GENOME ANNOTATION:
ADVANCED (EUKARYOTIC)
GENE PREDICTION

BIO 300/CMPSC 300
Dr. Kristen Webb
Spring 2016
Prediction Algorithms

- Alignment-based – sequence similarity to previously identified gene in another organism (BLAST)

- Sequence-based – search for specific sequences - e.g. ORF-finder – searches start and stop codons

- Content-based – search for patterns – e.g. nucleotide or codon frequency

- Probabilistic – combination of sequence- and content-based plus probability that sequence is part of a gene
Prokaryotic Gene Structure  Eukaryotic Gene Structure
Prokaryotic Gene Structure
Prokaryotic Gene Structure

- Highly conserved sequences 10 and 35 bases upstream (before) start codon
- Highly conserved Shine-Dalgarno sequence immediately before start codon

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Consensus (5' → 3')</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prokaryotes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>−10 sequence</td>
<td>TATAAT</td>
<td>RNA polymerase binds to start transcription</td>
</tr>
<tr>
<td>−35 sequence</td>
<td>TTGACA 17±2 from −10</td>
<td>RNA polymerase binds to start transcription</td>
</tr>
<tr>
<td>Shine-Dalgarno</td>
<td>AGGAGG 5±2 from ATG</td>
<td>Ribosome binds to find start codon</td>
</tr>
</tbody>
</table>

(A) prokaryotes
Prokaryotic Gene Structure

- Highly conserved sequences 10 and 35 bases upstream (before) start codon
- Highly conserved Shine-Dalgarno sequence immediately before start codon
- Genes rarely contain introns
  - Present as ORFs (start codon through stop codon – all protein-coding sequence)
Prokaryotic Gene Structure

- conserved -10 and -35 promoter sequences
- conserved Shine-Dalgarno sequence marks the start codon
- few interrupted ORFs (introns are rare)

Eukaryotic Gene Structure
Eukaryotic Gene Structure

- Variable promoter structure
  - not all promoter elements present in all gene
  - promoter element sequence can vary between genes
Eukaryotic Gene Structure

- Variable promoter structure
  - not all promoter elements present in all gene
  - promoter element sequence can vary between genes
  - no conserved Shine-Delgarno-like sequence
Eukaryotic Gene Structure

- Variable promoter structure
  - not all promoter elements present in all gene
  - promoter element sequence can vary between genes
  - no conserved Shine-Delgarno-like sequence

- makes identifying the start of the gene more difficult
Eukaryotic Gene Structure

- Most genes contain introns

- Only interested in coding region
  - Exons only
    - Sequence with codons
Eukaryotic Gene Structure

- Exon/Intron boundaries not highly conserved
Eukaryotic Gene Structure

- Exon/Intron boundaries not highly conserved
Eukaryotic Gene Structure

http://weblogo.berkeley.edu/logo.cgi

- Exon/Intron boundaries not highly conserved
<table>
<thead>
<tr>
<th>Prokaryotic Gene Structure</th>
<th>Eukaryotic Gene Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>• -10 and -35 promoter sequences</td>
<td>• Promoter sequences vary in number and sequence</td>
</tr>
<tr>
<td>• Shine-Dalgarno sequence marks the start codon</td>
<td>• No Shine-Dalgarno – unambiguous identification of transcriptional start site is difficult</td>
</tr>
<tr>
<td>• Few interrupted ORFs (introns are rare)</td>
<td>• Nearly all genes contain introns</td>
</tr>
<tr>
<td></td>
<td>• Intron/exon boundaries are hard to discern</td>
</tr>
</tbody>
</table>
Bioinformatics Solution?

• Content- and Probability-Based Gene Prediction
Bioinformatics Solution?

- Content- and Probability-Based Gene Prediction

- Content-based gene prediction
  - Alone not very precise
  - Better result when used in combination with other algorithms
Bioinformatics Solution?

