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:

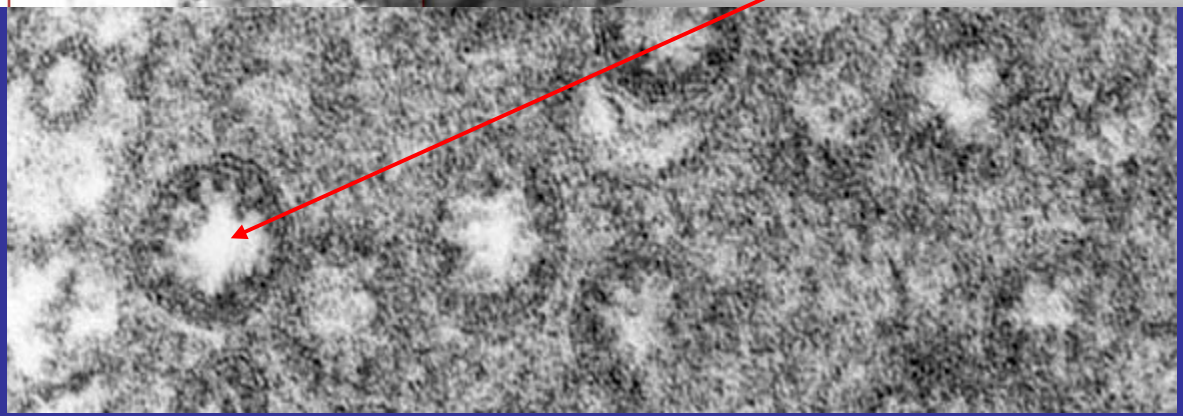
- SARS Virus Genomes and Palindromes
- Overlapping Genes and Frameshifting
- Predicted RNA Secondary Structures
- Necessity to Improve Prediction Speed

Next

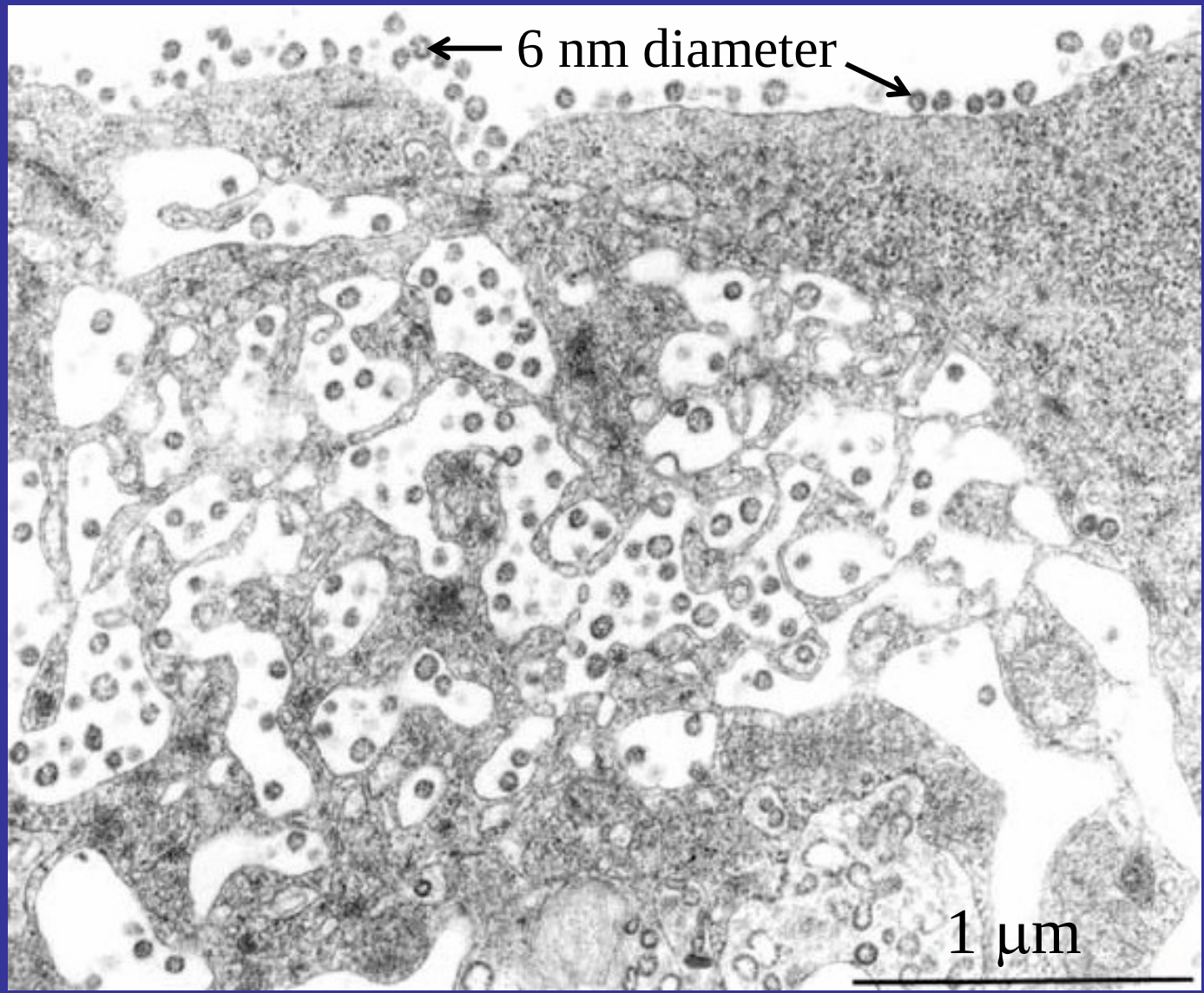
Severe acute respiratory syndrome (SARS) was first detected late last year and is believed to have originated in southern China. The mysterious disease is apparently resistant to standard treatments and has put health authorities worldwide in a spin to find ways to curb the intensifying outbreak.

