CSE 427
Computational Biology

Genes and Gene Prediction
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 these slides
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, mouse, a few others
  unlikely for most others
  and non-coding is poorly understood even in human
Biological Basics

Central Dogma:

\[
\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{Protein}
\]

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.)
Translation: mRNA → Protein

Watson, Gilman, Witkowski, & Zoller, 1992
Ribosomes
DNA (thin lines), RNA Pol (Arrow), mRNA with attached Ribosomes (dark circles)

Figure 3-7. Coupled transcription/translation in bacteria is visualized. Oscar Miller and colleagues lysed E. coli cells and immediately collected the cell contents on electron microscope grids. They saw threads of mRNA still associated with DNA (thin lines), and ribosomes—several at a time—were already translating protein along the mRNA. Thus, in bacterial cells, the picture of information recovery and use, at least in broad outline, was complete: mRNA was made on demand; ribosomes recognized the 5’ end of the mRNA, bound, and began protein synthesis even before the mRNA had been completely synthesized. (In this photo, the arrow indicates a presumptive RNA polymerase [the faint disk to the left of the first ribosome]. The DNA thread at the top is being copied into mRNA, but the one at the bottom is not. Both are presumably double stranded.) (Reprinted, with permission, from Miller et al. 1970 [©AAAS].)
## 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, Ser</td>
<td>Tyr, Cys, U</td>
</tr>
<tr>
<td></td>
<td>Phe, Ser</td>
<td>Tyr, Cys, C</td>
</tr>
<tr>
<td></td>
<td>Leu, Ser</td>
<td>Stop, Stop, A</td>
</tr>
<tr>
<td></td>
<td>Leu, Ser</td>
<td>Stop, Trp, G</td>
</tr>
<tr>
<td>C</td>
<td>Leu, Pro</td>
<td>His, Arg, U</td>
</tr>
<tr>
<td></td>
<td>Leu, Pro</td>
<td>His, Arg, C</td>
</tr>
<tr>
<td></td>
<td>Leu, Pro</td>
<td>Gln, Arg, A</td>
</tr>
<tr>
<td></td>
<td>Leu, Pro</td>
<td>Gln, Arg, G</td>
</tr>
<tr>
<td>A</td>
<td>Ile, Thr</td>
<td>Asn, Ser, U</td>
</tr>
<tr>
<td></td>
<td>Ile, Thr</td>
<td>Asn, Ser, C</td>
</tr>
<tr>
<td></td>
<td>Ile, Thr</td>
<td>Lys, Arg, A</td>
</tr>
<tr>
<td></td>
<td>Met/Start, Thr</td>
<td>Lys, Arg, G</td>
</tr>
<tr>
<td>G</td>
<td>Val, Ala</td>
<td>Asp, Gly, U</td>
</tr>
<tr>
<td></td>
<td>Val, Ala</td>
<td>Asp, Gly, C</td>
</tr>
<tr>
<td></td>
<td>Val, Ala</td>
<td>Glu, Gly, A</td>
</tr>
<tr>
<td></td>
<td>Val, Ala</td>
<td>Glu, Gly, G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acids</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Alanine</td>
</tr>
<tr>
<td>Arg</td>
<td>Arginine</td>
</tr>
<tr>
<td>Asn</td>
<td>Asparagine</td>
</tr>
<tr>
<td>Asp</td>
<td>Aspartic acid</td>
</tr>
<tr>
<td>Cys</td>
<td>Cysteine</td>
</tr>
<tr>
<td>Gln</td>
<td>Glutamine</td>
</tr>
<tr>
<td>Glu</td>
<td>Glutamic acid</td>
</tr>
<tr>
<td>Gly</td>
<td>Glycine</td>
</tr>
<tr>
<td>His</td>
<td>Histidine</td>
</tr>
<tr>
<td>Ile</td>
<td>Isoleucine</td>
</tr>
<tr>
<td>Leu</td>
<td>Leucine</td>
</tr>
<tr>
<td>Lys</td>
<td>Lysine</td>
</tr>
<tr>
<td>Met</td>
<td>Methionine</td>
</tr>
<tr>
<td>Phe</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>Pro</td>
<td>Proline</td>
</tr>
<tr>
<td>Ser</td>
<td>Serine</td>
</tr>
<tr>
<td>Thr</td>
<td>Threonine</td>
</tr>
<tr>
<td>Trp</td>
<td>Tryptophane</td>
</tr>
<tr>
<td>Tyr</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>Val</td>
<td>Valine</td>
</tr>
</tbody>
</table>
Idea #1: Find Long ORF’s

