Genes & Gene Finding

Ben Langmead

Please sign guestbook ([www.langmead-lab.org/teaching-materials](http://www.langmead-lab.org/teaching-materials)) to tell me briefly how you are using the slides. For original Keynote files, email me ([ben.langmead@gmail.com](mailto:ben.langmead@gmail.com)).
Gene finding

Recall the "Central Dogma" and the centrality of genes

DNA molecules contain information about how to create proteins; this is transcribed into RNA molecules, which, in turn, direct chemical machinery to translate the message into a protein.


Bacterial genes

Genes (colored arrows) packed tightly into the *Staphylococcus aureus* genome

In bacteria, one gene corresponds to one continuous interval on the genome

We have good methods for predicting where these genes are
## Bacterial gene finding

Approaches can identify exact bacterial genes with as high as 92% accuracy; can identity gene ends with ≥98% accuracy


### Table 3. Glimmer3 prediction accuracy with long-orf s training

<table>
<thead>
<tr>
<th>Organism</th>
<th>GC%</th>
<th># Genes</th>
<th>Glimmer3 Predictions</th>
<th>versus Glimmer2.13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G3%</td>
<td></td>
<td>3' Matches</td>
<td>5' &amp; 3' Matches</td>
</tr>
<tr>
<td><em>A. fulgidus</em></td>
<td>49</td>
<td>1165</td>
<td>1161</td>
<td>99.7%</td>
</tr>
<tr>
<td><em>B. anthracis</em></td>
<td>35</td>
<td>3132</td>
<td>3125</td>
<td>99.8%</td>
</tr>
<tr>
<td><em>B. subtilis</em></td>
<td>44</td>
<td>1576</td>
<td>1562</td>
<td>99.1%</td>
</tr>
<tr>
<td><em>C. tepidum</em></td>
<td>57</td>
<td>1292</td>
<td>1289</td>
<td>99.8%</td>
</tr>
<tr>
<td><em>C. perfringens</em></td>
<td>29</td>
<td>1504</td>
<td>1501</td>
<td>99.8%</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>51</td>
<td>3603</td>
<td>3534</td>
<td>98.1%</td>
</tr>
<tr>
<td><em>G. sulfurreducens</em></td>
<td>61</td>
<td>2351</td>
<td>2337</td>
<td>99.4%</td>
</tr>
<tr>
<td><em>H. pylori</em></td>
<td>39</td>
<td>915</td>
<td>910</td>
<td>99.5%</td>
</tr>
<tr>
<td><em>P. fluorescens</em></td>
<td>63</td>
<td>4535</td>
<td>4510</td>
<td>99.4%</td>
</tr>
<tr>
<td><em>R. solanacearum</em></td>
<td>67</td>
<td>2512</td>
<td>2485</td>
<td>98.9%</td>
</tr>
<tr>
<td><em>S. epidermidis</em></td>
<td>32</td>
<td>1650</td>
<td>1646</td>
<td>99.8%</td>
</tr>
<tr>
<td><em>T. pallidum</em></td>
<td>53</td>
<td>575</td>
<td>567</td>
<td>98.6%</td>
</tr>
<tr>
<td><em>U. parvum</em></td>
<td>26</td>
<td>327</td>
<td>324</td>
<td>99.1%</td>
</tr>
<tr>
<td>Averages:</td>
<td></td>
<td></td>
<td></td>
<td>99.3%</td>
</tr>
</tbody>
</table>

Genomes and columns are as in the preceding table. Glimmer3 was run by using the output of its long-orf s program to train an IMM. The output of an initial run of Glimmer3 was used to set start codon frequencies and to find a ribosome-binding-site motif. A second run of Glimmer3 using those values generated the above predictions. Glimmer2 was trained on the output of its version of the long-orf s program.
Eukaryotic genes

Eukaryotic genes are more complex than prokaryotic (bacterial) genes for several reasons, as we’ll see.

Likewise, finding eukaryotic genes computationally is harder.
Gene finding

During the Human Genome Project, there was public debate about how many genes were in the human genome.