AP Photos/World Health Organistion/CNN Graphics

SARS Viral Particles



SARS Virus Particles in Host Cell



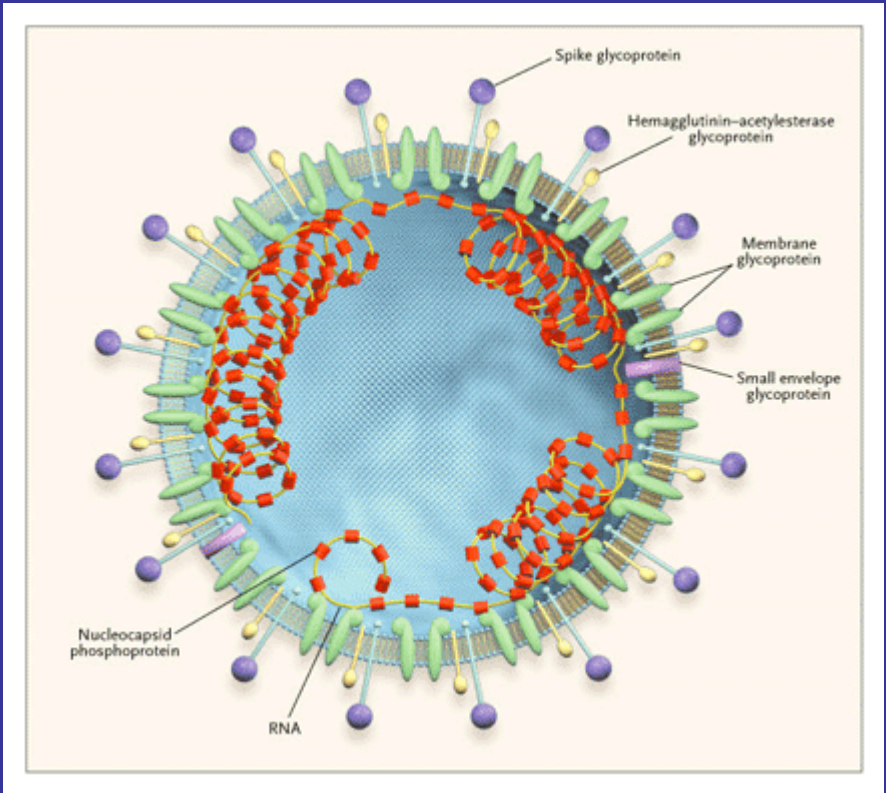
SARS Virus Particle

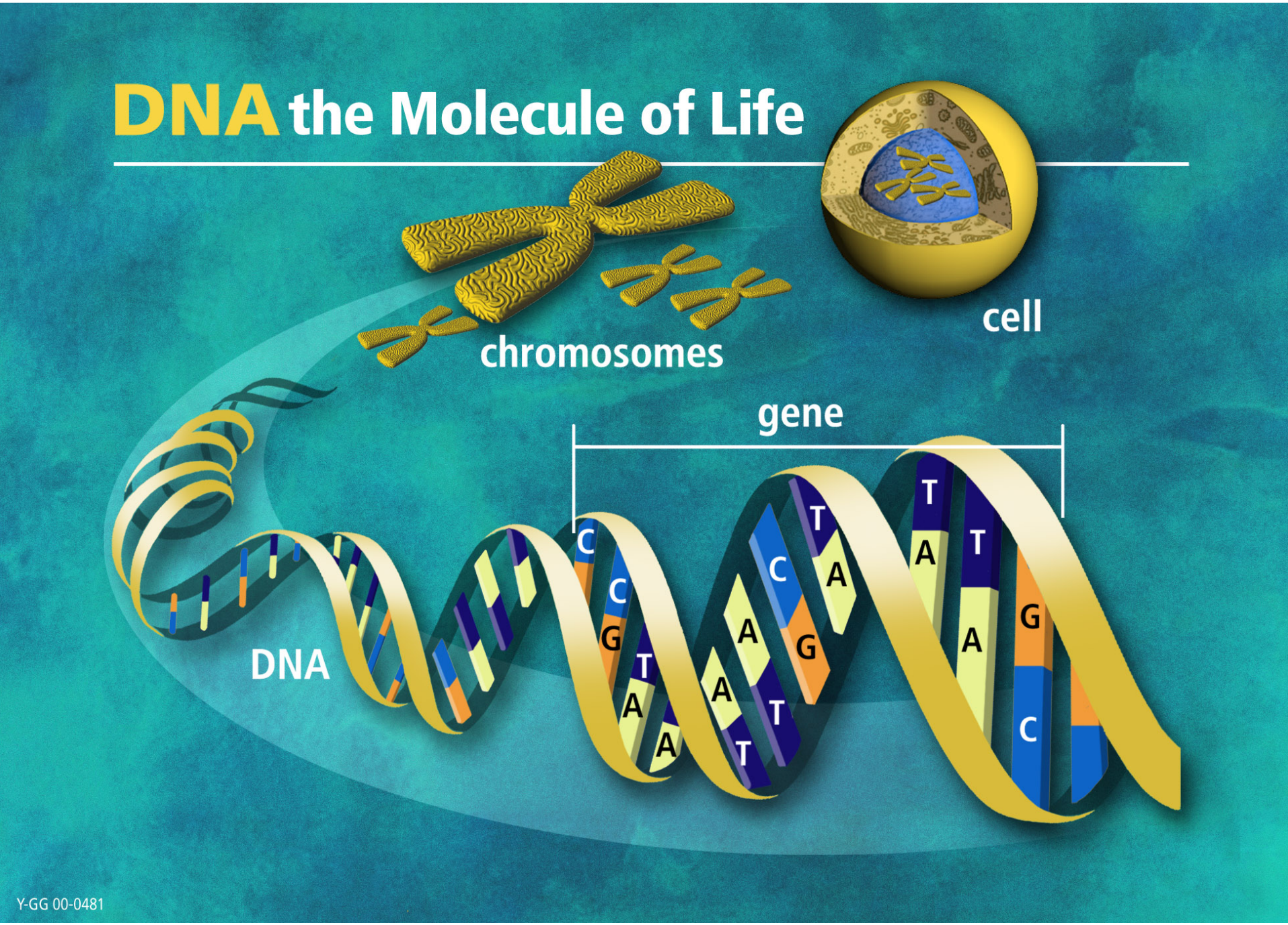
Host Cell

Spike Glycoprotein

- Single-stranded RNA genome with ~30 kilobases
- Only about 7% the genome size of Cytomegalovirus (CMV) with double-stranded DNA

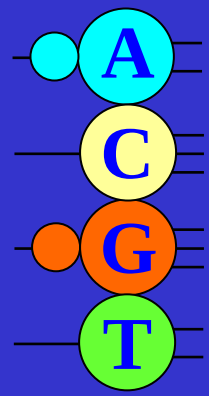
SARS Virus


Spike glycoprotein
Hemagglutinin–acetylesterase glycoprotein
Membrane glycoprotein
Small envelope glycoprotein
Nucleocapsid phosphoprotein
RNA

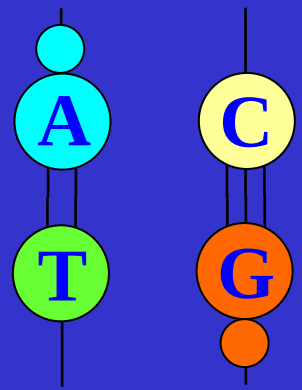

DNA the Molecule of Life
chromosomes
cell
gene
DNA
C
C
G
T
A
A
A
T
T
C
G
T
A
T
A
T
A
G
C
Y-GG 00-0481

