RNA 2\textsuperscript{nd} structure prediction
Why?

- RNA molecules can be classified in:
  - messenger (coding) RNA
  - non-coding RNA
- The non-coding RNAs have a wide range of functions that is (believed to be) determined by its tertiary structure
- The scaffold for the tertiary structure is provided by the 2nd structure
RNA sequence

- RNA (RiboNucleic Acid) molecules are very similar to DNA (DeoxyriboNucleic Acid) molecules.
- Each molecule is made of a chain of nucleotides (bases). There are only four nucleotides.
- Thus, the sequence (or primary structure) of the RNA molecule can be represented as a string over the alphabet \{A, C, G, U\}.

![adenine](image1)
![cytosine](image2)
![guanine](image3)
![uracil](image4)
RNA 2\textsuperscript{nd} structure

• Unlike DNA, RNA is produced as a single stranded molecule which then folds to form base pairs (2\textsuperscript{nd} structure)

• The typical base pairs are:
  • canonical (Watson - Crick):
    – A and U
    – C and G
  • non-canonical:
    – G and U

• RNA can have other base pairs but they are formed with very low frequency
RNA 2\textsuperscript{nd} structure representation

- **Primary structure**
  \[UUUGGAUAAAA\]

- **Sequence of base pairs**
  \{1 \cdot 11, 2 \cdot 10, 3 \cdot 9\}

- **Bracket**
  ((((((...))))))

- **Dome**

- **Standard graphical representation**
Base pairs

- Any base can take part in at most one base pair
- Two base pairs can be in one of three configurations

- Overlapping base pairs form a pseudoknot
- A 2\textsuperscript{nd} structure without pseudoknots can be represented as a planar graph
RNA 2\textsuperscript{nd} structure prediction

- Energy minimization
  - predict a 2\textsuperscript{nd} structure of least free energy
  - based on primary structure only
  - example Nussinov, Zuker's Mfold

- Comparative structure prediction
  - predict 2\textsuperscript{nd} structures for several sequences
  - based on a prior (reliable) alignment

- Probabilistic models
  - example SCFGs (stochastic context free grammars)
Nussinov

- Minimum energy ↔ maximum number of base pairs
- Calculate best structure for small subsequences and work outwards to larger and larger subsequences

**Notations**

- \(\text{seq}\) the RNA sequence
  
  (over alphabet \{A, C, G, U\})

- \(\text{seq}[i, j]\) the RNA sequence from position \(i\) to \(j\)

- \(\text{str}\) the best 2\(^{nd}\) structure for \(\text{seq}\)
  
  (over alphabet \{(, ), .\})

- \(\text{str}[i, j]\) the best 2\(^{nd}\) structure for \(\text{seq}[i, j]\)

- \(\text{score}[i, j]\) the number of base pairs in \(\text{str}[i, j]\)
Nussinov

- i unpaired and \( str[i+1, j] \)
- j unpaired and \( str[i, j-1] \)
- \( seq[i] \cdot seq[j] \) and \( str[i+1, j-1] \)
- \( str[i, k] \) and \( str[k+1, j] \) for some \( i < k < j \)
\[
\text{score}[i, j] = \begin{cases} 
0 & \text{if } j - i < 2 \\
\max & \text{else} \\
\text{score}[i+1, j] \\
\text{score}[i, j-1] \\
\text{score}[i+1, j-1] + 1 & \text{if } \text{seq}[i] \cdot \text{seq}[j] \\
\max_{i < k < j-1} (\text{score}[i, k] + \text{score}[k+1, j]) 
\end{cases}
\]
\[ \text{score}[i, j] = \begin{cases} 
0 & \text{if } j - i < 2 \\
\max \{ \text{score}[i+1, j], \text{score}[i, j-1], \text{score}[i+1, j-1] + 1 \text{ if } \text{seq}[i] \cdot \text{seq}[j], \max_{i < k < j-1}(\text{score}[i, k] + \text{score}[k+1, j]) \} 
\end{cases} \]
Nussinov

\[
\text{score}[i, j] = \begin{cases} 
0 & \text{if } j - i < 2 \\
\max & \text{otherwise}
\end{cases}
\]

\[
\begin{align*}
\text{score}[i, j] = & \text{score}[i + 1, j] \\
& \text{score}[i, j - 1] \\
& \text{score}[i + 1, j - 1] + 1 \text{ if } seq[i] \cdot seq[j] \\
& \max_{i < k < j - 1} \left( \text{score}[i, k] + \text{score}[k + 1, j] \right)
\end{align*}
\]

