GENOME ANNOTATION: SEQUENCE-BASED GENE PREDICTION

BIO 300/CMPSC 300
Dr. Kristen Webb
Spring 2016
Genome Projects

• Goals:
  • Determine complete genome sequence of an organism
Genome Projects

• Goals:
  • Determine complete genome sequence of an organism
  • Annotate protein-coding genes and other important genome-encoded features
    • find
    • identify
    • characterize
    • describe
Genome Projects

• Goals:
  • Determine complete genome sequence of an organism
  • Annotate protein-coding genes and other important genome-encoded features
    • find
    • identify
    • characterize
    • describe
  • computational predictions later confirmed at the lab bench
Prediction Algorithms

• Alignment-based – find genes/features based on conserved sequences is well-studied organisms (database searching)

  • Automatic assignment based on sequence similarity (best BLAST hit): gene name, protein name, function

  • Quality vs Quantity
Prediction Algorithms

• Sequence-based – find features based on specific sequences
Prediction Algorithms

• Sequence-based – find features based on specific sequences
Prediction Algorithms

- Content-based – consider overall properties of the sequence when making predictions
  - nucleotide frequency
  - codon frequency/codon bias

GC Content for all *V. vulnificus* and *V. naverensis* gene predictions
Prediction Algorithms

- Content-based – consider overall properties of the sequence when making predictions
  - nucleotide frequency
  - codon frequency/codon bias

GC Content for all *V. vulnificus* and *V. naverensis* gene predictions - Most of the genomes contained a high percentage of genes with GC contents between 45-50%.
Prediction Algorithms

• Probabilistic – combination of sequence-based and content-based plus probability
  • “annotation pipeline”
NCBI Prokaryotic Annotation Pipeline

- Combines sequence-based algorithm with alignment-based approach
  - Protein-coding genes
  - Structural RNAs (5S, 16S, 23S)
  - Transfer RNAs
  - Small non-coding RNAs
- Rely only on properties of DNA and training set of genes

NCBI Eukaryotic Annotation Pipeline

NCBI Eukaryotic Annotation Pipeline

1. **masking**
   - try to identify and ignore non-coding regions

2. **alignment-based predictions**

3. **sequence/content-based predictions from alignment-based**

4. **best selected (probability), named, and released**

Prokaryotic versus Eukaryotic Genomes

- **Prokaryotes**
  - 1 circular chromosome
  - “genome”
  - extra DNA in plasmids
  - smaller, self-replicating
Prokaryotic versus Eukaryotic Genomes

- **Eukaryotes**
  - Multiple linear chromosomes
  - “genome”
- Extra DNA in mitochondria/chloroplast
# Prokaryotic versus Eukaryotic Genomes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Amount of DNA (bp)</th>
<th># of genes</th>
<th>Genes per million bases</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli</em></td>
<td>4,600,000</td>
<td>4,400</td>
<td>950</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>12,000,000</td>
<td>5,800</td>
<td>480</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>180,000,000</td>
<td>13,700</td>
<td>76</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>2,600,000,000</td>
<td>25,000</td>
<td>11</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>2,900,000,000</td>
<td>25,000</td>
<td>10</td>
</tr>
</tbody>
</table>
Need to know feature structure

Comparison of Prok vs Euk transcription unit structures
## Consensus Sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Consensus (5’ → 3’)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prokaryotes</strong></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>
<tr>
<td><strong>Eukaryotes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TATA box</td>
<td>TATAWAW</td>
<td>Core promoter; binds TFIID</td>
</tr>
<tr>
<td>Inr sequence</td>
<td>YYCARR</td>
<td>Core promoter; contains +1 sequence (C)</td>
</tr>
<tr>
<td>GC box</td>
<td>GGGCGG</td>
<td>Transcription factor binding site</td>
</tr>
<tr>
<td>CAT box</td>
<td>CAAT</td>
<td>Transcription factor binding site</td>
</tr>
<tr>
<td>Kozak consensus</td>
<td>gccRccATGG</td>
<td>Context of start codon</td>
</tr>
<tr>
<td>5’ splice site</td>
<td>MAG</td>
<td>GTragt</td>
</tr>
<tr>
<td>3’ splice site</td>
<td>cAG</td>
<td>G</td>
</tr>
<tr>
<td>intron branch  site</td>
<td>CTRAY</td>
<td>3’ end of intron binds to mark for degradation</td>
</tr>
<tr>
<td>polyadenylation site</td>
<td>AAUAAA</td>
<td>Cleavage of mRNA for poly(A) tail</td>
</tr>
</tbody>
</table>
## IUB Ambiguity Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Base Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>Guanine</td>
<td>(A or G)</td>
</tr>
<tr>
<td>A</td>
<td>Adenine</td>
<td>(C or T or U)</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
<td>(A or C)</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
<td>(G or T)</td>
</tr>
<tr>
<td>R</td>
<td>Purine</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Pyrimidine</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Amino</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Ketone</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Strong interaction</td>
<td>(C or G)</td>
</tr>
<tr>
<td>W</td>
<td>Weak interaction</td>
<td>(A or T)</td>
</tr>
<tr>
<td>H</td>
<td>Not-G</td>
<td>(A or C or T)</td>
</tr>
<tr>
<td>B</td>
<td>Not-A</td>
<td>(C or G or T)</td>
</tr>
<tr>
<td>V</td>
<td>Not-T (not-U)</td>
<td>(A or C or G)</td>
</tr>
<tr>
<td>D</td>
<td>Not-C</td>
<td>(A or G or T)</td>
</tr>
<tr>
<td>N</td>
<td>Any</td>
<td>(A or C or G or T)</td>
</tr>
</tbody>
</table>
Open Reading Frame

(A) prokaryotes

DNA

-35  -10  +1  Shine-Dalgarno  gene A  gene B  gene C

operon

coding sequence (ORF)

(B) eukaryotes

DNA

transcription factor sites

TATA box

Lnr +1

core promoter

exon 1  intron 1  exon 2  intron 2  exon 3

coding sequence

ATG

TAG
Open Reading Frame (ORF)
Class Activity: NCBI – ORF Finder

• Use NCBI ORF Finder to annotate a plasmid
• Describe generic pattern-matching algorithm using ORF finder as an example