DNA

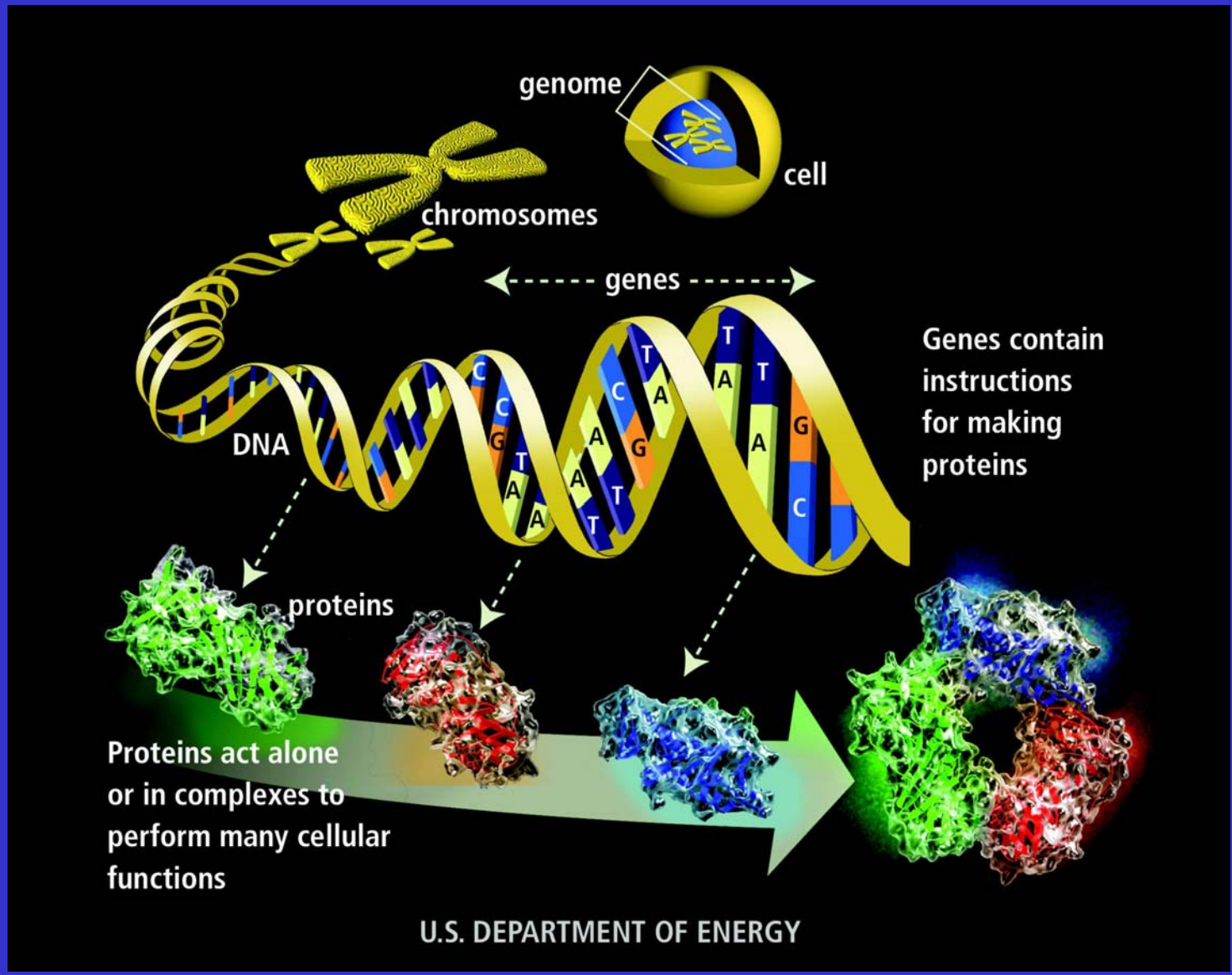
- **DNA is deoxyribonucleic acid, made up of 4 nucleotide bases**
  - **Adenine (A)**
  - **Cytosine (C)**
  - **Guanine (G)**
  - **Thymine (T)**



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



Genes and Genome

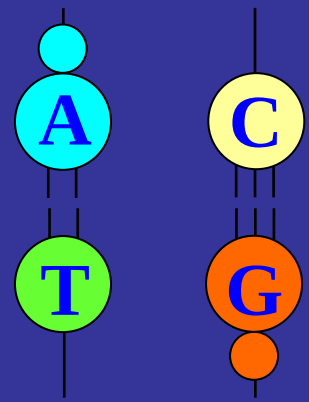


DNA and RNA

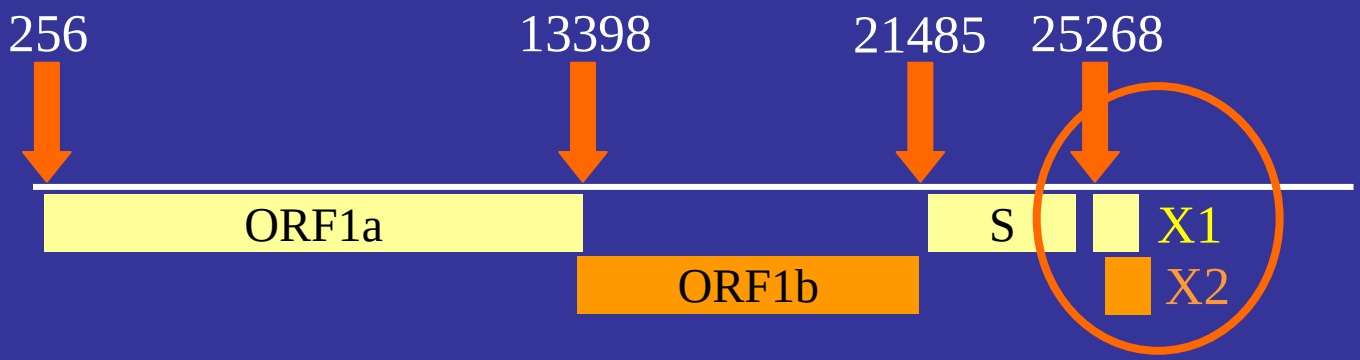
DNA is deoxyribonucleic acid, made up of 4 nucleotide bases Adenine, Cytosine, Guanine, and Thymine.

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

TCTTTAACAAGCTTGTTAAAGA

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:

“A nut for a jar of tuna”

remove spaces and capitalize

ANUTFORAJAROFTUNA

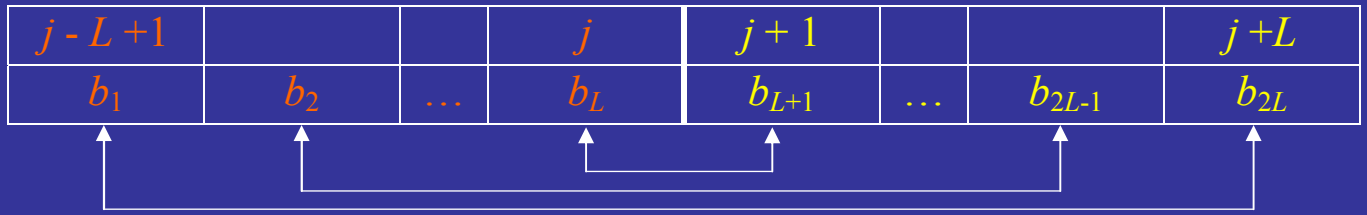
Even Palindrome:

