Some References

- A great online bib
  - http://www.nslij-genetics.org/gene/
- A good intro survey
- A gene finding bake-off

Motivation

- Sequence data flooding into Genbank
- What does it mean?
  - protein genes, RNA genes, mitochondria, chloroplast, regulation, replication, structure, repeats, transposons, unknown stuff, ...

Protein Coding Nuclear DNA

- Focus of next 2 lectures
- Goal: Automated annotation of new sequence data
- State of the Art:
  - predictions ~ 60% similar to real proteins
  - ~80% if database similarity used
  - lab verification still needed, still expensive
Biological Basics

- Central Dogma:
  DNA \( \rightarrow \) RNA \( \rightarrow \) Protein

- Codons: 3 bases code one amino acid
  - Start codon
  - Stop codons
  - 3', 5' Untranslated Regions (UTR's)

The Genetic Code

(a) RNA Codons for the Twenty Amino Acids

<table>
<thead>
<tr>
<th>mRNA Codon</th>
<th>Amino-acid</th>
<th>Amino-acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Pro</td>
<td>Ser</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Tyr</td>
</tr>
<tr>
<td>A</td>
<td>Val</td>
<td>His</td>
</tr>
<tr>
<td>G</td>
<td>Arg</td>
<td>Cys</td>
</tr>
</tbody>
</table>

Ribosomes
Idea #1: Find Long ORF’s
- **Reading frame**: which of the 3 possible sequences of triples does the ribosome read?
- **Open** Reading Frame: No stop codons
- In random DNA
  - average ORF = 64/3 = 21 triplets
  - 300bp ORF once per 36kbp per strand
- But average protein ~ 1000bp

Idea #2: Codon Frequency
- In random DNA
  - Leucine : Alanine : Tryptophan = 6 : 4 : 1
- But in real protein, ratios ~ 6.9 : 6.5 : 1
- So, coding DNA is not random
- Even more: synonym usage is biased
  (in a species dependant way)
  - examples known with 90% AT 3rd base

Recognizing Codon Bias
- Assume
  - Codon usage i.i.d.; abc with freq. f(abc)
  - $a_1a_2a_3a_4...a_{3n+2}$ is coding, unknown frame
- Calculate
  - $p_1 = f(a_1a_2a_3)f(a_4a_5a_6)...f(a_{3n-2}a_{3n-1}a_{3n})$  
  - $p_2 = f(a_2a_3a_4)f(a_5a_6a_7)...f(a_{3n-1}a_{3n},a_{3n+1})$  
  - $p_3 = f(a_3a_4a_5)f(a_6a_7a_8)...(a_{3n},a_{3n+1},a_{3n+2})$  
- $P_i = \frac{p_i}{(p_1+p_2+p_3)}$
- More generally: k-th order Markov model
  - k=5 or 6 is typical

Codon Usage in Φx174
Promoters, etc.

- In prokaryotes, most DNA coding
  - E.g. ~70% in *H. influenzae*
- Long ORFs + codon stats do well
- But obviously won’t be perfect
  - short genes
  - 5’ & 3’ UTR’s
- Can improve by modeling promoters & other
  - e.g. via WMM or higher-order Markov models

Eukaryotes

- As in prokaryotes (but maybe more variable)
  - promoters
  - start/stop transcription
  - start/stop translation
- New Features:
  - polyA site/tail
  - introns, exons, splicing
  - branch point signal
  - alternative splicing

Characteristics of human genes
(Nature, 2/2001, Table 21)

<table>
<thead>
<tr>
<th>Sample (size)</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal exon</td>
<td>122 bp</td>
<td>145 bp</td>
</tr>
<tr>
<td>Exon number</td>
<td>7</td>
<td>8.8</td>
</tr>
<tr>
<td>Introns</td>
<td>1,023 bp</td>
<td>3,365 bp</td>
</tr>
<tr>
<td>3’ UTR</td>
<td>400 bp</td>
<td>770 bp</td>
</tr>
<tr>
<td>5’ UTR</td>
<td>240 bp</td>
<td>300 bp</td>
</tr>
<tr>
<td>Coding seq</td>
<td>1,100 bp</td>
<td>1340bp</td>
</tr>
<tr>
<td>(CDS)</td>
<td>367 aa</td>
<td>447 aa</td>
</tr>
<tr>
<td>Genomic extent</td>
<td>14 kb</td>
<td>27 kb</td>
</tr>
</tbody>
</table>

