Gene Finding

BCH339N Systems Biology / Bioinformatics – Spring 2016
Edward Marcotte, Univ of Texas at Austin

Lots of genes in every genome

Do humans really have the biggest genomes?

Lots of genes in every genome

6.5 Gb (2X human)
17K genes

gigabases (not gigabytes)

The Telegraph

World’s largest genome belongs to slow-growing mountain flower

An unremarkable and slow-growing plant has stunned scientists after they found it had the world’s largest genome – 50 times bigger than that of our own species.
Where are the genes? How can we find them?

A toy HMM for 5′ splice site recognition (from Sean Eddy's NBT primer linked on the course web page)

Remember this?
Let’s start with prokaryotic genes

What elements should we build into an HMM to find bacterial genes?

Can be polycistronic:
A CpG island model might look like:

\[
P(\text{X} \mid \text{CpG island}) \quad \frac{P(\text{X} \mid \text{CpG island})}{P(\text{X} \mid \text{not CpG island})}
\]

\(p(C \rightarrow G)\) is higher

could calculate (or log ratio) along a sliding window, just like the fair/biased coin test

One way to build a minimal gene finding Markov model

\[
P(\text{X} \mid \text{coding}) \quad \frac{P(\text{X} \mid \text{coding})}{P(\text{X} \mid \text{not coding})}
\]

Transition probabilities reflect codons

Transition probabilities reflect intergenic DNA

Could calculate (or log ratio) along a sliding window, just like the fair/biased coin test
Really, we’ll want to detect codons. The usual trick is to use a higher-order Markov process.

A standard Markov process only considers the current position in calculating transition probabilities.

An \( n \text{th-order} \) Markov process takes into account the past \( n \) nucleotides, e.g. as for a 5\( \text{th} \) order:

\[
\text{Codon 1} \quad \begin{array}{cccccc}
    i-5 & i-4 & i-3 & i-2 & i-1 & i \\
\end{array} \quad \text{Codon 2}
\]

5\( \text{th} \) order Markov chain, using models of coding vs. non-coding using the classic algorithm GenMark.
An HMM version of GenMark

For example, accounting for variation in start codons...

The probabilities of the start codons were defined in agreement with the *E.coli* genome statistics: $P(\text{ATG}) = 0.905$, $P(\text{GTG}) = 0.090$, $P(\text{TTG}) = 0.005$. The probability of transition from a non-coding state to a Typical (Atypical) coding state was set to 0.85 (0.15).
... and variation in gene lengths

Length distributions (in # of nucleotides)

Coding (ORFs)  Non-coding (intergenic)

(Placing these curves on top of each other)

Short ORFS occur often by chance

Long ORFS tend to be real protein coding genes

Protein-coding genes <100 aa’s are hard to find

GeneMark.hmm: new solutions for gene finding

Alexander V. Lukashov and Mark Baranowski

Model for a ribosome binding site
(based on ~300 known RBS’s)

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Position 1</th>
<th>Position 2</th>
<th>Position 3</th>
<th>Position 4</th>
<th>Position 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.161</td>
<td>0.050</td>
<td>0.012</td>
<td>0.071</td>
<td>0.115</td>
</tr>
<tr>
<td>C</td>
<td>0.077</td>
<td>0.037</td>
<td>0.012</td>
<td>0.025</td>
<td>0.046</td>
</tr>
<tr>
<td>A</td>
<td><strong>0.681</strong></td>
<td>0.105</td>
<td>0.015</td>
<td><strong>0.861</strong></td>
<td>0.164</td>
</tr>
<tr>
<td>G</td>
<td>0.077</td>
<td><strong>0.808</strong></td>
<td><strong>0.960</strong></td>
<td>0.043</td>
<td><strong>0.659</strong></td>
</tr>
</tbody>
</table>

How well does it do on well-characterized genomes?

<table>
<thead>
<tr>
<th>Genome</th>
<th>Genes annotated</th>
<th>Genes predicted</th>
<th>Exact prediction (%)</th>
<th>Missing genes (%)</th>
<th>Wrong genes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. fumigatus</td>
<td>2407</td>
<td>2530</td>
<td>73.1</td>
<td>10.8 (2.0)</td>
<td>15.1</td>
</tr>
<tr>
<td>B. subtilis</td>
<td>4101</td>
<td>4384</td>
<td>77.5</td>
<td>3.6 (2.8)</td>
<td>9.8</td>
</tr>
<tr>
<td>E. coli</td>
<td>4288</td>
<td>4440</td>
<td>75.4</td>
<td>5.0 (2.7)</td>
<td>8.2</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>1718</td>
<td>1840</td>
<td>86.7</td>
<td>3.8 (3.2)</td>
<td>10.2</td>
</tr>
<tr>
<td>H. pylori</td>
<td>1566</td>
<td>1612</td>
<td>79.7</td>
<td>6.0 (4.4)</td>
<td>8.7</td>
</tr>
<tr>
<td>M. genitalium</td>
<td>467</td>
<td>509</td>
<td>78.4</td>
<td>9.9 (1.7)</td>
<td>17.3</td>
</tr>
<tr>
<td>M. jannaschii</td>
<td>1680</td>
<td>1841</td>
<td>72.7</td>
<td>4.6 (0.8)</td>
<td>12.9</td>
</tr>
<tr>
<td>M. pneumonicum</td>
<td>678</td>
<td>734</td>
<td>70.1</td>
<td>7.8 (4.1)</td>
<td>13.6</td>
</tr>
<tr>
<td>M. thermoaerophilicum</td>
<td>1869</td>
<td>1944</td>
<td>70.9</td>
<td>5.0 (3.5)</td>
<td>8.6</td>
</tr>
<tr>
<td>S. mutans</td>
<td>3169</td>
<td>3360</td>
<td>89.6</td>
<td>4.0 (1.5)</td>
<td>9.4</td>
</tr>
<tr>
<td>Averaged</td>
<td>21 943</td>
<td>23 194</td>
<td>78.1</td>
<td>5.4 (2.7)</td>
<td>10.4</td>
</tr>
</tbody>
</table>