“Step on no pets”

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'



We say that a palindrome of length $2L$ occurs at position j when the $(j\text{-}i+1)$st and the $(j+i)$th bases are complementary to each other for $i=1,\ldots, L$. 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

$$p = \left[2(\hat{p}_A \hat{p}_T + \hat{p}_C \hat{p}_G)\right]^{11}$$

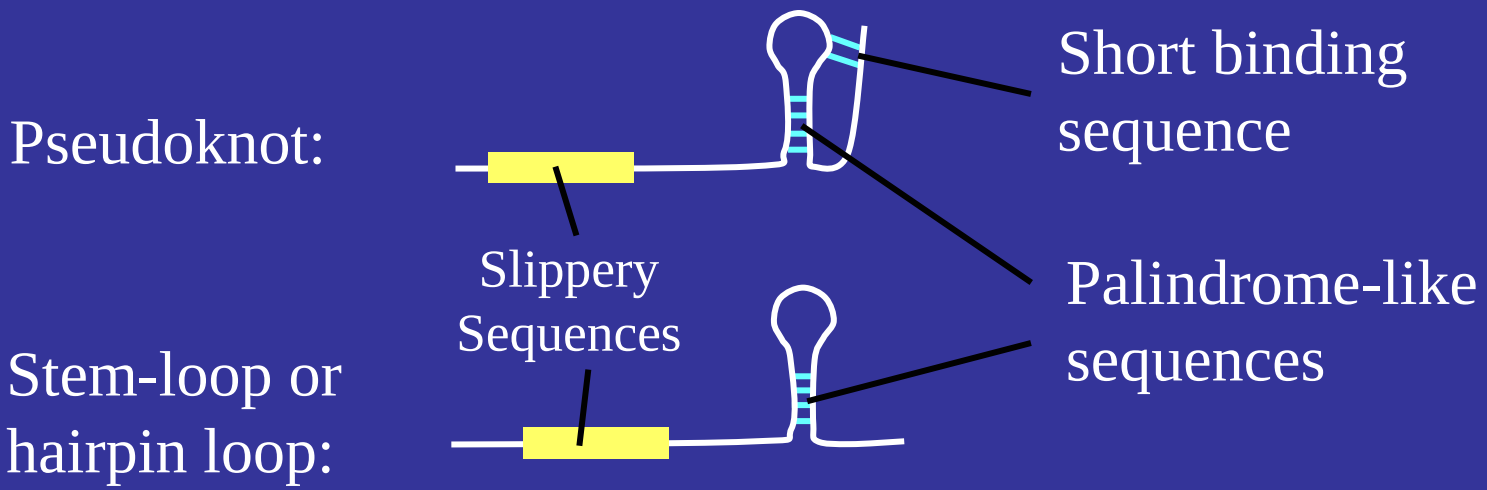
- The probability of the occurrence of at least one length 22 palindrome in the genome is

$$1 - e^{-0.01008} \approx 0.01$$

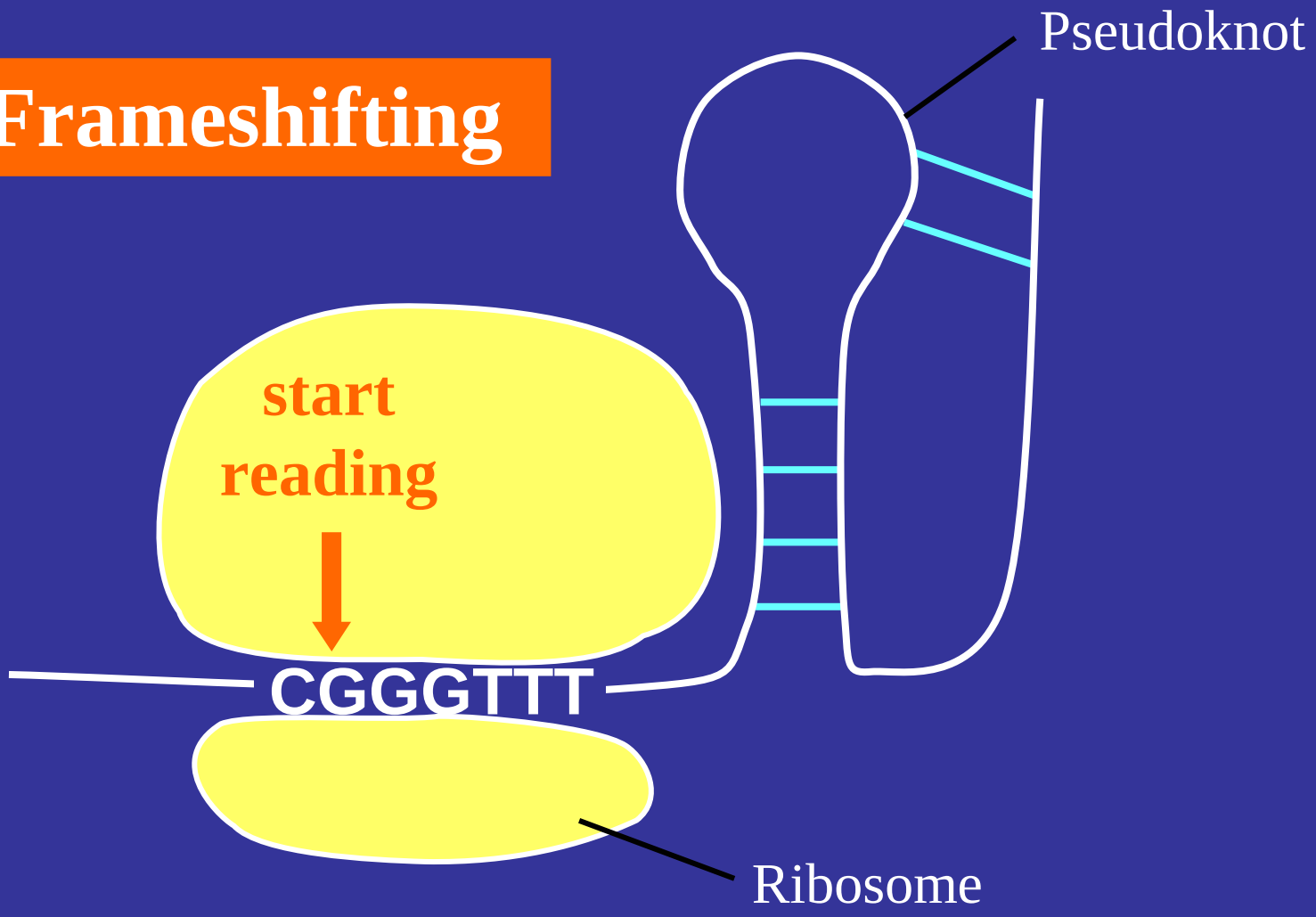
Expression of Overlapping Genes Requires Frameshifting in Reading