A range of predictions were made: ~40K to ~100K.

Answer turned out to be ~20K, and the number of protein-coding genes has slowly but steadily decreased since then.

But how did they find the genes given the genome sequence?
Homo sapiens hemoglobin, beta (HBB)
Genes

Sequence models will allow us to "see" gene sequences in the ocean of the genome

Recall a few things we've learned about genes and transcription

Transcription produces an RNA copy of a stretch of DNA but with Ts (thymine) replaced by Us (uracil)

Genes

Triples of nucleotides ("codons") are translated into amino acids via the genetic code

Genes

Some codons are special ("stop codons"), signaling that translation of the protein should stop.

"Start codon" (AUG) signals translation should begin; also codes for amino acid Methionine.
**Genes**

*Splicing* is a process by which some portions of the mRNA are cut out prior to translation

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCCTAAGCCAGTGCAGAAGAGGTAGAGGACAGGTACGGCTGTCATCACTTAGACCTCACTGCCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTAACTAA

GCTCGCTTTCTTGCTGTCCAAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTAAAAAGGGAATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATATCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTAACATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTC

TGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTTATCTTCCTCCCACAGTCTCCTGGGCAACGTGCTGGTCTGTG

TGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATACTAA
Splicing is a process by which some portions of the mRNA are cut out prior to translation.
Genes

Splicing happens at certain nucleotide patterns: **GU** and **AG**

Animation of splicing:  [https://evolutionnews.org/2013/09/the_spliceosome_1/](https://evolutionnews.org/2013/09/the_spliceosome_1/)
Transcription

Genome → ... → Pre-mRNA → Mature mRNA

Transcription

Splicing

Coding segment

Translated by ribosome

AUG → GU AG → GU AG → UAA

Start codon → codons → Donor → codons → Acceptor → codons → Stop codon

: exon  : DNA  : intron  : untranslated (UTR)
A human gene

chr11:5246500-5248500 (reverse strand):

ATATCTTAGGAGGCTGAGGTGTTYGGAAGTCTGGAACACTCTCTAAGCCAGTGGGACAGCTCTGCTGACATCTCTAGGCTTATCATTGCTTCTAGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGCTATCGGAGGAGGGTATCAGGGGTTGGCCATCTTGGTCACCCCTGGGGTGCTGGGGCATAAAGTCCAACTACTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTATTTATTTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTAAAAGGGAATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATACTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTAATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTCCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC

Start codon
Donor
Acceptor
Stop codon
These are the signals we want to capture with a sequence model.

Probabilistic model is appropriate since the signals are "fuzzy" to various degrees.

There are a lot of start codons, stop codons, donors and acceptors in the genome that have nothing to do with translation or splicing.

Only some donor or acceptors in a gene are involved in splicing.

Exception: stop codon in a gene always stops translation.
Can HMM help us find eukaryotic genes and their constituent parts?

What will be the states? Emissions?
Eukaryotic gene finding

Emissions are nucleotides

I = intron
E = exon
N = intergenic
(between genes)

q0 is a start state;
guarantees we start in
the N (intergenic) state

Model captures:
Exons and introns and
space between genes

Does not capture:
Acceptors & donors,
start & stop codons,
other codons

Eukaryotic gene finding

What if we wanted to model the three codon positions separately?

Eukaryotic gene finding

As before:

I = intron
E = exon
N = intergenic
    (between genes)

Can we additionally model
start & stop codons and
donors & acceptors?

Eukaryotic gene finding

Emission probability is 1.0 for bases shown (AUG), 0.0 for others

Eukaryotic gene finding

Eukaryotic gene finding

Which nodes have non-trivial emission probabilities that we must learn?

All non-motif nodes

Eukaryotic gene finding

Which edges have non-trivial transition probabilities that we must learn?

Edges outgoing from nodes with multiple outgoing edges.

Eukaryotic gene finding

Given a trained model and emission string, is it possible for a backtrace (not necessarily the Viterbi backtrace) to have probability 0?

Yes: e.g. any backtrace that puts us in state 2 at a step where the emission string does not have “A”