- Content- and Probability-Based Gene Prediction

- Content-based gene prediction
  - Alone not very precise
  - Better result when used in combination with other algorithms

- Content
  - Codon usage
  - CpG Islands
**Codon Usage**

synonymous codons for same amino acid not used with equal frequency

<table>
<thead>
<tr>
<th>Codon</th>
<th>E. coli K12 Fraction</th>
<th>Homo sapiens Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>0.57</td>
<td>0.46</td>
</tr>
<tr>
<td>UUC</td>
<td>0.43</td>
<td>0.54</td>
</tr>
<tr>
<td>UUA</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>UUG</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>CUU</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>CUC</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>CUA</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>CUG</td>
<td>0.46</td>
<td>0.40</td>
</tr>
<tr>
<td>AUA</td>
<td>0.07</td>
<td>0.17</td>
</tr>
<tr>
<td>AUG</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>GUA</td>
<td>0.17</td>
<td>0.24</td>
</tr>
<tr>
<td>GUG</td>
<td>0.40</td>
<td>0.46</td>
</tr>
<tr>
<td>GCU</td>
<td>0.11</td>
<td>0.27</td>
</tr>
<tr>
<td>GCC</td>
<td>0.18</td>
<td>0.24</td>
</tr>
<tr>
<td>GCA</td>
<td>0.21</td>
<td>0.12</td>
</tr>
<tr>
<td>GCG</td>
<td>0.38</td>
<td>0.46</td>
</tr>
<tr>
<td>UAU</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>UAC</td>
<td>0.47</td>
<td>0.45</td>
</tr>
<tr>
<td>UAA</td>
<td>0.64</td>
<td>0.30</td>
</tr>
<tr>
<td>UAG</td>
<td>0.00</td>
<td>0.24</td>
</tr>
<tr>
<td>UGU</td>
<td>0.42</td>
<td>0.47</td>
</tr>
<tr>
<td>UGC</td>
<td>0.58</td>
<td>0.56</td>
</tr>
<tr>
<td>UGA</td>
<td>0.36</td>
<td>0.47</td>
</tr>
<tr>
<td>UGG</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

[Codon/a.a./fraction per codon per a.a.]
E. coli K12 data from the Codon Usage Database

Homo sapiens data from the Codon Usage Database
Codon Usage

some amino acids are much more common in proteins than others
Using Codon Freq to find Exon/Intron Boundaries

• Expectations:
  
  • Exon – codon frequency closely matches expected frequency for a gene
  • Intron – “codon” frequency poorly matches expected frequency for a gene (because not really codons!!)
  • Boundary – point where frequencies shift
Using Codon Freq to find Exon/Intron Boundaries

- Sliding-Window Approach

As the sliding window advances, the slice of its input data changes. Here the algorithm uses the current sliding window data to compute the sum of the window’s elements.
Using Codon Freq to find Exon/Intron Boundaries

- **CBI** = codon bias index – compares usage of most common codons to random occurrence of those codons
  - 0 = random codon usage
  - 1 = exclusive usage of preferred codons

---

- **Window 1A**
  - CBI = 0.77
  - Putative exon-intron boundary

- **Window 1B**
  - CBI = 0.18

- **Window 2A**
  - CBI = 0.16

- **Window 2B**
  - CBI = 0.21
Finding Promoters using CpG Islands

- Promoter regions tend to have a higher frequency of C and G nucleotides relative to A and T nucleotides.
- The CG dinucleotide occurs in promoter regions more frequently than would be expected by chance.
- CpG targets for methylation and epigenetic regulation of gene expression.
Finding Promoters using CpG islands

**Left:** CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG constitutes the start codon.

**Right:** CpG sites present at every 1/100 nucleotides, constituting a more normal example of the genome - a non-coding region.
Finding Promoters using CpG islands

- Sliding-window approach + pattern matching algorithm
  - Just one window

\[
\text{CpG ratio: } \frac{\text{observed CG pairs}}{C \text{ nucleotides } \times G \text{ nucleotides/total nucleotides}}
\]