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

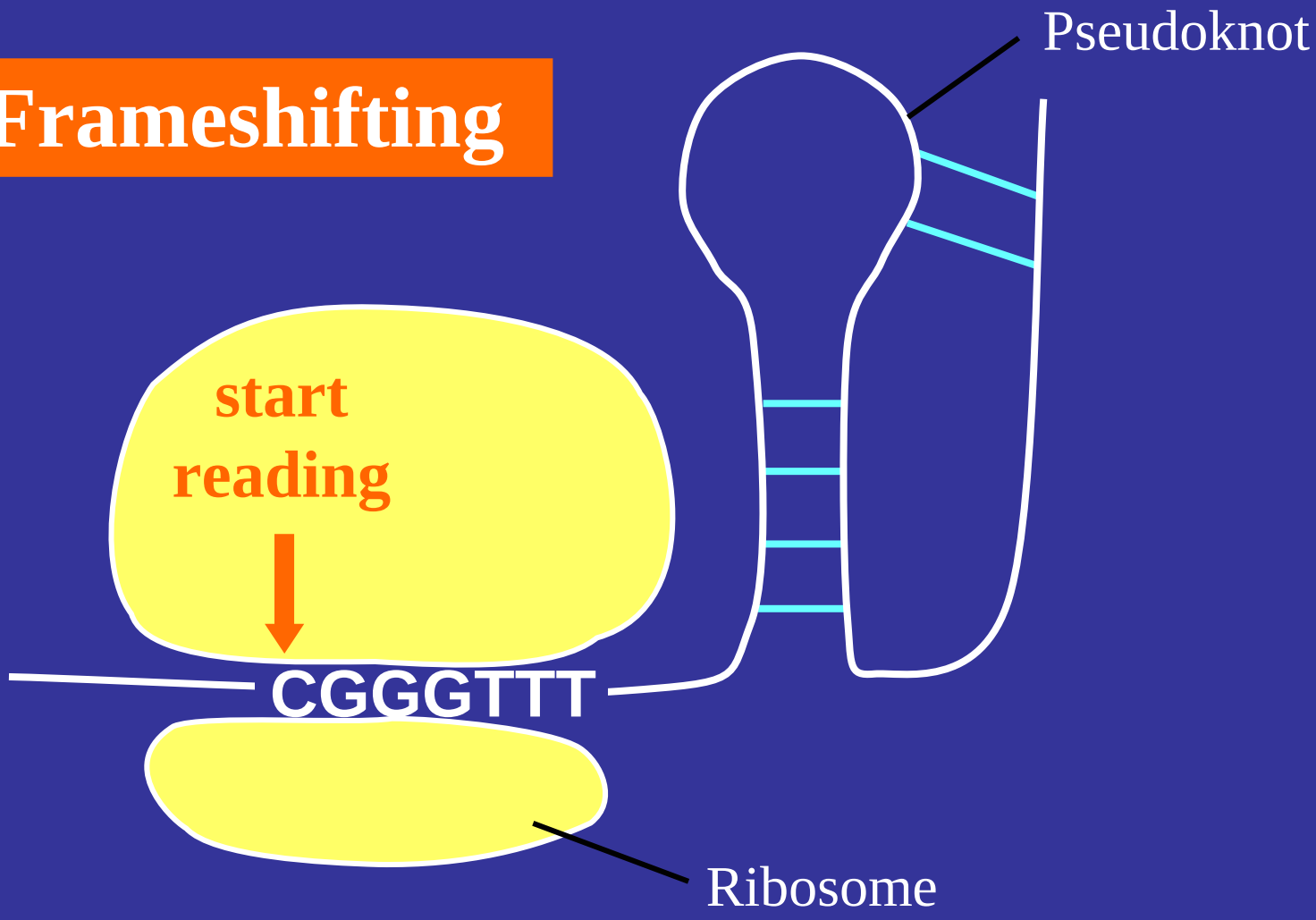


-1 Frameshifting



- 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 XXXYYYN
where X = A,T, or G; Y = A or T; and N = A, T, or C

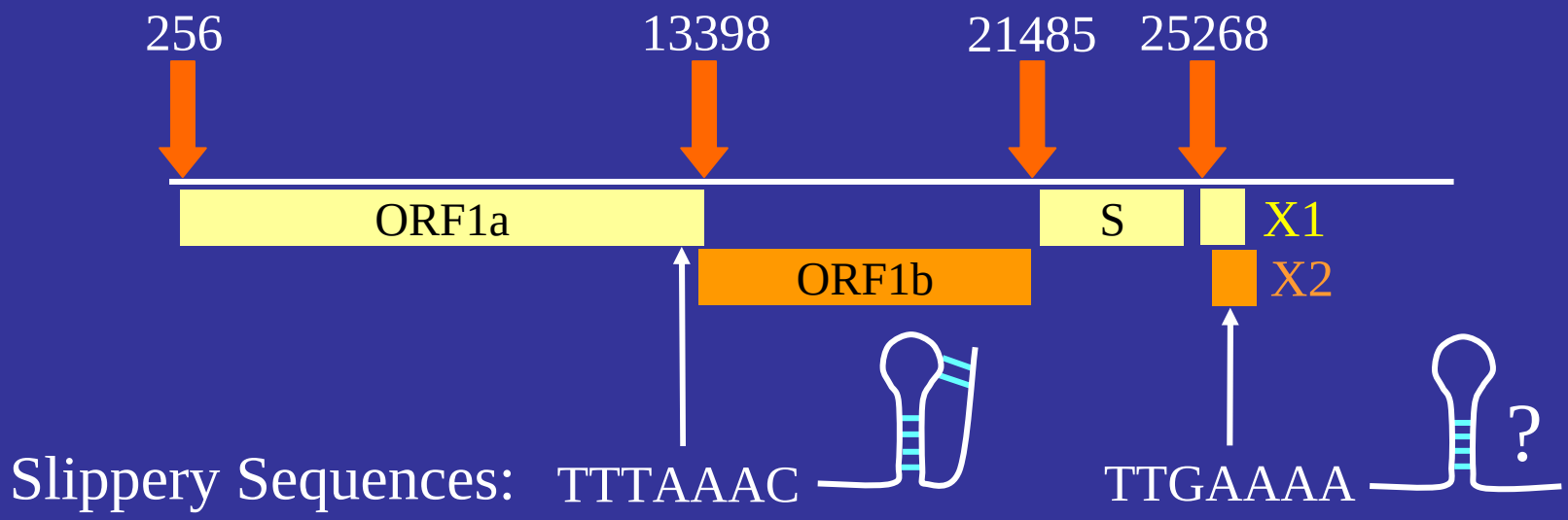
- ORF1a and ORF1b (Overlap: 13398, 1 base only)