**Reading frame**: which of the 3 possible sequences of triples does the ribosome read?

**Open Reading Frame**: No internal stop codons

In random DNA

- average ORF $\frac{64}{3} \approx 21$ 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 3\textsuperscript{rd} base
  Why? E.g. efficiency, histone, enhancer, splice interactions
Idea #3: Non-Independence

Not only is codon usage biased, but residues (aa or nt) in one position are not independent of neighbors.

How to model this? Markov models.
A Markov Model (1st order)

States: A, C, G, T
Emissions: corresponding letter
Transitions: $a_{st} = P(x_i = t \mid x_{i-1} = s)$
$k^{th}$ Order Markov Model

$4^k$ states, each in/out-degree 4, joined as follows:

- $A_{xyz}^{k-1}$ emit $A$
- $C_{xyz}^{k-1}$ emit $C$
- $G_{xyz}^{k-1}$ emit $G$
- $T_{xyz}^{k-1}$ emit $T$

(where $xyz$ is a length $k-1$ string over $\{A,C,G,T\}$, and, e.g., $A_{xyz}$ and $xyzA$ are the same state when $xyz=A^{k-1}$)
E.g., a 2\textsuperscript{nd} order Markov Model
Summary: Prokaryotic Genes

In prokaryotes, most DNA is coding

E.g. ~70% in *H. influenzae* (vs <2% in humans)

Long ORFs + codon/nucleotide stats do well

Can improve by modeling associated features (TATA boxes, promoters, etc.)

  e.g. via WMM or higher-order Markov models

But obviously won’t be perfect

  short genes, frame shifts, 5’ & 3’ UTR’s, …
GENES, PART II
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.
Hints to Origins?

Tetrahymena thermophila
As in prokaryotes
promoters
start/stop transcription
start/stop translation

New Features:
introns, exons, splicing
branch point signal
alternative splicing
polyA site/tail
### 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><strong>Internal exon</strong></td>
<td>122 bp</td>
<td>145 bp</td>
<td>RefSeq alignments to draft genome sequence, with</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>confirmed intron boundaries (43,317 exons)</td>
</tr>
<tr>
<td><strong>Exon number</strong></td>
<td>7</td>
<td>8.8</td>
<td>RefSeq alignments to finished seq (3,501 genes)</td>
</tr>
<tr>
<td><strong>Introns</strong></td>
<td>1,023 bp</td>
<td>3,365 bp</td>
<td>RefSeq alignments to finished seq (27,238 introns)</td>
</tr>
<tr>
<td><strong>3’ UTR</strong></td>
<td>400 bp</td>
<td>770 bp</td>
<td>Confirmed by mRNA or EST on chromo 22 (689)</td>
</tr>
<tr>
<td><strong>5’ UTR</strong></td>
<td>240 bp</td>
<td>300 bp</td>
<td>Confirmed by mRNA or EST on chromo 22 (463)</td>
</tr>
<tr>
<td><strong>Coding seq</strong></td>
<td>1,100 bp</td>
<td>1340 bp</td>
<td>Selected RefSeq entries (1,804)*</td>
</tr>
<tr>
<td><strong>(CDS)</strong></td>
<td>367 aa</td>
<td>447 aa</td>
<td></td>
</tr>
<tr>
<td><strong>Genomic span</strong></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. Variation in size distribution of coding sequences and exons is less extreme, but 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
Exons

Introns

Nature 2/2001
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.
Other Relevant Features

PolyA Tails
100-300 A’s typically added to the 3′ end of the mRNA after transcription—*not* templated by DNA

Processed pseudogenes
Sometimes mRNA (*after* splicing + polyA) is reverse-transcribed into DNA and re-integrated into genome
~14,000 in human genome
Alternative Splicing

- Exon skipping/inclusion
- Alternative 3’ splice site
- Alternative 5’ splice site
- Mutually exclusive exons
- Intron retention

These are *regulated*, not just errors
Other Features (cont)

Alternative start sites (5’ ends)
Alternative PolyA sites (near 3’ ends)
Alternative splicing

