RNA Secondary Structure Prediction: taking conservation into account
Assumptions of the RNA secondary structure prediction algorithm, based on MFE:

1. The most likely structure of the RNA molecule is identical or similar to the energetically most stable structure.

2. The energy associated with any position in the structure is only influenced by local sequence structure.

3. The structure is assumed to be formed by folding of the chain back on itself in a manner that does not produce any pseudoknots.

“Legal” structural elements

“Illegal” structural elements
A recursive solution

\[ L(i, j) = \max_{i \leq q \leq j} \left( \max_{i \leq q \leq j} \left( \text{[diagram]} \right) \right) \]
The Nussinov-Jacobson Algorithm

- A DP algorithm which performs a bottom-up computation of the recurrence.

- Uses a table $M$ which stores solutions for subsequences: $M[i,j] = L(i,j)$.

- Upon reaching $M[i,j]$, all entries which are needed for the computation of $L(i,j)$ have already been computed and stored in $M$. 
Main approaches to RNA secondary structure prediction

- **Energy minimization (Single-strand Folding)**
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure (could be based on parameter-learning)

- **Comparative structure analysis**
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
Main approaches to RNA secondary structure prediction

- **Energy minimization (Single-strand Folding)**
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure (could be based on parameter-learning)

- **Comparative structure analysis**
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
Given an RNA sequence, predict its energetically most stable structure (minimal free energy).

**The RNA Folding Problem:**

Given an RNA sequence, predict its energetically most stable structure (minimal free energy).

AUCCCCGUAUCGAUC
AAAAAUCCCAUGGGGUAC
CCUAGUGAAAGUGUA
UAUACGUGCUCUGAU
UCUUUACUGAGGAGU
CAGUGAACGAACUGA
The Nussinov-Jacobson Algorithm

- Space complexity: $O(n^2)$
- Time complexity: $O(n^3)$
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
On the MFE approach

- “some species can remain kinetically trapped in nonequilibrium states... we expect that most RNA’s exist naturally in their thermodynamically most stable configurations” — Tinoco and Bustamante, J. Mol. Biol. 1999.
Energy Minimization Methods

Given the energy tables, and a folding, the free energy can be calculated for a structure.
Energy Minimization

• Thermodynamic Stability
  – Mathews&Turner measured the energy behavior of structural components using experimental techniques
  – Theory: Most Stable is the Most likely

• Zuker and Stiegler’s Algorithm $O(n^3)$
  Computes RNA structure by finding the conformation with the minimum energy based on Mathews&Turner’s formulations.
  – MFOLD: http://www.rpi.edu/~zukerm
  – Vienna RNA Package: http://www.tbi.univie.ac.at/~ivo/RNA
A. Single-stranded RNA

B. Double-stranded RNA helix of stacked base pairs

C. Stem and loop or hairpin loop.

D. Bulge loop

E. Interior loop

F. Junctions or multi-loops.

\[
L(i, j) = \max_{i < q \leq j} \left( \begin{array}{c}
\text{max} \\
\end{array} \right)
\]
Tables 5.2. Predicted free-energy values (kcal/mole at 37°C) for base pairs and other features of predicted RNA secondary structures

<table>
<thead>
<tr>
<th></th>
<th>A/U</th>
<th>C/G</th>
<th>G/C</th>
<th>U/A</th>
<th>G/U</th>
<th>U/G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/U</td>
<td>−0.9</td>
<td>−1.8</td>
<td>−2.3</td>
<td>−1.1</td>
<td>−1.1</td>
<td>−0.8</td>
</tr>
<tr>
<td>C/G</td>
<td>−1.7</td>
<td>−2.9</td>
<td>−3.4</td>
<td>−2.3</td>
<td>−2.1</td>
<td>−1.4</td>
</tr>
<tr>
<td>G/C</td>
<td>−2.1</td>
<td>−2.0</td>
<td>−2.9</td>
<td>−1.8</td>
<td>−1.9</td>
<td>−1.2</td>
</tr>
<tr>
<td>U/A</td>
<td>−0.9</td>
<td>−1.7</td>
<td>−2.1</td>
<td>−0.9</td>
<td>−1.0</td>
<td>−0.5</td>
</tr>
<tr>
<td>G/U</td>
<td>−0.5</td>
<td>−1.2</td>
<td>−1.4</td>
<td>−0.8</td>
<td>−0.4</td>
<td>−0.2</td>
</tr>
<tr>
<td>U/G</td>
<td>−1.0</td>
<td>−1.9</td>
<td>−2.1</td>
<td>−1.1</td>
<td>−1.5</td>
<td>−0.4</td>
</tr>
</tbody>
</table>

B. Destabilizing energies for loops

<table>
<thead>
<tr>
<th>Number of bases</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>–</td>
<td>5.3</td>
<td>6.6</td>
<td>7.0</td>
<td>7.4</td>
</tr>
<tr>
<td>Bulge</td>
<td>3.9</td>
<td>4.8</td>
<td>5.5</td>
<td>6.3</td>
<td>6.7</td>
</tr>
<tr>
<td>Hairpin</td>
<td>–</td>
<td>4.4</td>
<td>5.3</td>
<td>6.1</td>
<td>6.5</td>
</tr>
</tbody>
</table>

(Upper) Stacking energy in double-stranded region when base pair listed in left column is followed by base pair listed in top row. C/G followed by U/A is therefore the dinucleotide 5’ CU 3’ paired to 5’ AG 3’.

(Lower) Destabilizing energies associated with loops. Hairpin loops occur at the end of a double-stranded region, internal loops are unpaired regions flanked by paired regions, and a bulge loop is a bulge of one strand in an otherwise paired region (Fig. 5.2). An updated and more detailed list of energy parameters may be found at the Web site of M. Zuker (http://bioinfo.math.rpi.edu/~zuker/rna/energy/).

Comparison of Methods
Suboptimal Folds

- The correct structure is not necessarily the structure with optimal free energy
- Within a certain threshold of the calculated minimum energy
- MFOLD updated to report suboptimal folds
יתכן והשוש לבוש יוצר
מפתרון אחד,닫ה במכות
אגוריה דומיה מתכזרים
פרטונות מעוד שומימ,
יש להתייחס ליבגי בפסגון

Figure 5.9. Model of RNA secondary structure of the human adenovirus pre-terminal protein. This model is one of several alternative structures represented by the above energy plot and provided as an output by the current versions of MFOLD. (A) Simple text representation of one of the predicted structures. Each stem-and-loop structure is shown separately and the left end of each structure is placed below the point of connection to the one above. (B) More detailed rendition of one part of the predicted structures. The structure continues beyond the right side of the page.
MFOLD

The MFOLD program allows for the calculation of RNA secondary structures. It is based on the thermodynamic principle that RNA structures are determined by the minimization of free energy. The program takes an RNA sequence as input and predicts the most stable secondary structure. The predictions are based on a set of energy parameters that are calculated using various thermodynamic models. The program outputs the predicted structure in a format that can be visualized using graphics software.

Figure 5.9. Model of RNA secondary structure of the human adenovirus pre-terminal protein. This model is one of several alternative structures represented by the above energy plot and provided as output by the current version of MFOLD. (A) Simple text representation of one of the predicted structures. Each stem-and-loop structure is shown separately and the left end of each structure is placed below the point of connection to the one above. (B) More detailed rendition of one part of the predicted structures. The structure continues beyond the right side of the page.
לא פרקים

2)

1)

 Пет סMouseDown insanely

 יפ הינב

2) עשת זהב כולל בציר

 לקס

 אם זה

13 June 2006
Scoring models

- Nussinov-Jacobson (1978): each base-pair gets one “point”

- Zuker-Stiegler (1981): specific scores for stacked base-pairs \((i, j)\), which depend on the nucleotides at positions \(i, i+1, j-1, \) and \(j\). Additional scoring terms for hairpins and internal loops, based on their lengths

- Turner, Mathews, et al. (1999, 2004): extends the “feature vocabulary” by examining more structural elements, and refining the scores based on intensive lab measurements

- Andronescu et al. (2007): refined scores for the “Turner model”, obtained by applying a machine learning procedure to a comprehensive dataset of known RNA structures
Context Fold

- *Context Fold* is a new RNA folding suite
  [Zakov, Goldberg, Elhadad, Ziv-Ukelson (RECOMB 2011)]

- Improvements with respect to previous tools:
  - It allows for “rich parameterization”: flexible scoring models with much larger feature spaces
  - A built-in state of the art machine learning algorithm: the fast and effective “Online Passive Aggressive” algorithm [Crammer et al. (2006)]
  - Incorporation of advanced sparsification techniques for significant time complexity improvements [Backofen, Tsur, Zakov, Ziv-Ukelson (2010)]
The folding prediction problem

AAUAACUAUAAACUGUCUAAAGGCUUUUAGUAUUGUUUUAAAG
Feature-based scoring model

Score = 31
Standard Features
Assigning feature weights

- Measuring free energies in lab experiments
  - Tinoco et al. 1971, 1973
  - Mathews et al. 1999, 2004

- Using *Machine Learning* techniques
  - Sakakibara et al. 1994
  - Dowell and Eddy 2004
  - Do et al. 2006, 2007
  - Andronescu et al. 2007, 2011
FEATURE-RICH MODELS
FEATURE-RICH MODELS
FEATURE-RICH MODELS
More sequential contexts

FEATURE-RICH MODELS
More structural contexts: “bulge hierarchy”

- Internal loop
- Right bulge
- Left bulge

0-left-il  1-left-il  2-left-il  3-left-il  Internal loop
The **Structured Online Passive-Aggressive** algorithm

- **Iteration:**
  - Predict **one** sample from the training set
  - Compare the features of the predicted structure and the true one, and update weights accordingly

\[
\phi(x,y) \cdot w = \text{score}
\]

Crammer et al. 2006

\[
w^i = \begin{cases} 
    w^{i-1}, & \rho(y, \hat{y}) = 0, \\
    w^{i-1} + \tau_i \Phi(x, y) - \tau_i \Phi(x, \hat{y}), & \text{otherwise.}
\end{cases}
\]

where:

\[
\tau_i = \min \left( 1, \frac{\Phi(x, \hat{y})^T \cdot w^{i-1} - \Phi(x, y)^T \cdot w^{i-1} + \sqrt{\rho(y, \hat{y})}}{||\Phi(x, \hat{y}) - \Phi(x, y)||^2} \right)
\]
Results

- Testing dataset: 659 structures, average length of 279

- Training datasets:
  - [1] 2586 structures, average length of 267
  - [2] 2518 structures, average length of 331
  - [3] 1291 structures with measured free energies, average length of 19

<table>
<thead>
<tr>
<th>Model</th>
<th>#param</th>
<th>$F_1$(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Turner 99</td>
<td>363</td>
<td>59.8</td>
</tr>
<tr>
<td>* CG [1], [3]</td>
<td>363</td>
<td>66.8</td>
</tr>
<tr>
<td>* BL [1], [3]</td>
<td>363</td>
<td>67.7</td>
</tr>
<tr>
<td>* BL-FR [1], [3]</td>
<td>7726</td>
<td>69.5</td>
</tr>
<tr>
<td>* CONTRAfold [2]</td>
<td>714</td>
<td>65.5</td>
</tr>
<tr>
<td>$\text{St}^{\text{low}}\text{Co}^{\text{low}}$ [1], [3]</td>
<td>~220</td>
<td>57.9</td>
</tr>
<tr>
<td>$\text{St}^{\text{med}}\text{Co}^{\text{med}}$ [1], [3]</td>
<td>4k</td>
<td>68.9</td>
</tr>
<tr>
<td>$\text{St}^{\text{high}}\text{Co}^{\text{med}}$ [1], [3]</td>
<td>7k</td>
<td>72.5</td>
</tr>
<tr>
<td>$\text{St}^{\text{med}}\text{Co}^{\text{high}}$ [1], [3]</td>
<td>38k</td>
<td>81.0</td>
</tr>
<tr>
<td>$\text{St}^{\text{high}}\text{Co}^{\text{high}}$ [1], [3]</td>
<td>70k</td>
<td>84.1</td>
</tr>
</tbody>
</table>

* Taken Andronescu et al. 2011
Conclusions

- Exploiting modern datasets and ML algorithms, it is possible to efficiently train feature-rich models

- Feature-rich models significantly improve RNA folding prediction accuracy
Folding an RNA Sequence of length $n$.

1. Classical, $O(n^3)$,
   [Nussinov et. Al. 1978, 1980]
   [Waterman and Smith 1978]
   [Zuker and Stiegler 1981]
   **MFOLD**: [http://www.rpi.edu/~zukerm](http://www.rpi.edu/~zukerm)
   **Vienna RNA Package**: [http://www.tbi.univie.ac.at/~ivo/RNA](http://www.tbi.univie.ac.at/~ivo/RNA)

2. Complex worst case speedup based on Fast Matrix Multiplication:
   $O(n^3 \log^3 \log n / (\log^2 n)$ [Akutsu 1999]
   Based on Four Russians: $O(n^3 / (\log n)$ [Frid&Gusfield]
   $O(n^3 / (\log^2 n)$ [Pinhas*, Zakov*, Tsur and Ziv-Ukelson 2013]

3. **Sparsification**: Practical speed up: $O(nZ)$ where $Z$ is in $[n, n^2]$
   [Wexler, Zilberstein, Ziv-Ukelson 2007]
   [Backofen, Tsur, Zakov, Ziv-Ukelson 2009]
Context Fold vs. other RNA folders

“... rankings based on the RNAstrand data set were dominated by single-sequence methods: in 7 out of 10 rankings, ContextFold was the best performing method.”
ContextFold:

RECOMB 2011 Best Paper Award
This is to certify that

S. Zahor, Y. Goldberg, H. Ethalad, and M. Ziv-Ukelson

have been awarded the

RECOMB 2011 Best Paper Award

for their contribution

Rich Parameterization Improves RNA Structure Prediction

Vancouver, April 1, 2011

Martin Vigron
Chair, RECOMB Steering Committee
Main approaches to RNA secondary structure prediction

- Energy minimization (Single-strand Folding)
  - does not require prior sequence alignment
  - require estimation of energy terms contributing to secondary structure (could be based on parameter-learning)

- Comparative structure analysis
  - Using sequence alignment to find conserved residues and covariant base pairs.
  - most trusted
Why is RNA Structure Interesting?

Structure

Function
Accessible RNA binding sites
Binding site accessibility

A motif has to be accessible to binding
Binding site accessibility

A motif has to be accessible to binding
Binding site accessibility

A target-site motif accessible to binding
Bioinformatic Structural witnesses for RNA functionality

Witness 1: Structure Stability.

Witness 2: Sequence/Structure Conservation.
(within the structural context).

Structural Cis-Elements: Purine Riboswitch
[Mandal et al., 2003] predicted a potential pseudoknot between the two arms of the purine riboswitch aptamer.
Journal of Computational Biology

A Structure-Based Flexible Search Method for Motifs in RNA

To cite this article:

Published in Volume: 14 Issue 7: September 5, 2007
Witness 1: Stability of Structure (2D, predicted)

AUCCCCGUAUCGAUC
AAAAAUCCAUUUGGUAC
CCUAGUGAAAGUGUA
UAUACGUGCUCUGCUGAU
UCUUUAACUGAGGAGU
CAGUGAAGCAACUGA

RNA Secondary Structure Prediction: O(N^3):

**Mfold**: http://www.rpi.edu/~zukerm

**Vienna RNA Package**: http://www.tbi.univie.ac.at/~ivo/RNA
Witness 2: Sequence Conservation (e.g. in binding sites)

Lactobacillus acidophilus

Lactobacillus delbrueckii

GGUAU \(\Rightarrow\) GGUAU

CCGUA \(\Rightarrow\) CCGUA
Witness 3: Compensatory Mutations (in stems)

**Lactobacillus acidophilus**

**Lactobacillus delbrueckii**

G-U → U-A
Witness 3: Compensatory Mutations (in stems)

Lactobacillus acidophilus

Lactobacillus delbrueckii

G-C → C-G
RNAAlifold (Washietl, Hofacker and Stadler)

**Consistent mutations** - mutations that conserve the stem structure.

**Compensatory mutations** - joint events where a mutation in one nucleotide was compensated by a corresponding mutation in the paired nucleotide in order to conserve the structure.
Another source of information: Conserved Structure

- RNA more conserved by structure than by sequence. Conserved base-pairs should be taken more seriously than an unconserved ones, even though the energy is the same.
Inferring Structure By Comparative Sequence Analysis

- most reliable computational method for determining RNA secondary structure

- consider the example from Durbin, et al., p 266
מונח: Covariation - עקביות. אם שני אירועים מתחזקים יחד,.shutdown
ונﳚים המחשה של פונקצייתiras ה関係
-between שני לאנים, ניתן להניח שהן מייצגות

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>G</th>
<th>G</th>
<th>G</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ron Unger – Bar-Ilan University 2009
Inferring Structure By Comparative Sequence Analysis

- first step is to calculate a multiple sequence alignment
- Requires sequences be similar enough so that they can be initially aligned
- Sequences should be dissimilar enough for covarying substitutions to be detected
Escherichia coli
Hildenbrandia rubra
Bangia fuscopurpurea
Rhodochaete parvula
Cordyceps kanzashiana
Stichococcus bacillarii
Graphiola phoenicis