13201	CATCCAAATC	CTAAAGGATT	CTGTGACTTG	AAAGGTAAGT	ACGTCCAAAT
13251	ACCTACCACT	TGTGCTAATG	ACCCAGTGGG	TTTTACACTT	AGAAACACAG
13301	TCTGTACCGT	CTGCGGAATG	TGGAAAGGTT	ATGGCTGTAG	TTGTGACCAA
13351	CTCCGCGAAC	CCTTGATGCA	GTCTGCGGAT	GCATCAACGT	TTTTAAACGG
13401	GTTTGCGGTG	TAAGTGCAGC	CCGTCTTACA	CCGTGCGGCA	CAGGCACTAG
13451	TACTGATGTC	GTCTACAGGG	CTTTTGATAT	TTACAACGAA	AAAGTTGCTG
13501	GTTTTGCAAA	GTTCCTAAAA	ACTAATTGCT	GTCGCTTCCA	GGAGAAGGAT
13551	GAGGAAGGCA	ATTTATTAGA	CTCTTACTTT	GTAGTTAAGA	GGCATACTAT

- X1 and X2 (Overlap: 25689-26089, 401 bases)

25651	GCTTTGTTGG	AAGTGCAAAT	CCAAGAACCC	ATTACTTTAT	GATGCCAACT
25701	ACTTTGTTTG	CTGGCACACA	CATAACTATG	ACTACTGTAT	ACCATATAAC
25751	AGTGTCACAG	ATACAATTGT	CGTTACTGAA	GGTGACGGCA	TTTCAACACC
25801	AAAACTCAAA	GAAGACTACC	AAATTGGTGG	TTATTCTGAG	GATAGGCACT
25851	CAGGTGTTAA	AGACTATGTC	GTTGTACATG	GCTATTTCAC	CGAAGTTTAC
25901	TACCAGCTTG	AGTCTACACA	AATTACTACA	GACACTGGTA	TTGAAAATGC
25951	TACATTCTTC	ATCTTTAACA	AGCTTGTTAA	AGACCCACCG	AATGTGCAAA
26001	TACACACAAT	CGACGGCTCT	TCAGGAGTTG	CTAATCCAGC	AATGGATCCA
26051	ATTTATGATG	AGCCGACGAC	GACTACTAGC	GTGCCTTTGT	AAGCACAAGA
26101	AAGTGAGTAC	GAACTTATGT	ACTCATTCGT	TTCGGAAGAA	ACAGGTACGT

Locations of 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

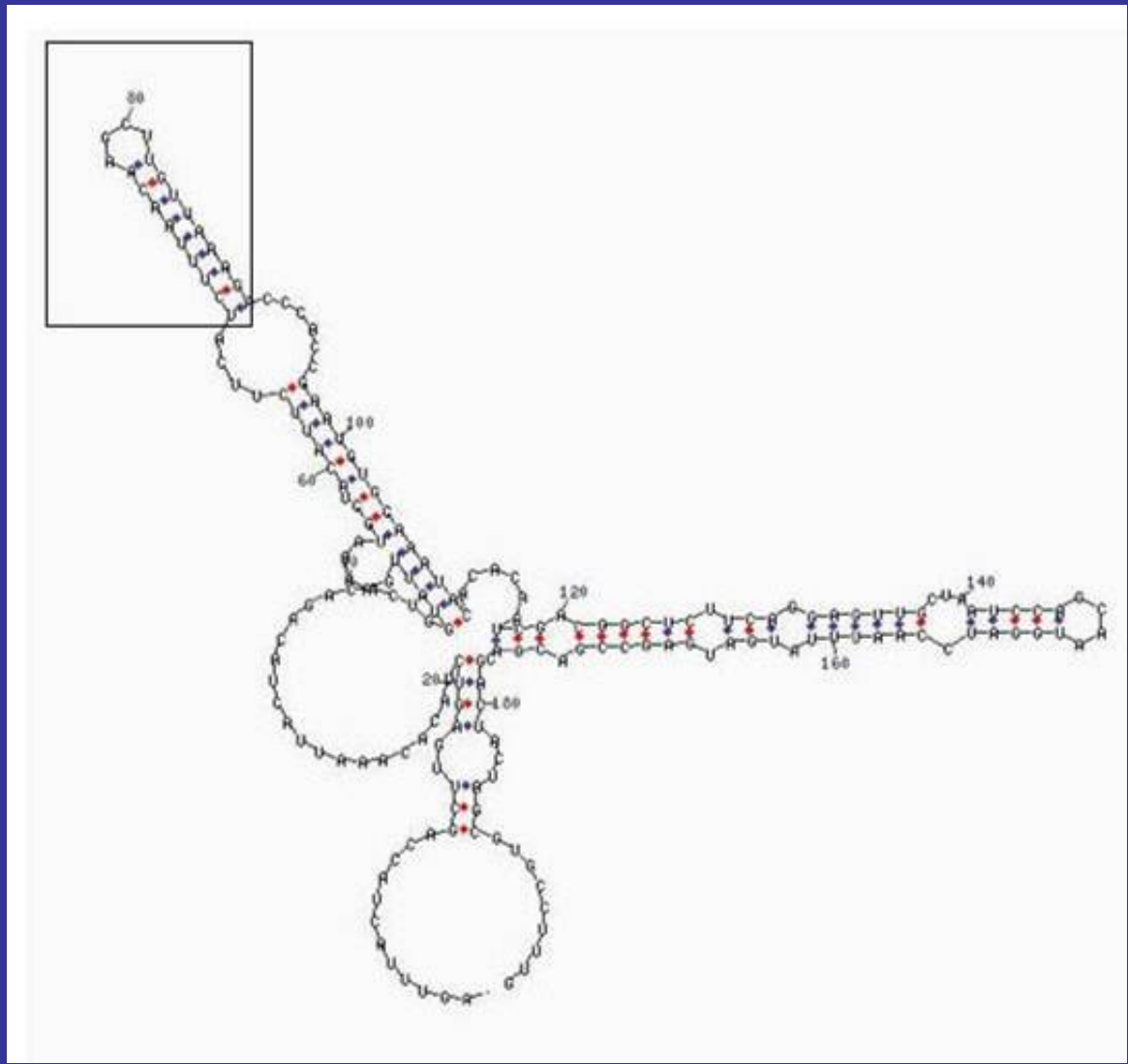
Program	Website
Mfold	http://www.bioinfo.rpi.edu/applications/mfold/old/rna/
Pknots*	http://www.genetics.wustl.edu/eddy/software/#pk
KineFold	http://kinefold.u-strasbg.fr/
RNAFold	http://www.tbi.univie.ac.at/~ivo/RNA/
PknotsRG	http://bibiserv.techfak.uni-bielefeld.de/bibi/Tools_RNA_Studio.html

*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*



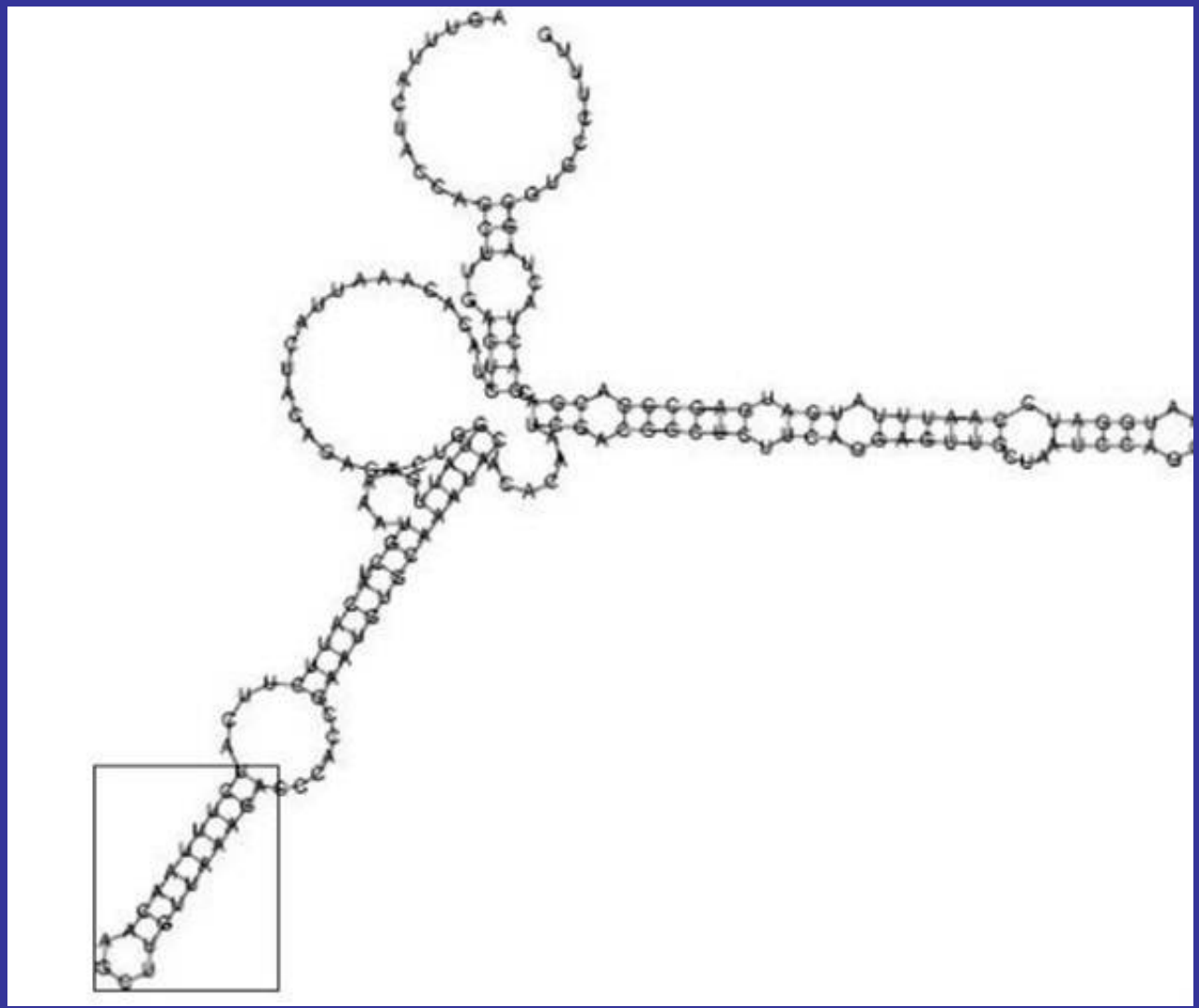
ΔG = -55.2 kcal/mol.

Hairpin Loop Predicted by *Pknots*

```
PKNOTS: optimal minimum-energy RNA folding with pseudoknots and coaxial energies
        PKNOTS 1.01 (JUN 2000) using squid 1.5m (Sept 1997)
-------------------------------------------------
Folding sequences from:  overlap.txt
-------------------------------------------------

       1 A   G   U   U   U   A   C   U   A   C   C   A   G   C   U   U   G   A   G   U
         0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
         .   .  51  50  49  48   .  46  45  44  43  42  41   .   .   .   .   .  39  38

      21 C   U   A   C   A   C   A   A   A   U   U   A   C   U   A   C   A   G   A   C
        20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
        37  36   .   .   .   .   .   .   .   .   .   .   .   .   .   .  21  20  19  18

      41 A   C   U   G   G   U   A   U   U   G   A   A   A   A   U   G   C   U   A   C
        40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59
         .  12  11  10   9   8   7   .   5   4   3   2   .   .  104 103 102   . 101 100

      61 A   U   U   C   U   U   C   A   U   C   U   U   U   A   A   C   A   A   G   C
        60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79
        98  98  97  96   .   .   .   .  88  88  87  86  85  84  83  82  81   .   .   .

      81 U   U   G   U   U   A   A   A   G   A   C   C   C   A   C   C   G   A   A   U
        80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99
         .  76  75  74  73  72  71  70  69  68   .   .   .   .   .   .  63  62  61  60

     101 G   U   G   C   A   A   A   U   A   C   A   C   A   C   A   A   U   C   G   A
       100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
        59  58  56  55  54   .   .   .   .   .   .   .   .   .   .   .  175 174 173   .

     121 C   G   G   C   U   C   U   U   C   A   G   G   A   G   U   U   G   C   U   A
       120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139
       171 170 169 168 167 166   . 164 163 162   . 160 159 158 157 156 155   .   .   .

     141 A   U   C   C   A   G   C   A   A   U   G   G   A   U   C   C   A   A   U   U
       140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159
       153 152 151 150 149   .   .   .  144 143 142 141 140   . 136 135 134 133 132

     161 U   A   U   G   A   U   G   A   G   C   C   G   A   C   G   A   C   G   A   C
       160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179
       131   . 129 128 127   . 125 124 123 122 121 120   . 118 117 116 195 194 193   .

     181 U   A   C   U   A   G   C   G   U   G   C   C   U   U   U   G
       180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195
       189 188 187   .   .   .  182 181 180   .   .  178 177 176
```