But this was a long time ago!
What elements should we build into an HMM to find eukaryotic genes?
We’ll look at the GenScan eukaryotic gene annotation model:

Zoomed in on the forward strand model...
Introns and different flavors of exons all have different typical lengths.

Taking into account donor splice sites.
An example of an annotated gene...

How well do these programs work?
We can measure how well an algorithm works using these:

<table>
<thead>
<tr>
<th>Algorithm predicts:</th>
<th>True answer:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive positive</td>
<td>True positive</td>
</tr>
<tr>
<td>Negative positive</td>
<td>False positive</td>
</tr>
<tr>
<td>False negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>

Specificity = $\frac{TP}{TP + FP}$
Sensitivity = $\frac{TP}{TP + FN}$
How well do these programs work? How good are our current gene models?

Accuracy per base

<table>
<thead>
<tr>
<th>Program</th>
<th>Sequences</th>
<th>Sn</th>
<th>Sp</th>
<th>Accuracy per exon</th>
<th>Sn</th>
<th>Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENSCAN</td>
<td>570 (8)</td>
<td>0.93</td>
<td>0.93</td>
<td>0.78</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>FGENEH</td>
<td>569 (22)</td>
<td>0.77</td>
<td>0.88</td>
<td>0.51</td>
<td>0.61</td>
<td>0.64</td>
</tr>
<tr>
<td>GeneID</td>
<td>570 (2)</td>
<td>0.63</td>
<td>0.81</td>
<td>0.44</td>
<td>0.46</td>
<td>0.48</td>
</tr>
<tr>
<td>Genie</td>
<td>570 (0)</td>
<td>0.76</td>
<td>0.77</td>
<td>0.55</td>
<td>0.51</td>
<td>0.48</td>
</tr>
<tr>
<td>GenLang</td>
<td>570 (30)</td>
<td>0.72</td>
<td>0.79</td>
<td>0.35</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td>GeneParser2</td>
<td>562 (0)</td>
<td>0.66</td>
<td>0.79</td>
<td>0.36</td>
<td>0.43</td>
<td>0.47</td>
</tr>
<tr>
<td>GRAIL2</td>
<td>570 (23)</td>
<td>0.72</td>
<td>0.87</td>
<td>0.42</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>SORFIND</td>
<td>561 (0)</td>
<td>0.71</td>
<td>0.85</td>
<td>0.15</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>Xpound</td>
<td>570 (28)</td>
<td>0.61</td>
<td>0.87</td>
<td>0.73</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>GeneID+</td>
<td>478 (1)</td>
<td>0.91</td>
<td>0.91</td>
<td>0.56</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>GeneParser3</td>
<td>478 (1)</td>
<td>0.86</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In general, we can do better with more data, such as mRNA and conservation.

**Box 2: Gene prediction versus gene annotation**

In the year 2000, scientists from around the world held a contest ("GASP") to predict genes in part of the fly genome, then compare them to experimentally determined "truth".
How well do we know the genes now?

“In the year 2000...”

“Over 95% of the coding nucleotides ... were correctly identified by the majority of the gene finders.”

“...the correct intron/exon structures were predicted for >40% of the genes.”

Most promoters were missed; many were wrong.

“Integrating gene finding and cDNA/EST alignments with promoter predictions decreases the number of false-positive classifications but discovers less than one-third of the promoters in the region.”

EGASP: the Project

Assessment

= scientists predict genes

We discussed these earlier

18 groups
36 programs
So how did they do?

- “The best methods had at least one gene transcript correctly predicted for close to **70%** of the annotated genes.”

- “…taking into account alternative splicing, … only approximately **40%** to **50%** accuracy.

- At the coding nucleotide level, the best programs reached an accuracy of **90%** in both sensitivity and specificity.”

At the gene level, most genes have errors
How well do we know the genes now?

nGASP – the nematode genome annotation assessment project

- scientists from around the world held a contest (“NGASP”) to predict genes in part of the worm genome, then compare them to experimentally determined “truth”

- 17 groups from around the world competed

- “Median gene level sensitivity ... was 78%”

- “their specificity was 42%, comparable to human

For example:
How well do we know the genes now?

**GENCODE: The reference human genome annotation for The ENCODE Project**

- a large consortium of scientists trying to annotate the human genome using a combination of experiment and prediction.

Best estimate of the current state of human genes.

Quality of evidence used to support automatic, manually, and merged annotated transcripts (probably reflective of transcript quality)

- 23,855 transcripts
- 89,669 transcripts
- 22,535 transcripts
How well do we know the genes now?

The bottom line:
• Gene prediction and annotation are hard
• Annotations for all organisms are still buggy
• Few genes are 100% correct; expect multiple errors per gene
• Most organisms’ gene annotations are probably much worse than for humans