Gene Finding

BCH394P/374C Systems Biology / Bioinformatics
Edward Marcotte, Univ of Texas at Austin

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.
Genome size ranges vary widely across organisms


A pine tree

Us

Height (not area) is proportional to genome size

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:

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

\[ p(C \rightarrow G) \] is lower

\[ P(X | \text{CpG island}) \]

\[ P(X | \text{not CpG island}) \]

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

\[ P(X | \text{coding}) \]

\[ P(X | \text{not coding}) \]

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 $n^{th}$-order Markov process takes into account the past $n$ nucleotides, e.g. as for a 5th order:

![Diagram of a 5th order Markov chain](image)

5th order Markov chain, using models of coding vs. non-coding using the classic algorithm GenMark

![Graphs showing data distribution for different reading frames and strands](image)
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
Model for a ribosome binding site
(based on ~300 known RBS’s)

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td></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></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></td>
<td>0.681</td>
<td>0.105</td>
<td>0.015</td>
<td>0.861</td>
<td>0.164</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td>0.077</td>
<td>0.808</td>
<td>0.960</td>
<td>0.043</td>
<td>0.659</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. fucisideris</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. pneumoniae</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>Synechocystis</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!
Eukaryotic genes

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...

21

22
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></td>
<td>Positive</td>
</tr>
<tr>
<td>Negative positive</td>
<td>True positive</td>
</tr>
<tr>
<td>False negative</td>
<td>False positive</td>
</tr>
</tbody>
</table>

Specificity = TP / (TP + FP)
Sensitivity = TP / (TP + FN)
How well do these programs work?
How good are our current gene models?

![Reference gene model diagram]

<table>
<thead>
<tr>
<th>Program</th>
<th>Sequences</th>
<th>Accuracy per base</th>
<th>Accuracy per exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENSCAN</td>
<td>570 (8)</td>
<td>0.93 0.93</td>
<td>0.78 0.81</td>
</tr>
<tr>
<td>FGENEH</td>
<td>569 (22)</td>
<td>0.77 0.88</td>
<td>0.61 0.64</td>
</tr>
<tr>
<td>GeneID</td>
<td>570 (2)</td>
<td>0.63 0.81</td>
<td>0.44 0.46</td>
</tr>
<tr>
<td>Genie</td>
<td>570 (0)</td>
<td>0.76 0.77</td>
<td>0.55 0.48</td>
</tr>
<tr>
<td>GenLang</td>
<td>570 (30)</td>
<td>0.72 0.79</td>
<td>0.51 0.52</td>
</tr>
<tr>
<td>GeneParser2</td>
<td>562 (0)</td>
<td>0.66 0.79</td>
<td>0.35 0.40</td>
</tr>
<tr>
<td>GRAIL2</td>
<td>570 (23)</td>
<td>0.72 0.87</td>
<td>0.36 0.43</td>
</tr>
<tr>
<td>SORFIND</td>
<td>561 (0)</td>
<td>0.71 0.85</td>
<td>0.42 0.47</td>
</tr>
<tr>
<td>Xpound</td>
<td>570 (28)</td>
<td>0.61 0.87</td>
<td>0.15 0.18</td>
</tr>
<tr>
<td>GeneID+</td>
<td>478 (1)</td>
<td>0.91 0.91</td>
<td>0.73 0.70</td>
</tr>
<tr>
<td>GeneParser3</td>
<td>478 (1)</td>
<td>0.86 0.91</td>
<td>0.56 0.58</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**

<table>
<thead>
<tr>
<th>Gene prediction (SNAP)</th>
<th>mRNA or EST evidence (Exonerate)</th>
<th>Protein evidence (BLASTX)</th>
<th>Gene annotation resulting from synthesizing all available evidence (two alternative splice forms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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".

** Genome Annotation Assessment in Drosophila melanogaster**

<table>
<thead>
<tr>
<th>Program name</th>
<th>Gene ID</th>
<th>Fusion recognition</th>
<th>EST-c DNA alignment</th>
<th>Protein similarity</th>
<th>Repeat</th>
<th>Gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENIA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>GeneID</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EnsemblGene</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>SUCCINO</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flagon</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAFPIE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>TAVF</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenomeTgenera</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCBItranscript</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genefac</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How well do we know the genes now?

“In the year 2000, how well do we know the genes now?“

“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

= scientists from around the world held a contest (“EGASP”) to predict genes in part of the human genome, then compare them to experimentally determined “truth.”

18 groups
36 programs

We discussed these earlier.

In the year 2006
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?

**In the year 2008**

**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:**

[Image of a diagram showing gene annotations and predictions]
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.*

How well do we know the genes now?

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

- **A**: 23,855 transcripts
- **B**: 89,669 transcripts
- **C**: 22,535 transcripts

*Genome Res. 2012 22: 1760-1774*
How well do we know the genes now?

**In the year 2015**

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

What about the current state of prokaryote gene models?
Here’s the overlap in gene predictions from 4 algs on 20 test strains:

AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions
Deepank R. Korandla, Jacob M. Wozniak, Anaamika Campeau, David J. Gonzalez and Erik S. Wright

*Bioinformatics, 36(4), 2020, 1022–1029*
What about the current state of prokaryote gene models?

• “We applied AssessORF to compare gene predictions offered by GenBank, GeneMarkS-2, Glimmer and Prodigal on genomes spanning the prokaryotic tree of life.

• Gene predictions were 88–95% in agreement with the available evidence, with Glimmer performing the worst but no clear winner.

• All programs were biased towards selecting start codons that were upstream of the actual start.”

Bioinformatics, 36(4), 2020, 1022–1029

In practice, gene finding and genome annotation combines all lines of evidence, e.g. as for the frog genome:

Align frog RNA sequencing data (ESTs and cDNA) & BLAST genes from other animals vs. frog assembly → Define gene segments

Integrate ab initio gene predictions & BLAST hits using Fgenesh and GenomeScan (= GenScan successor, Genome Research 11:803 (2001))

Refine with RNA-seq and H3K4me3 data

Refine vs final genome assembly

Manually curate 412 gene models → Estimate 96% accuracy overall

Session et al., Nature 2016
Supplementary Info, pg. 22