A Note on HW #3

\[ \mu = 14.8 \]
\[ \sigma = 1 \]

\[ \mu = 23.1 \]
\[ \sigma = 1 \]

\[ \mu = 49.4 \]
\[ \sigma = 1 \]

3% change in LL may look small, but \( \exp(4.3) = 73.7 \) time more likely.
Gene Finding: Motivation

Sequence data flooding in
What does it mean?

protein genes, RNA genes, mitochondria, chloroplast, regulation, replication, structure, repeats, transposons, unknown stuff, …

More generally, how do you: learn from complex data in an unknown language, leverage what’s known to help discover what’s not
Protein Coding Nuclear DNA

Focus of this lecture
Goal: Automated annotation of new seq data
State of the Art:
  In Eukaryotes:
    predictions ~ 60% similar to real proteins
    ~80% if database similarity used
  Prokaryotes
    better, but still imperfect
Lab verification still needed, still expensive
Largely done for Human; unlikely for most others
Biological Basics

Central Dogma:

DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{translation}}$ Protein

Codons: 3 bases code one amino acid

- Start codon
- Stop codons
- 3’, 5’ Untranslated Regions (UTR’s)
RNA Transcription

(This gene is heavily transcribed, but many are not.)
# Codons & The Genetic Code

<table>
<thead>
<tr>
<th>First Base</th>
<th>Second Base</th>
<th>Third Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Phe</td>
<td>Ala : Alanine</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>Arg : Arginine</td>
</tr>
<tr>
<td></td>
<td>Tyr</td>
<td>Asn : Asparagine</td>
</tr>
<tr>
<td></td>
<td>Cys</td>
<td>Asp : Aspartic acid</td>
</tr>
<tr>
<td>Leu</td>
<td>Stop</td>
<td>Cys : Cysteine</td>
</tr>
<tr>
<td>Leu</td>
<td>Stop</td>
<td>Gln : Glutamine</td>
</tr>
<tr>
<td></td>
<td>Trp</td>
<td>Glu : Glutamic acid</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Gly : Glycine</td>
</tr>
<tr>
<td></td>
<td>Pro</td>
<td>His : Histidine</td>
</tr>
<tr>
<td></td>
<td>His</td>
<td>Ile : Isoleucine</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Leu : Leucine</td>
</tr>
<tr>
<td></td>
<td>Gln</td>
<td>Lys : Lysine</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Met : Methionine</td>
</tr>
<tr>
<td></td>
<td>Gln</td>
<td>Phe : Phenylalanine</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>Pro : Proline</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Ser : Serine</td>
</tr>
<tr>
<td></td>
<td>Thr</td>
<td>Thr : Threonine</td>
</tr>
<tr>
<td></td>
<td>Asn</td>
<td>Trp : Tryptophane</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>Tyr : Tyrosine</td>
</tr>
<tr>
<td></td>
<td>Met/Start</td>
<td>Val : Valine</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala : Alanine</td>
</tr>
<tr>
<td></td>
<td>Ala</td>
<td>Arg : Arginine</td>
</tr>
<tr>
<td></td>
<td>Asp</td>
<td>Asn : Asparagine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Asp : Aspartic acid</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Cys : Cysteine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Gln : Glutamine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Glu : Glutamic acid</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Gly : Glycine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>His : Histidine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Ile : Isoleucine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Leu : Leucine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Lys : Lysine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Met : Methionine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Phe : Phenylalanine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Pro : Proline</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Ser : Serine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Thr : Threonine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Trp : Tryptophane</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Tyr : Tyrosine</td>
</tr>
<tr>
<td></td>
<td>Gly</td>
<td>Val : Valine</td>
</tr>
</tbody>
</table>
Translation: mRNA → Protein
Ribosomes

Watson, Gilman, Witkowski, & Zoller, 1992
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 $\sim \frac{64}{3} = 21$ triplets
- 300bp ORF once per 36kbp per strand

But average protein $\sim 1000$bp
A Simple ORF finder

start at left end
scan triplet-by-non-overlapping triplet for AUG
then continue scan for STOP
repeat until right end
repeat all starting at offset 1
repeat all starting at offset 2
then do it again on the other strand
Scanning for ORFs

* In bacteria, GUG is sometimes a start codon…
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
    Why? E.g. efficiency, histone, enhancer, splice interactions
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

\[p_1 = f(a_1a_2a_3)f(a_4a_5a_6)...f(a_{3n-2}a_{3n-1}a_{3n})\]
\[p_2 = f(a_2a_3a_4)f(a_5a_6a_7)...f(a_{3n-1}a_{3n}a_{3n+1})\]
\[p_3 = f(a_3a_4a_5)f(a_6a_7a_8)...f(a_{3n}a_{3n+1}a_{3n+2})\]
\[P_i = p_i / (p_1+p_2+p_3)\]

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, etc.
   e.g. via WMM or higher-order Markov models
Eukaryotes

As in prokaryotes (but maybe more variable)
promoters
start/stop transcription
start/stop translation
And then…

Nobel Prize of the week: P. Sharp, 1993, Splicing
Mechanical Devices of the Spliceosome: Motors, Clocks, Springs, and Things

Jonathan P. Staley and Christine Guthrie

CELL Volume 92, Issue 3, 6 February 1998, Pages 315-326
Figure 2. Spliceosome Assembly, Rearrangement, and Disassembly Requires ATP, Numerous DExD/H box Proteins, and Prp24. The snRNPs are depicted as circles. The pathway for *S. cerevisiae* is shown.
Figure 3. Splicing Requires Numerous Rearrangements

E.g.: exchange of U1 for U6
Figure 6. A Paradigm for Unwindase Specificity and Timing? The DExD/H box protein UAP56 (orange) binds U2AF65 (pink) through its linker region (L). U2 binds the branch point. Y's indicate the polypyrimidine stretch; RS, RRM as in Figure 5A. Sequences are from mammals.
Hints to Origins?