* 1,804 selected RefSeq entries were those with full-length unambiguous alignment to finished sequence

Big Genes

- Many genes are over 100 kb long,
- Max known: dystrophin gene (DMD), 2.4 Mb.
- The variation in the size distribution of coding sequences and exons is less extreme, although there are remarkable outliers.
  - The titin gene has the longest currently known coding sequence at 80,780 bp; it also has the largest number of exons (178) and longest single exon (17,106 bp).

RNApol rate: 2.5 kb/min
Figure 36 GC content. a, Distribution of GC content in genes and in the genome. For 9,315 known genes mapped to the draft genome sequence, the local GC content was calculated in a window covering either the whole alignment or 20,000 bp centered around the midpoint of the alignment, whichever was larger. Ns in the sequence were not counted. GC content for the genome was calculated for adjacent nonoverlapping 20,000-bp windows across the sequence. Both the gene and genome distributions have been normalized to sum to one.

b, Gene density as a function of GC content, obtained by taking the ratio of the data in a. Values are less accurate at higher GC levels because the denominator is small. c, Dependence of mean exon and intron lengths on GC content. For exons and introns, the local GC content was derived from alignments to finished sequence only, and were calculated from windows covering the feature or 10,000 bp centered on the feature, whichever was larger.

A Case Study -- Genscan


Training Data

- 238 multi-exon genes
- 142 single-exon genes
- total of 1492 exons
- total of 1254 introns
- total of 2.5 Mb

- NO alternate splicing, none > 30kb, ...
Performance Comparison

<table>
<thead>
<tr>
<th>Program</th>
<th>Sn per exon</th>
<th>Sp per exon</th>
<th>Sn per nuc.</th>
<th>Sp per nuc.</th>
<th>Avg.</th>
<th>ME per exon</th>
<th>WE per exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENSCAN</td>
<td>0.93</td>
<td>0.93</td>
<td>0.78</td>
<td>0.81</td>
<td>0.8</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>FGENEH</td>
<td>0.77</td>
<td>0.88</td>
<td>0.61</td>
<td>0.64</td>
<td>0.64</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>GeneID</td>
<td>0.63</td>
<td>0.81</td>
<td>0.44</td>
<td>0.46</td>
<td>0.46</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>Genie</td>
<td>0.76</td>
<td>0.77</td>
<td>0.55</td>
<td>0.48</td>
<td>0.51</td>
<td>0.17</td>
<td>0.33</td>
</tr>
<tr>
<td>GenLang</td>
<td>0.72</td>
<td>0.79</td>
<td>0.51</td>
<td>0.52</td>
<td>0.52</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td>GeneParser2</td>
<td>0.66</td>
<td>0.79</td>
<td>0.35</td>
<td>0.4</td>
<td>0.37</td>
<td>0.34</td>
<td>0.17</td>
</tr>
<tr>
<td>GRAIL2</td>
<td>0.72</td>
<td>0.87</td>
<td>0.36</td>
<td>0.43</td>
<td>0.4</td>
<td>0.25</td>
<td>0.11</td>
</tr>
<tr>
<td>SORFIND</td>
<td>0.71</td>
<td>0.85</td>
<td>0.42</td>
<td>0.47</td>
<td>0.45</td>
<td>0.24</td>
<td>0.14</td>
</tr>
<tr>
<td>Xpound</td>
<td>0.61</td>
<td>0.87</td>
<td>0.15</td>
<td>0.18</td>
<td>0.17</td>
<td>0.33</td>
<td>0.13</td>
</tr>
<tr>
<td>GeneID‡</td>
<td>0.91</td>
<td>0.91</td>
<td>0.73</td>
<td>0.7</td>
<td>0.71</td>
<td>0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>GeneParser3</td>
<td>0.86</td>
<td>0.91</td>
<td>0.56</td>
<td>0.58</td>
<td>0.57</td>
<td>0.14</td>
<td>0.09</td>
</tr>
</tbody>
</table>

After Burge & Karlin, Table 1. Sensitivity, Sn = TP/AP; Specificity, Sp = TP/PP.

