Improving the Algorithm to Predict RNA Structures for Frameshifting in the Expression of Overlapping Viral Genes

Ming-Ying Leung
Department of Mathematical Sciences
University of Texas at El Paso (UTEP)

Outline:

• SARS Virus Genomes and Palindromes
• Overlapping Genes and Frameshifting
• Predicted RNA Secondary Structures
• Necessity to Improve Prediction Speed
Severe acute respiratory syndrome (SARS) was first detected late last year and is believed to have originated in southern China. The mysterious disease is apparently resistant to standard treatments and has put health authorities worldwide in a spin to find ways to curb the intensifying outbreak.
SARS Virus Particles in Host Cell

6 nm diameter

1 µm
SARS Virus Particle

- Single-stranded RNA genome with ~30 kilobases
- Only about 7% the genome size of Cytomegalovirus (CMV) with double-stranded DNA
SARS Virus
• 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.
Genes and Genome

DNA

Proteins act alone or in complexes to perform many cellular functions

Genes contain instructions for making proteins

U.S. DEPARTMENT OF ENERGY
DNA and RNA

DNA is deoxyribonucleic acid, made up of 4 nucleotide bases Adenine, Cytosine, Guanine, and Thymine.

RNA is ribonucleic acid, made up of 4 nucleotide bases Adenine, Cytosine, Guanine, and Uracil.

For uniformity of notation, all DNA and RNA data sequences deposited in GenBank are represented as sequences of A, C, G, and T. The bases A and T form a complementary pair, so are C and G.
SARS Virus Genome Map

- Replicase (1a and 1b), spike glycoprotein (S), X1 and X2 occupy 87% of the genome
- Two pairs of overlapping ORFs, 1a & 1b and X1 & X2 (as designated by Rota et al. 2003), are predicted in this region
- 1a and 1b are standard in all coronaviruses, X1 and X2 are unique to SARS. Whether X1 and X2 do code for proteins is still unconfirmed
A Long Palindrome in X1 and X2

TCTTTCAAACAAGCTTGTTAAAGA

Positions: 25962-25983 (22 bases)

• Found in SARS but not in other 6 coronavirus genomes (Chew et al. 2004)
• The next longest palindrome in SARS is 14 bases long
• In the overlapping region of X1 and X2
Palindromes in Letter Sequences

Odd Palindrome:

“\text{A nut for a jar of tuna}”
remove spaces and capitalize

\text{ANUTFORAJAROFTUNA}

Even Palindrome:

“\text{Step on no pets}”

\text{STEPONNOPETS}
**Palindrome:** A string of nucleotide bases that reads the same as its reverse complement. A palindrome must be even in length, e.g. palindrome of length 10:

5’ ..... GCAATATTGC .....3’

<table>
<thead>
<tr>
<th>j - L +1</th>
<th>j</th>
<th>j + 1</th>
<th>j +L</th>
</tr>
</thead>
<tbody>
<tr>
<td>b_1</td>
<td>b_L</td>
<td>b_{L+1}</td>
<td>b_{2L}</td>
</tr>
</tbody>
</table>

We say that a palindrome of length 2L occurs at position j when the (j-i+1)st and the (j+i)th bases are complementary to each other for i=1,…, L. In an i.i.d. sequence model this occurs with probability

\[
\left[ 2\left( p_A p_T + p_C p_G \right) \right]^L.
\]
Probability of Observing a Length 22 Palindrome

- Approximate the palindrome distribution by a Poisson process with rate
  \[ \lambda = np = 0.01008 \]
- Here \( n = \) genome length = 29727, and
  \[ p = \left[ 2(\hat{p}_A\hat{p}_T + \hat{p}_C\hat{p}_G) \right]^{11} \]
- The probability of the occurrence of at least one length 22 palindrome in the genome is
  \[ 1 - e^{-0.01008} \approx 0.01 \]
Expression of Overlapping Genes Requires Frameshifting in Reading

Frameshifting must have the following elements:

- **Slippery Sequence** - a mechanism that allows the “reader” (called ribosome) to slip
- **Stimulatory Element** - a pseudoknot or stem-loop structure that blocks the ribosome

![Diagram showing pseudoknot and slippery sequences](image)

- Pseudoknot: Palindrome-like sequences
- Stem-loop or hairpin loop: Short binding sequence
-1 Frameshifting

- Reading with default frame: GGG then TTT
-1 Frameshifting

- Reading with default frame: GGG then TTT
- Reading after -1 frameshifting: CGG then GTT
# Heptanucleotide Slippery Sequences

A string in the form of \text{XXXYYYN}
where \text{X} = \text{A, T, or G}; \text{Y} = \text{A or T}; \text{and N} = \text{A, T, or C}

- **ORF1a and ORF1b (Overlap: 13398, 1 base only)**
  
<table>
<thead>
<tr>
<th>Strand</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>13201</td>
<td>CATCCAAATC</td>
<td>CTAAAGGATT</td>
<td>CTGTGACCTTG</td>
<td>AAAGGTAAGT</td>
<td>ACCTTCCAAAT</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>13251</td>
<td>ACCTACCACT</td>
<td>TGTGCTAATG</td>
<td>ACCCAGTGGG</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>13301</td>
<td>TCTGTACGGT</td>
<td>CTGCGGAAATG</td>
<td>GGAAAGGTT</td>
<td>ATGGCTGTAG</td>
<td>TTGTGACCAA</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>13351</td>
<td>CTCCGCGAAC</td>
<td>CCTTGTAGCA</td>
<td>GTCTGCGGAT</td>
<td>GCATCAACGT</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
<td>TTTTACACTT</td>
</tr>
<tr>
<td>13401</td>
<td>GTTTGCGGGT</td>
<td>TAAGTCGGAC</td>
<td>CCGTCTTTACA</td>
<td>CCGTGCGLGCA</td>
<td>CAGGCACATG</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>13451</td>
<td>TACTGATGTC</td>
<td>GTCTACAGGG</td>
<td>CTTTGTATTT</td>
<td>TTTAAGGAA</td>
<td>AAAGTTGCTG</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>13501</td>
<td>GTTTTGCAAA</td>
<td>GTTCCTAAAA</td>
<td>ACTAATTGCT</td>
<td>GTCGCTTTCA</td>
<td>GGAGAGGAT</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>13551</td>
<td>GAGGAAGGCCA</td>
<td>ATTTATTAGA</td>
<td>CTCTTTACTTT</td>
<td>GTAGTTAAGA</td>
<td>GGCATACTAT</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
</tbody>
</table>

- **X1 and X2 (Overlap: 25689-26089, 401 bases)**
  
<table>
<thead>
<tr>
<th>Strand</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>25651</td>
<td>GCTTTTGTGG</td>
<td>AAGTGCAAAT</td>
<td>CCAAGAAACC</td>
<td>ATTACTTTAT</td>
<td>GATGCCAACT</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>25701</td>
<td>ACTTTTTTGG</td>
<td>CTGGCACAAC</td>
<td>CATAACTATG</td>
<td>ACTACTGTAT</td>
<td>ACCATATAAC</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>25751</td>
<td>AGTGTCACTG</td>
<td>ATACATTGT</td>
<td>CGTTACTGAA</td>
<td>GGTGACGGCA</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
<td>TTTTACACTT</td>
</tr>
<tr>
<td>25801</td>
<td>AAAACTCAAA</td>
<td>GAAGACTACC</td>
<td>AAATTTGGTG</td>
<td>TTTAATCTGAG</td>
<td>GATAGGCAC</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>25851</td>
<td>GAGTTTTTAA</td>
<td>AGACTATGTC</td>
<td>GTTGTACAGT</td>
<td>GCTATTCCA</td>
<td>CGAAGTTTAC</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>25901</td>
<td>TACCAGCTTG</td>
<td>AGTCTACACA</td>
<td>AATTACTACA</td>
<td>GACACTGTTA</td>
<td>TTGAAAAATGC</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>25951</td>
<td>TACATTCTTC</td>
<td>ATCTTTAACAC</td>
<td>AGCTGGTTAA</td>
<td>AGACCCACCG</td>
<td>AATGTGACAA</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>26001</td>
<td>TACACACAAT</td>
<td>CGACGGCTCT</td>
<td>TCAGGAGTTG</td>
<td>CTAATCCAGC</td>
<td>AATGGATCCA</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>26051</td>
<td>ATTTATGATG</td>
<td>AGCGGCACGC</td>
<td>GACTACTAGC</td>
<td>GTCGCTTTGT</td>
<td>AAGCCACAAGA</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
<tr>
<td>26101</td>
<td>AAGTGAGTAC</td>
<td>GAACCTTATGT</td>
<td>ACTCTTCCTG</td>
<td>TTGCGGAAGAA</td>
<td>ACAGTGACGT</td>
<td>TTTTACACTT</td>
<td>AGAAACACAG</td>
</tr>
</tbody>
</table>
Slippery Sequences:

- Right preceding the overlapping base between 1a and 1b, there is a slippery sequence followed by a pseudoknot (Theil et al. 2003)
- Possible slippery sequences are detected in the overlapping region of X1 and X2; any pseudoknot or stem-loop structure in close proximity downstream?
## 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>Pknotts*</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>PknottsRG</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)
Secondary Structure Prediction

• Any program capable of predicting pseudoknots (*KineFold, Pknots, PknotsRG*) will not allow long sequences

• None of X1, X2 or even their overlapping region is accepted by the pseudoknot prediction programs

• Focus on a segment of about 200 bases in the overlapping region containing the 22-base palindrome at bases 25962-25983, with a slippery sequence located at 25941-25947
Hairpin Loop Predicted by *Mfold*

\[ \Delta G = -55.2 \text{ kcal/mol} \]
Hairpin Loop Predicted by *Pknuts*

\[ \Delta G = -54.8 \text{ kcal/mol} \]
Hairpin Loop Predicted by *RNAFold*

\[ \Delta G = -58.75 \text{ kcal/mol.} \]
Pseudoknot Predicted by *PknutsRG*

\[ \Delta G = -55.84 \text{ kcal/mol.} \]
Pseudoknot Predicted by *KineFold*

\[ \Delta G = -49.9 \text{ kcal/mol}. \]
Predicted Structures

- All programs consistently indicate that the 22-base palindrome folds into a hair-pin loop. Free energy of predicted structures ranges from -58.75 to -49.9 kcal/mol.
- *RNAFOLD* predicts that the palindrome is part of a stem-loop structure (lowest $\Delta G = -58.75$ kcal/mol).
- *PknotsRG* predicts that the palindrome is part of a pseudoknot (2nd lowest $\Delta G = -55.84$ kcal/mol).
- In each case above, the structure predicted immediately follows the slippery sequence at 25941-25947.
Sequence Selection and Parallelization

• Unusual palindrome and slippery sequence help to select sequence segment for structural prediction.
• Run program *Pknots* on one processor of the IBM p690 parallel processor continuously for over 4 days for the selected segment of about 200 bases. Currently attempting to parallelize the algorithm to run on multiple processors or distributed systems.
Conclusion and Question

• It seems likely that these ORFs do code for real proteins, but X2 should probably starts at base 25947 rather than 25689.

• Can the prediction of frameshifting mechanism be improved by applying the appropriate secondary structure prediction algorithms based on energy minimization or stochastic context free grammar?