Tetrahymena thermophila
Genes in Eukaryotes

As in prokaryotes \( \text{(but maybe more variable)} \)
- 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></th>
<th>Median</th>
<th>Mean</th>
<th>Sample (size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal exon</td>
<td>122 bp</td>
<td>145 bp</td>
<td>RefSeq alignments to draft genome sequence, with confirmed intron boundaries (43,317 exons)</td>
</tr>
<tr>
<td>Exon number</td>
<td>7</td>
<td>8.8</td>
<td>RefSeq alignments to finished seq (3,501 genes)</td>
</tr>
<tr>
<td>Introns</td>
<td>1,023 bp</td>
<td>3,365 bp</td>
<td>RefSeq alignments to finished seq (27,238 introns)</td>
</tr>
<tr>
<td>3' UTR</td>
<td>400 bp</td>
<td>770 bp</td>
<td>Confirmed by mRNA or EST on chromo 22 (689)</td>
</tr>
<tr>
<td>5' UTR</td>
<td>240 bp</td>
<td>300 bp</td>
<td>Confirmed by mRNA or EST on chromo 22 (463)</td>
</tr>
<tr>
<td>Coding seq</td>
<td>1,100 bp</td>
<td>1340 bp</td>
<td>Selected RefSeq entries (1,804)*</td>
</tr>
<tr>
<td>(CDS)</td>
<td>367 aa</td>
<td>447 aa</td>
<td></td>
</tr>
<tr>
<td>Genomic span</td>
<td>14 kb</td>
<td>27 kb</td>
<td>Selected RefSeq entries (1,804)*</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: 1.2-2.5 kb/min = >16 hours to transcribe DMD
**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 on 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 distributions normalized to sum to one.

**b:** Gene density as a function of GC content (= ratios of data in a. Less accurate at high GC because the denominator is small)

**c:** Dependence of mean exon and intron lengths on GC content. The local GC content, based on alignments to finished sequence only, calculated from windows covering the larger of feature size or 10,000 bp centered on it
Computational Gene Finding?

How do we algorithmically account for all this complexity…
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>per nuc. Sn</th>
<th>per nuc. Sp</th>
<th>per exon Sn</th>
<th>per exon Sp</th>
<th>Avg.</th>
<th>ME</th>
<th>WE</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>0.80</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>0.45</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>0.40</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>0.40</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>0.70</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* gen’ed by submodel

Given length \( L \)

1. Pick start state \( q_1 \) (\( \sim \pi \))
2. While \( \sum d_i < L \)
   - Pick \( d_i \) & string \( s_i \) of length \( d_i \) \( \sim \) submodel for \( q_i \)
   - Pick next state \( q_{i+1} \) (\( \sim a_{ij} \))
3. Output \( s_1 s_2 \ldots \)
Decoding

A “parse” $\phi$ of $s = s_1 s_2 \ldots s_L$ is a pair $d = d_1 d_2 \ldots d_k$, $q = q_1 q_2 \ldots q_k$ with $\sum d_i = L$

A forward/backward-like alg calculates, e.g.:

$$Pr(\text{generate } s_1 s_2 \ldots s_i \& \text{end in state } q_k)$$

(summing over possible predecessor states $q_{k-1}$ and possible $d_k$, etc.)

$$Pr(\phi(s)) = \frac{Pr(\phi \ast s)}{Pr(s)} \ldots$$
GHMM Structure
Figure 4. Length distributions are shown for (a) 1254 introns; (b) 238 initial exons; (c) 1151 internal exons; and (d) 238 terminal exons from the 238 multi-exon genes of the learning set $\mathcal{G}$. Histograms (continuous lines) were derived with a bin size of 300 bp in (a), and 25 bp in (b), (c), (d). The broken line in (a) shows a geometric (exponential) distribution with parameters derived from the mean of the intron lengths; broken lines in (b), (c) and (d) are the smoothed empirical distributions of exon lengths used by GENSCAN (details given by Burge, 1997). Note different horizontal and vertical scales are used in (a), (b), (c), (d) and that multimodality in (b) and (d) may, in part, reflect relatively
# 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>Codeilen: single-exon genes (bp)</td>
<td>1130</td>
<td>1251</td>
<td>1304</td>
<td>1137</td>
</tr>
<tr>
<td>Codeilen: 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>
</tbody>
</table>

**Initial probabilities:**

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<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>
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.
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?

intron

exon     5'

exon

3'

donor

intron

acceptor
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
Are A & B independent?

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>not B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>not A</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}
\]

“Expected” means expected assuming independence, e.g. expect B 10/20; A 12/20; both 120/400*20 = 6, etc.

Look up in table (or approximate as normal).
\( \chi^2 \) test for independence

<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
GHMM Structure
Summary of Burge & Karlin

Coding DNA & control signals are 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 important 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?