Space? $O(n^2)$  
Time? $O(n^3)$
\[
\text{score}[i, j] = \begin{cases} 
0 & \text{if } j - i < 2 \\
\max \left\{ \begin{array}{l}
\text{score}[i + 1, j] \\
\text{score}[i, j - 1] \\
\text{score}[i + 1, j - 1] + 1 & \text{if } \text{seq}[i] \cdot \text{seq}[j] \\
\max_{i < k < j - 1} (\text{score}[i, k] + \text{score}[k + 1, j]) 
\end{array} \right. 
\end{cases}
\]
Nussinov

\[
\text{score}[i, j] = \begin{cases} 
0 & \text{if } j - i < 2, \\
\max 
& \left\{ \begin{array}{l}
\text{score}[i + 1, j] \\
\text{score}[i, j - 1] \\
\text{score}[i + 1, j - 1] + 1 & \text{if } \text{seq}[i] \cdot \text{seq}[j] \\
\max_{i < k < j - 1} (\text{score}[i, k] + \text{score}[k + 1, j])
\end{array} \right.
\end{cases}
\]
**Backtracking**

![Backtracking Diagram]

<table>
<thead>
<tr>
<th>U</th>
<th>U</th>
<th>U</th>
<th>G</th>
<th>G</th>
<th>A</th>
<th>U</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>U</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\text{score}[i, j] = \begin{cases} 
0 & \text{if } j - i < 2 \\
\max \left\{ \begin{array}{l}
\text{score}[i + 1, j] \\
\text{score}[i, j - 1] \\
\text{score}[i + 1, j - 1] + 1 & \text{if } \text{seq}[i] \cdot \text{seq}[j] \\
\max_{i < k < j - 1} (\text{score}[i, k] + \text{score}[k + 1, j])
\end{array} \right. 
\end{cases}
\]
Backtracking

\[
\text{score}[i,j] = \begin{cases} 
0 & \text{if } j - i < 2 \\
\max \left\{ \text{score}[i+1,j], \text{score}[i,j-1], \text{score}[i+1,j-1] + 1 \right\} & \text{if } \text{seq}[i] \cdot \text{seq}[j] > 0 \\
\max_{i < k < j-1} \left( \text{score}[i,k] + \text{score}[k+1,j] \right) & \text{otherwise}
\end{cases}
\]
\[
\text{score}[i, j] = \begin{cases} 
0 & \text{if } j - i < 2 \\
\max \left\{ \begin{array}{l}
\text{score}[i + 1, j] \\
\text{score}[i, j - 1] \\
\text{score}[i + 1, j - 1] + 1 \text{ if } seq[i] \cdot seq[j] \\
\max_{i < k < j - 1} (\text{score}[i, k] + \text{score}[k + 1, j])
\end{array} \right. 
\end{cases}
\]

Time? \(O(n^2)\)

\[
\left( \frac{4}{3} \left( \frac{2}{1} \right) \left( \frac{1}{1} \right) \right) \left( \frac{2}{3} \right) \left( \frac{3}{3} \right) = \left( \frac{4}{3} \right) \left( \frac{2}{1} \right) \left( \frac{1}{1} \right) \left( \frac{2}{3} \right) \left( \frac{3}{3} \right) = \left( \frac{4}{3} \right) \left( \frac{2}{1} \right) \left( \frac{1}{1} \right) \left( \frac{2}{3} \right) \left( \frac{3}{3} \right).
\]
Nussinov

(((...)))

VS

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