Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

From genes to proteins

Encyclopedia of DNA Elements (ENCODE)

Gene definitions

• Definition 1910s: Gene as a distinct locus
• Definition 1940s: Gene as a blueprint for a protein
• Definition 1950s: Gene as a physical molecule
• Definition 1960s: Gene as transcribed code
• Definition 1970s–1980s: Gene as open reading frame (ORF) sequence pattern
• Definition 1990s–2000s: Annotated genomic entity, enumerated in the databanks (current view, pre-ENCODE)
• A current computational metaphor: Genes as “subroutines” in the genomic operating system

Genomic region (ENCODE)
Gene concept problems

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Description</th>
<th>Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene structure and function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intronic genes</td>
<td>A gene may contain an exon of more than one ORF.</td>
<td>Gene boundary issues</td>
</tr>
<tr>
<td>DNA regions</td>
<td>DNA regions may be two different from each other, but are adjacent.</td>
<td>DNA sequence comparison between ORF and protein sequence</td>
</tr>
<tr>
<td>Defective regulatory variants</td>
<td>Epistatic interactions between ORF and regulatory sequence.</td>
<td>Gene regulation issues</td>
</tr>
<tr>
<td>Expressed variant</td>
<td>Allelic variation</td>
<td>Allelic variation issues</td>
</tr>
<tr>
<td>Gene transposition variants</td>
<td>DNA transposition variants</td>
<td>DNA transposition issues</td>
</tr>
<tr>
<td>Copy number variants</td>
<td>Copy number variation in human DNA</td>
<td>Copy number variation issues</td>
</tr>
<tr>
<td>Epigenetic and chromatin structure</td>
<td>Epigenetic modifications</td>
<td>Epigenetic modifications issues</td>
</tr>
</tbody>
</table>

ENCODE definition of gene

- DNA
  - Primary transcripts
  - Spliced transcripts
  - Functional products
    - Genes

Chromosome 19 gene map

Computational Gene Prediction

- Where the genes are unlikely to be located?
- How do transcription factors know where to bind a region of DNA?
- Where are the transcription, splicing, and translation start and stop signals?
- What does coding region do (and non-coding regions do not)?
- Can we learn from examples?
- Does this sequence look familiar?
Computational Gene Prediction

**Spliced Alignment (Procrustes)**

- New genomic sequence
- Selection of candidate exons
  - AUG → GU initial exons
  - AG → GU internal exons
  - AG → UAA or UAG or UGA terminal exons
- Filtration (based on the codon usage statistics)
- Construction of all possible chains of candidate exons
- Finding a chain with the maximum global similarity to the target protein

**Predicted Exon Assembly (Procrustes)**
PCR Primers Prediction (GenePrimer)

Exon 1085..1182 (98) hit using first 2 primers
Exon 1628..1676 (49) missed
Exon 1900..2001 (102) hit using first 8 primers
Exon 2110..2184 (75) missed
Exon 2516..2722 (207) hit using first 4 primers
Exon 3385..3472 (88) missed
Exon 3546..3746 (201) hit using first primer
...

Suboptimal Solutions for the Human Growth Hormone Gene (GeneParser)

Measures of Prediction Accuracy

Nucleotide Level

<table>
<thead>
<tr>
<th>REALITY</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDICTION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity

\[ S_n = \frac{TP}{(TP + FN)} \]

Specificity

\[ S_p = \frac{TP}{(TP + FP)} \]

Exon Level

<table>
<thead>
<tr>
<th>REALITY</th>
<th>WRONG EXON</th>
<th>CORRECT EXON</th>
<th>MISSING EXON</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDICTION</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity

\[ S_n = \frac{\text{number of correct exons}}{\text{number of actual exons}} \]

Specificity

\[ S_p = \frac{\text{number of correct exons}}{\text{number of predicted exons}} \]
Computational Gene Prediction

<table>
<thead>
<tr>
<th>Tool</th>
<th>Start</th>
<th>End</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpliceView</td>
<td>1573</td>
<td>2085</td>
<td>512</td>
</tr>
<tr>
<td>NNetGenez</td>
<td>1600</td>
<td>1763</td>
<td>163</td>
</tr>
<tr>
<td>BDGP-NNSIP</td>
<td>1680</td>
<td>1763</td>
<td>83</td>
</tr>
<tr>
<td>SPl</td>
<td>1680</td>
<td>1763</td>
<td>83</td>
</tr>
</tbody>
</table>

GeneMark Accuracy Evaluation

Errors in genome annotation