Genome 373: Gene Prediction 1

Doug Fowler
Outline

• Review of gene structure

• Scale of the problem

• Solutions
  – Empirical methods
  – *Ab initio* prediction
What is a gene?

Gene Structure Varies by Organism

- Eubacteria, archae, and single-celled fungi have simple genes:
  - No or few introns (and these are short)
  - Transcription start and stop sites relatively well-defined in primary sequence.

- Plants and animals have complex genes:
  - Many introns (and these can be long)
  - Transcription start/stop sites are poorly defined (hard to find) in primary sequence
  - Intron/exon boundaries are poorly defined in primary sequence
A Sample from *S. cerevisiae* (yeast)

No introns, genes tightly packed.

(there actually are introns in yeast, but very few and none for these particular genes)
A Sample from *C. elegans* (worm)

10 kb region

Note many introns, variable length.
A Sample from *Homo sapiens*

- Large number of long introns, highly variable length
- (note different scale from previous slide)
- A single gene, CFTR (defects cause cystic fibrosis):
Outline

• Review of gene structure

• Scale of the problem

• Solutions
  – Empirical methods
  – Ab initio prediction
# How Big are Genomes?

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (Mb)</th>
<th># of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epstein-Barr virus (EBV)</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>Yeast</td>
<td>12.5</td>
<td></td>
</tr>
<tr>
<td>Worm</td>
<td>100.2</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>3,300</td>
<td></td>
</tr>
</tbody>
</table>

**Guesses about the number of genes?**

[http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html](http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html)
### What About Introns?

Humans have many introns, relative to most other organisms.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genomic size (Mb)</th>
<th>Number of introns per gene</th>
<th>Size of individual introns</th>
<th>Total intron size per kb cds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All genes</td>
<td>Homologous genes</td>
<td>All genes</td>
</tr>
<tr>
<td>Human</td>
<td>3400</td>
<td>4.0</td>
<td>5.54</td>
<td>3413.4</td>
</tr>
<tr>
<td>Mouse</td>
<td>3454</td>
<td>3.1</td>
<td>6.68</td>
<td>1321.4</td>
</tr>
<tr>
<td>Rat</td>
<td>2900</td>
<td>3.0</td>
<td>3.77</td>
<td>1091.7</td>
</tr>
<tr>
<td>Chicken</td>
<td>1200</td>
<td>3.2</td>
<td>3.69</td>
<td>706.3</td>
</tr>
<tr>
<td>Drosophila</td>
<td>180</td>
<td>2.5</td>
<td>2.44</td>
<td>563.9</td>
</tr>
<tr>
<td>C. elegans</td>
<td>100</td>
<td>4.2</td>
<td>4.26</td>
<td>466.6</td>
</tr>
<tr>
<td>Cress</td>
<td>100</td>
<td>4.8</td>
<td>4.10</td>
<td>239.7</td>
</tr>
<tr>
<td>Corn</td>
<td>5000</td>
<td>4.2</td>
<td>4.09</td>
<td>327.5</td>
</tr>
<tr>
<td>S. pombe</td>
<td>14</td>
<td>2.2</td>
<td>2.39</td>
<td>92.7</td>
</tr>
<tr>
<td>Aspergillus</td>
<td>13</td>
<td>3.1</td>
<td>5.27</td>
<td>72.2</td>
</tr>
</tbody>
</table>

## What About Introns?

<table>
<thead>
<tr>
<th>Species</th>
<th>Genomic size (Mbp)</th>
<th>Number of introns per gene</th>
<th>Size of individual introns</th>
<th>Total intron size per kb cds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All genes</td>
<td>Homologous genes</td>
<td>All genes</td>
</tr>
<tr>
<td>Human</td>
<td>3400</td>
<td>4.0</td>
<td>5.54</td>
<td>3413.4</td>
</tr>
<tr>
<td>Mouse</td>
<td>3454</td>
<td>3.1</td>
<td>6.68</td>
<td>1321.4</td>
</tr>
<tr>
<td>Rat</td>
<td>2900</td>
<td>3.0</td>
<td>3.77</td>
<td>1091.7</td>
</tr>
<tr>
<td>Chicken</td>
<td>1200</td>
<td>3.2</td>
<td>3.69</td>
<td>706.3</td>
</tr>
<tr>
<td>Drosophila</td>
<td>180</td>
<td>2.5</td>
<td>2.44</td>
<td>563.9</td>
</tr>
<tr>
<td>C. elegans</td>
<td>100</td>
<td>4.2</td>
<td>4.26</td>
<td>466.6</td>
</tr>
<tr>
<td>Cress</td>
<td>100</td>
<td>4.8</td>
<td>4.10</td>
<td>239.7</td>
</tr>
<tr>
<td>Corn</td>
<td>5000</td>
<td>4.2</td>
<td>4.09</td>
<td>327.5</td>
</tr>
<tr>
<td>S. pombe</td>
<td>14</td>
<td>2.2</td>
<td>2.39</td>
<td>92.7</td>
</tr>
<tr>
<td>Aspergillus</td>
<td>13</td>
<td>3.1</td>
<td>5.27</td>
<td>72.2</td>
</tr>
</tbody>
</table>

Human introns are much longer than other organism’s introns.

Why such diversity in intron number/size?
And Remember, We’re Sequencing New Genomes Rapidly
When We Assemble A New Genome, Finding the Genes is a Key Problem

Assembling a genome is just the beginning…
When We Assemble A New Genome, Finding the Genes is a Key Problem

...because if we want to do much with it we have to **annotate** it, identifying the location of functional elements including genes
When We Assemble A New Genome, Finding the Genes is a Key Problem

This is part of the original human genome paper in 2001. You can see six chromosomes. Each little annotation at the bottom is a gene.
When We Assemble A New Genome, Finding the Genes is a Key Problem

So, genome annotation generally and gene finding specifically is a big problem!!!
Outline

• Review of gene structure

• Scale of the problem

• Solutions
  – Empirical methods
  – Ab initio prediction
The Goal of Gene Finding

To create a **model** for every gene in a genome

A gene model describes the “structure” of a gene, comprising the locations in DNA sequence of all key gene features (introns and exons, transcription start, etc.).
How Would You Find Genes?
Ab Initio Gene Prediction

Here, we define sequence features of real genes based on experimental evidence

- Open reading frame model
- Splice donor sequence model
- Splice acceptor sequence model
- Intron/exon length distribution
- Requirement that introns maintain the reading frame

Then, we use these sequence features to obtain the best interpretation of where genes are in any region from sequence alone

Ab initio = from first principles
Example #2: Splice Donor and Acceptors

Splice donor and acceptor sites have characteristic sequences
Where Would You Infer Introns?

Size of arrows indicates strength of match to donor/acceptor motif sequence.
Where would you infer introns?

Size of arrows indicates strength of match to donor/acceptor motif sequence.

splice donor candidates
splice acceptor candidates

open reading frames (above threshold length, plus strand)

(example cont.)

stop codon before highest scoring splice donor!
Where would you infer introns?

Size of arrows indicates strength of match to donor/acceptor motif sequence.

---

**Example cont.**

Stop codon before highest scoring splice donor!

---

Open reading frames (above threshold length, plus strand)

Reinterpreted (avoids stop codon by using lower scoring splice donor):

- splice donor candidates
- splice acceptor candidates

---

Open reading frames (above threshold length, plus strand)
How Do We Actually Accomplish This Task?

It turns out that we can make a Hidden Markov Model that, given a particular sequence, can return the most likely gene model.