\) and \(j\) then \(M_{ij} = 0\).

- The total energy for RNA is set to a linear combination of measured free energy plus the covariance contribution
Other Secondary Prediction Methods

- **Vienna:**
  - Uses the same recursive method in search the folding space
  - Added the option of computing the population of RNA secondary structures by the equilibrium partition function
  - Specific heat of an RNA can be calculated by numerical differentiation from the equilibrium partition function

- **RNACAD:**
  - http://www.cse.ucsc.edu/research/compbio/ssurrna.html
  - An effort in improving multiple RNA sequence alignment by taking into both primary as well secondary structure information
  - Uses Stochastic Context-Free Grammars (SCFGs), an extension of Hidden Markov models (HMMs) method

  - This work treats RNA as heteropolymer and uses a simplified Go-like model to provide an exact solution for RNA transition between its native and molten phase.

Predicting RNA 3D Structures

- Currently available RNA 3D structure prediction programs make use the fact that a tertiary structure is built upon preformed secondary structures
- So once a solid secondary structure can be predicted, it is possible to predict its 3D structure
- The chances of obtaining a valid 3D structure can be increased by known space constraints among the different secondary segments (e.g. cross-linking, NMR results).
- However, there are far less thermodynamic data on 3-D RNA structures which makes 3-D structure prediction challenging.

Mc-Sym

- Mc-Sym uses “backtracking” method to solve a general problem in computer science called the constraint satisfaction problem (CSP)
- Backtracking algorithm organizes the search space as a tree where each node corresponds to the application of an operator
- At each application, if the partially folded RNA structure is consistent with its RNA conformational database, the next operator is applied, otherwise the entire attached branch is pruned and the algorithm backtracks to the previous node.
Mc-Sym (Continued)

• The selection of a spanning tree for a particular RNA is left to the user, but it is suggested that the nucleotides imposing the most constraints are introduced first

• Users also supply a particular Mc-Sym "conformation" for each nucleotide. These "conformers" are derived from currently available 3D databases

Mc-Sym (Continued)

Sample script:

```
SEQUENCE

1       A       r       GAAUGCCUGCGAGCAUCCC

DECLARE

RELATIONS

18      helix    *      19
17      helix    *      18
16      helix    *      17
5       helix    *      6
4       helix    *      5
3       helix    *      4
2       helix    *      3
1       helix    *      2

BUILD

19      18      17      16      15      14      13      12
12      11      10      9       8       7       6       5
5
4
3
2
1

CONSTRAINTS

18     2     3.0
```

RNA-protein Interactions

• There is currently no computational method that can predict the RNA-protein interaction interfaces;

• Statistical methods have been applied to identify structure features at the protein-RNA interface. For instance, ENTRANCL finds that most atoms contributed from a protein to recognizing an RNA are from main chains (C, O, N, H), not from side chains! But much remain to be done;

• Electrostatic potential has primary importance in protein-RNA recognition due to the negatively charged phosphate backbones. Efforts are made to quantify electrostatic potential at the molecular surface of a protein and RNA in order to predict the site of RNA interaction. This often provides good prediction at least for the site on the protein.
References

Predicting RNA secondary structures:

good reviews


obtaining experimental thermodynamic parameters:


thermodynamics theory for RNA structure prediction: