Inversion Distribution in RNA Sequences

Ming-Ying Leung, Ph.D.
Professor, Department of Mathematical Sciences
Director, Bioinformatics Program
The University of Texas at El Paso
El Paso, Texas, U.S.A.
DNA

• DNA is deoxyribonucleic acid, made up of 4 nucleotide bases
  – Adenine (A)
  – Cytosine (C)
  – Guanine (G)
  – Thymine (T)

• The bases A and T form a complementary pair, so are C and G.

***RNA is ribonucleic acid with uracil (U) replacing the T.
Examples of Inversions in English and RNA
Biological Functions of RNA

- Ribosomal, messenger, and transfer RNA
- miRNA, siRNA for gene regulation
- Viral genomic RNA.
SARS Virus
RNA Secondary Structures

- 3D structure of an RNA molecule is often the key to its function.
- Experimental determination of 3D structures can be time-consuming and costly.
- Useful information about the molecule can be gained from its secondary structure.
RNA Secondary Structure Elements

- Secondary structure refers to the collection of hydrogen-bonded base pairs in the molecule.
- Elements classified into 2 basic categories:
  - stem-loops
  - pseudoknots
Various Types of RNA Secondary Structures

<table>
<thead>
<tr>
<th>Junctions</th>
<th>Internal loops</th>
<th>Bulge</th>
<th>Hairpin loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-stem</td>
<td>5' 3'</td>
<td>3' 5'</td>
<td>5' 3'</td>
</tr>
<tr>
<td>Three-stem</td>
<td>5' 3'</td>
<td>3' 5'</td>
<td>3' 5'</td>
</tr>
<tr>
<td>Four-stem</td>
<td>5' 3'</td>
<td>3' 5'</td>
<td>3' 5'</td>
</tr>
</tbody>
</table>

- **Junctions**
  - Two-stem: 5' → 3'
  - Three-stem: 5' → 3' → 5'
  - Four-stem: 5' → 3' → 5' → 3'

- **Internal loops**
  - 3' 5'
  - asymmetric
  - symmetric

- **Bulge**
  - 3' 5'

- **Hairpin loop**
  - 5' 3'

Color Coding:
- Green: AU, UA, CG, GC, GU, or UG
- Red: AA, CC, GG, UU, AG, GA, AC, CA, CU, or UC
- Orange: A, U, C, or G
Computational Prediction of RNA Secondary Structures

- Development of mathematical models and computational prediction algorithms for stem-loop structures has started as early as the 1980’s.
- These algorithms are typically based on finding the most “stable” secondary structure.
- Overall stability of an RNA structural element is determined by the amount of energy needed to completely break all of the base pairs holding it together. It is expressed as “minimal free energy.”
Representation of RNA Secondary Structures

Bracket view – for computational analysis

::((((((:::[[[[[[[))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))::]]]:))))))))))))))))):

Schematic view – for human interpretation
Secondary Structure Prediction
Programs for the X1-X2 Overlap

<table>
<thead>
<tr>
<th>Program</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mfold</td>
<td><a href="http://www.bioinfo.rpi.edu/applications/mfold/old/rna/">http://www.bioinfo.rpi.edu/applications/mfold/old/rna/</a></td>
</tr>
<tr>
<td>Pknots*</td>
<td><a href="http://www.genetics.wustl.edu/eddy/software/#pk">http://www.genetics.wustl.edu/eddy/software/#pk</a></td>
</tr>
<tr>
<td>KineFold</td>
<td><a href="http://kinefold.u-strasbg.fr/">http://kinefold.u-strasbg.fr/</a></td>
</tr>
<tr>
<td>RNAFold</td>
<td><a href="http://www.tbi.univie.ac.at/~ivo/RNA/">http://www.tbi.univie.ac.at/~ivo/RNA/</a></td>
</tr>
<tr>
<td>PknotsRG</td>
<td><a href="http://bibiserv.techfak.uni-bielefeld.de/bibi/Tools_RNA_Studio.html">http://bibiserv.techfak.uni-bielefeld.de/bibi/Tools_RNA_Studio.html</a></td>
</tr>
</tbody>
</table>

*compilation required (others are web-based)
Challenge in Pseudoknot Prediction

- Pseudoknot prediction requires large amounts of memory and computing time to obtain the optimal and suboptimal structures with minimal free energies.
- Various alternative algorithms restrict the types of pseudoknots to be predicted to keep computation time and storage size under control.
- A large variety of pseudoknots occur in nature. Their omission from computation methods might significantly affect the prediction accuracy.
Grid Computing Approach

- The RNAVLab (RNA Virtual Laboratory) project is initiated to build a grid computing system that identifies computing resources across campus to predict secondary structures for multiple RNA sequences simultaneously.
- Cut a large RNA molecule into shorter segments. The secondary structures of the segments can be predicted individually by different computers.
- Individual predictions for the small pieces are then assembled to give a predicted structure for the original molecule.
Advantage and Challenge of Grid Computing Approach

• The grid computing approach can accommodate a variety of existing and new prediction algorithms in a heterogeneous workflow.
• The challenge lies in ensuring the predicted results of the smaller pieces are sufficiently consistent with one another so that they can be assembled to generate a reasonable structure for the original molecule.
Prediction of 2 Overlapping Segments

(a) PKNOTS-RE prediction of SARS segment 25884 to 25983

(b) PKNOTS-RE prediction of SARS segment 25923 to 26022
Motif Identifier Algorithm

• We proposed a motif identifier algorithm where each overlapping segment carries equal weight in determining the overall structure.

• As inversions are involved in both stem-loops and pseudoknots, it would be reasonable that segments containing a significantly high concentration of inversions should be given more weight. Knowledge of the distribution of inversions in random nucleotide sequences is needed.
Inversions with Stem Length $L$

\[
inversion_{\text{Gap} = 2, \ Center = i + 1/2}
\]

\[
inversion_{\text{Gap} = 3, \ Center = i}
\]
Probability of Observing an Inversion

Except for centers close to the beginning or end of an i.i.d. RNA sequence, the probability $P$ of observing an inversion of minimum stem length $L$ and exact gap size $g$ can be written as:

(a) For $g = 0$ or $1$, $P = \theta^L$, where $\theta = 2(p_A p_U + p_C p_G)$

(b) For even $g \geq 2$ and odd $g \geq 3$, $P = (1 - \theta)\theta^L$

In a long sequence of length $n$, we expect approx. $n\theta^L[1 + \theta + (1 - \theta)g_{\text{max}}]$ inversions of minimum stem $L$ and maximum gap $g_{\text{max}} \leq 2L$. 
The Average Number of Inversions Versus Cumulative Gap Size
The Average Number of Inversions Versus Exact Gap Size
Q-Q plot of Distances Between Successive Inversions
Poisson Process Approximation

- The occurrence of inversions in i.i.d. random RNA sequences can be approximated by a Poisson process for small gap size.
- The approximation becomes less accurate as gap size increases.
Acknowledgements

Collaborators:
Michela Taufer, University of Delaware
Kyle Johnson, UTEP

PhD Students: Clemente Aguilar, Jonathan Mohl
MS Students: Abel Licon, Sravya Tamma
Undergrads: Alexandria Melendez, David Mireles

Funding Support:
Texas Advanced Research Program: 003661-0013-2007
National Science Foundation: DMS0800272
National Institutes of Health: S06GM08012-35,
  5G12RR008124-11, 1R01AI077413, and
  1T36GM078000-01.