Central dogma of molecular biology

• Information is stored in DNA

• Genome is processed into messenger RNA molecules (transcription)

• RNA molecules are processed to form proteins (translation)
Open reading frames

• Generally defined as regions in genes between a start (ATG) and stop (eg. TGA) codon.

• Size is a multiple of 3

• Six possibilities given any DNA sequence
  – 0 offset, + strand; 1 offset, + strand, 2 offset, + strand
  – 0 offset, - strand; 1 offset, - strand, 2 offset, - strand
Long ORFs

- At random, we’d expect a stop codon every 64 nucleotides.
- Many bacteria genes are much longer than this.
- These can be used to train a statistical model.
IMM

• Interpolated Markov Models (IMMs) overcome the training problem by generating models of variable order.

• Bias is put towards higher models if and only if there is enough training data.

• Achieved via a linear combination of probabilities based on varied lengths.
Simple linear interpolation

\[ P_{\text{IMM}}(x_i \mid x_{i-n},...,x_{i-1}) = \lambda_0 P(x_i) \]
\[ + \lambda_1 P(x_i \mid x_{i-1}) \]
\[ ... \]
\[ + \lambda_n P(x_i \mid x_{i-n},...,x_{i-1}) \]

where \( \sum_i \lambda_i = 1 \)
GLIMMER

- Addressed the fundamental training problem of markov models

- As mentioned before, we want the highest order model possible

- However, a $k^{th}$ order model requires $4^{k+1}$ probabilities to be estimated
  - Impractical for small genomes
Table 1. Comparison of the IMM model used in GLIMMER to a 5th-order Markov model

<table>
<thead>
<tr>
<th>Model</th>
<th>Genes found</th>
<th>Genes missed</th>
<th>Additional genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLIMMER IMM</td>
<td>1680 (97.8%)</td>
<td>37</td>
<td>209</td>
</tr>
<tr>
<td>5th-Order Markov</td>
<td>1574 (91.7%)</td>
<td>143</td>
<td>104</td>
</tr>
</tbody>
</table>

The first column indicates how many of the 1717 annotated genes in *H. influenzae* were found by each algorithm. The ‘additional genes’ column shows how many extra genes, not included in the 1717 annotated entries, were called genes by each method.

Only 1 of the 37 genes missed by GLIMMER was found by the 5th order model

On the other hand, it found 107 more true genes
Pipmaker

- [http://www.bx.psu.edu/miller_lab/](http://www.bx.psu.edu/miller_lab/)

- Visualization of BLASTZ alignments

- Although pips are compact and informative, they do not show alignment information for the second sequence.
  - Dotplots are used to see relevant differences