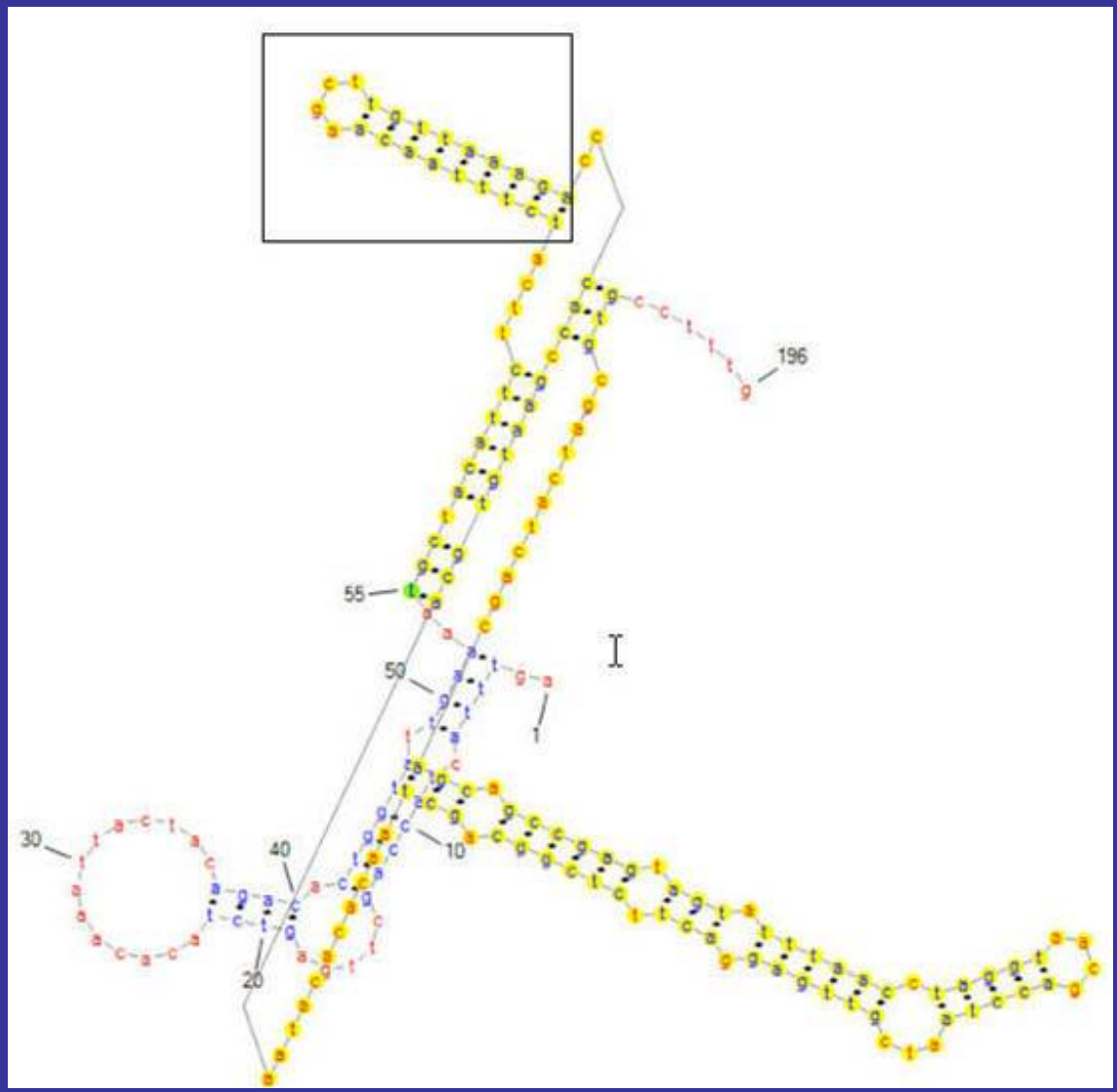
ΔG = -54.8 kcal/mol.

Hairpin Loop Predicted by *RNAFold*



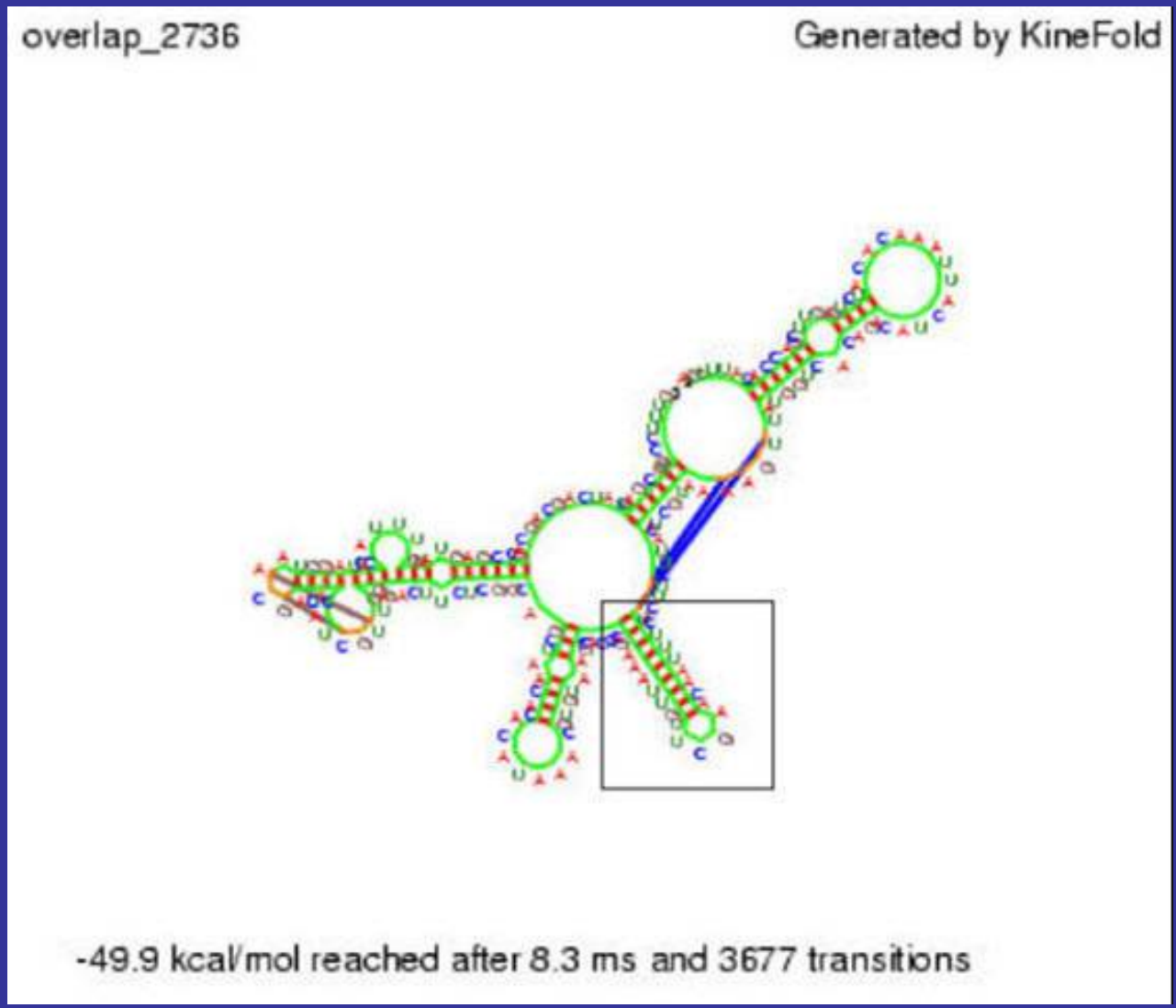
ΔG = -58.75 kcal/mol.

Pseudoknot Predicted by *PknotsRG*



ΔG = -55.84 kcal/mol.

Pseudoknot Predicted by *KineFold*



ΔG = -49.9 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 ΔG = -58.75 kcal/mol)
- *PknotsRG* predicts that the palindrome is part of a pseudoknot (2nd lowest ΔG = -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?