Generalized Hidden Markov Models

- $\pi$: Initial state distribution
- $a_{ij}$: Transition probabilities
- One submodel per state
- Outputs are strings generated by submodel
- Given length L
  - Pick start state $q_1$ (random)
  - While $\sum d_i < L$
    - Pick string $s_i$ of length $d_i = |s_i|$ ~ submodel for $q_i$
    - Pick next state $q_{i+1}$ ($a_{ij}$)
  - Output $s_1 s_2 ...$

GHMM Structure
### Submodels

- **5' UTR**
  - $L \sim \text{geometric}(769 \text{ bp})$, $s \sim \text{MM}(5)$

- **3' UTR**
  - $L \sim \text{geometric}(457 \text{ bp})$, $s \sim \text{MM}(5)$

- **Intergenic**
  - $L \sim \text{geometric}(\text{GC-dependent})$, $s \sim \text{MM}(5)$

- **Introns**
  - $L \sim \text{geometric}(\text{GC-dependent})$, $s \sim \text{MM}(5)$

#### Submodel: Exons

- Inhomogenous 3-periodic 5th order Markov models
- Separate models for low GC (<43%), high GC
- Track “phase” of exons, i.e. reading frame.

### Effect of G+C Content

<table>
<thead>
<tr>
<th>Group</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>C + G% range</td>
<td>&lt;43</td>
<td>43-51</td>
<td>51-57</td>
<td>&gt;57</td>
</tr>
<tr>
<td>Number of genes</td>
<td>65</td>
<td>115</td>
<td>99</td>
<td>101</td>
</tr>
<tr>
<td>Est. proportion single-exon genes</td>
<td>0.16</td>
<td>0.19</td>
<td>0.23</td>
<td>0.16</td>
</tr>
<tr>
<td>Codelen: single-exon genes (bp)</td>
<td>1130</td>
<td>1251</td>
<td>1304</td>
<td>1137</td>
</tr>
<tr>
<td>Codelen: multi-exon genes (bp)</td>
<td>902</td>
<td>908</td>
<td>1118</td>
<td>1165</td>
</tr>
<tr>
<td>Introns per multi-exon gene</td>
<td>5.1</td>
<td>4.9</td>
<td>5.5</td>
<td>5.6</td>
</tr>
<tr>
<td>Mean intron length (bp)</td>
<td>2069</td>
<td>1086</td>
<td>801</td>
<td>518</td>
</tr>
<tr>
<td>Est. mean transcript length (bp)</td>
<td>10866</td>
<td>6504</td>
<td>5781</td>
<td>4833</td>
</tr>
<tr>
<td>Isochore</td>
<td>L1+L2</td>
<td>H1+H2</td>
<td>H3</td>
<td>H3</td>
</tr>
<tr>
<td>DNA amount in genome (Mb)</td>
<td>2074</td>
<td>1054</td>
<td>102</td>
<td>68</td>
</tr>
<tr>
<td>Estimated gene number</td>
<td>22100</td>
<td>24700</td>
<td>9100</td>
<td>9100</td>
</tr>
<tr>
<td>Est. mean intergenic length</td>
<td>83000</td>
<td>36000</td>
<td>5400</td>
<td>2600</td>
</tr>
<tr>
<td>Initial probabilities:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intergenic (N)</td>
<td>0.892</td>
<td>0.867</td>
<td>0.54</td>
<td>0.418</td>
</tr>
<tr>
<td>Intron (I+, I-)</td>
<td>0.095</td>
<td>0.103</td>
<td>0.338</td>
<td>0.388</td>
</tr>
<tr>
<td>5' Untranslated region (F+, F-)</td>
<td>0.008</td>
<td>0.018</td>
<td>0.077</td>
<td>0.122</td>
</tr>
<tr>
<td>3' Untranslated region (T+, T-)</td>
<td>0.005</td>
<td>0.011</td>
<td>0.045</td>
<td>0.072</td>
</tr>
</tbody>
</table>
Signal Models I: WMM’s
- Polyadenylation
  - 6 bp, consensus AATAAA
- Translation Start
  - 12 bp, starting 6 bp before start codon
- Translation stop
  - A stop codon, then 3 bp WMM

Signal Models II: more WMM’s
- Promoter
  - 70% TATA
    - 15 bp TATA WMM
    - s ~ null, L ~ Unif(14-20)
    - 8 bp cap signal WMM
- 30% TATA-less
  - 40 bp null

Signal Models III: W/WAM’s
- Acceptor Splice Site (3’ end of intron)
  - [-20..+3] relative to splice site modeled by “1st order weight array model”
- Branch point & polypyrimidine tract
  - Hard. Even weak consensus like YYRAY found in [-40..-21] in only 30% of training
  - “Windowed WAM”: 2nd order WAM, but averaged over 5 preceding positions
    “captures weak but detectable tendency toward YYY triplets and certain branch point related triplets like TGA, TAA, ...”

What's in the Primary Sequence?
Signal Models IV: Maximum Dependence Decomposition

- Donor splice sites (5' end of intron) show dependencies between non-adjacent positions, e.g., poor match at one end compensated by strong match at other end, 6 bp away.
- Model is basically a decision tree.
- Uses $\chi^2$ test to quantitate dependence.

<table>
<thead>
<tr>
<th>i</th>
<th>Con</th>
<th>j:</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>+3</th>
<th>+4</th>
<th>+5</th>
<th>+6</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td>c/a</td>
<td>---</td>
<td>61.8*</td>
<td>14.9</td>
<td>5.8</td>
<td>20.2*</td>
<td>11.2</td>
<td>18.0*</td>
<td>131.8*</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>A</td>
<td>115.6*</td>
<td>---</td>
<td>40.5*</td>
<td>20.3*</td>
<td>57.5*</td>
<td>59.7*</td>
<td>42.9*</td>
<td>336.5*</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>G</td>
<td>15.4</td>
<td>82.8*</td>
<td>---</td>
<td>13.0</td>
<td>61.5*</td>
<td>41.4*</td>
<td>96.6*</td>
<td>310.8*</td>
<td></td>
</tr>
<tr>
<td>+3</td>
<td>a/g</td>
<td>8.6</td>
<td>17.5*</td>
<td>13.1</td>
<td>---</td>
<td>19.3*</td>
<td>1.8</td>
<td>0.1</td>
<td>60.5*</td>
<td></td>
</tr>
<tr>
<td>+4</td>
<td>A</td>
<td>21.8*</td>
<td>56.0*</td>
<td>62.1*</td>
<td>64.1*</td>
<td>---</td>
<td>56.8*</td>
<td>0.2</td>
<td>260.9*</td>
<td></td>
</tr>
<tr>
<td>+5</td>
<td>G</td>
<td>11.6</td>
<td>60.1*</td>
<td>41.9*</td>
<td>93.6*</td>
<td>146.6*</td>
<td>---</td>
<td>33.6*</td>
<td>387.3*</td>
<td></td>
</tr>
<tr>
<td>+6</td>
<td>t</td>
<td>22.2*</td>
<td>40.7*</td>
<td>103.8*</td>
<td>26.5*</td>
<td>17.8*</td>
<td>32.6*</td>
<td>---</td>
<td>243.6*</td>
<td></td>
</tr>
</tbody>
</table>

* means chi-squared p-value < .001

$\chi^2$ test for independence:

$\chi^2 = \sum \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$

"expected" means expected assuming independence.

GHMM Structure
Summary of Burge & Karlin

- Coding DNA & control signals nonrandom
  - Weight matrices, WAMs, etc. for controls
  - Codon frequency, etc. for coding
- GHMM nice for overall architecture
- Careful attention to small details pays

Problems with BK training set

- 1 gene per sequence
- Annotation errors
- Single exon genes over-represented?
- Highly expressed genes over-represented?
- Moderate sized genes over-represented?
  (none > 30 kb) …
- Similar problems with other training sets, too

Problems with all methods

- Pseudo genes
- Short ORFs
- Sequencing errors
- Non-coding RNA genes & spliced UTR’s
- Overlapping genes
- Alternative splicing/polyadenylation
- Hard to find novel stuff -- not in training
- Species-specific weirdness -- spliced leaders, polycistronic transcripts, RNA editing…

Other ideas

- Database search - does gene you’re predicting look anything like a known protein?
- Comparative genomics - what does this region look like in related organisms?