Collectively, these affect an estimated 95% of genes, with ~10 (a wild guess) isoforms per gene (but can be huge; fly Dscam: 38,016, potentially)

Trans-splicing and gene fusions
  (rare in humans but important in other organisms & some tumors)
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

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

<table>
<thead>
<tr>
<th>Program</th>
<th>Sn</th>
<th>Sp</th>
<th>Sn</th>
<th>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>GenID</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>

<table>
<thead>
<tr>
<th>Program</th>
<th>Sn</th>
<th>Sp</th>
<th>Sn</th>
<th>Sp</th>
<th>Avg.</th>
<th>ME</th>
<th>WE</th>
</tr>
</thead>
<tbody>
<tr>
<td>per exon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>per nuc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy

After Burge & Karlin, Table 1. Sensitivity, $Sn = TP/AP$; Specificity, $Sp = TP/PP$
Generalized Hidden Markov Models

States: 1, 2, …

$\pi$: Initial state distribution

$a_{ij}$: Transition probabilities

One submodel per state

Output string generated by submodel not just 1 letter, potentially variable length

Given length $L$

Pick start state $q_1$ ($\sim \pi$)

While $\sum d_i < L$

Pick $d_i$ & string $s_i$ of length $d_i \sim$ submodel for $q_i$

Pick next state $q_{i+1}$ ($\sim a_{ij}$)

Output $s_1s_2…$
GHMM Structure in Genscan
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 $D$. 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><strong>C + G% range</strong></td>
<td>&lt;43</td>
<td>43-51</td>
<td>51-57</td>
<td>&gt;57</td>
</tr>
<tr>
<td><strong>Number of genes</strong></td>
<td>65</td>
<td>115</td>
<td>99</td>
<td>101</td>
</tr>
<tr>
<td><strong>Est. proportion single-exon genes</strong></td>
<td>0.16</td>
<td>0.19</td>
<td>0.23</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Codelen: single-exon genes (bp)</strong></td>
<td>1130</td>
<td>1251</td>
<td>1304</td>
<td>1137</td>
</tr>
<tr>
<td><strong>Codelen: multi-exon genes (bp)</strong></td>
<td>902</td>
<td>908</td>
<td>1118</td>
<td>1165</td>
</tr>
<tr>
<td><strong>Introns per multi-exon gene</strong></td>
<td>5.1</td>
<td>4.9</td>
<td>5.5</td>
<td>5.6</td>
</tr>
<tr>
<td><strong>Mean intron length (bp)</strong></td>
<td>2069</td>
<td>1086</td>
<td>801</td>
<td>518</td>
</tr>
<tr>
<td><strong>Est. mean transcript length (bp)</strong></td>
<td>10866</td>
<td>6504</td>
<td>5781</td>
<td>4833</td>
</tr>
<tr>
<td><strong>Isochore</strong></td>
<td>L1+L2</td>
<td>H1+H2</td>
<td>H3</td>
<td>H3</td>
</tr>
<tr>
<td><strong>DNA amount in genome (Mb)</strong></td>
<td>2074</td>
<td>1054</td>
<td>102</td>
<td>68</td>
</tr>
<tr>
<td><strong>Estimated gene number</strong></td>
<td>22100</td>
<td>24700</td>
<td>9100</td>
<td>9100</td>
</tr>
<tr>
<td><strong>Est. mean intergenic length</strong></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><strong>Intergenic (N)</strong></td>
<td>0.892</td>
<td>0.867</td>
<td>0.54</td>
<td>0.418</td>
</tr>
<tr>
<td><strong>Intron (I+, I-)</strong></td>
<td>0.095</td>
<td>0.103</td>
<td>0.338</td>
<td>0.388</td>
</tr>
<tr>
<td><strong>5’ Untranslated region (F+, F-)</strong></td>
<td>0.008</td>
<td>0.018</td>
<td>0.077</td>
<td>0.122</td>
</tr>
<tr>
<td><strong>3’ Untranslated region (T+, T-)</strong></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
L ~ Unif(14-20), s ~ null (indp, intergenic freqs)
8 bp cap signal WMM

30% TATA-less

40 bp null
What do splice sites look like?

5’ exon

intron

donor

acceptor

3’ exon
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 & polypyrimididine 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, …”
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
Many dependencies, such as 5'/3' compensation, e.g. G₁ vs G₅/H₅.
$\chi^2$ test: Are events A & B independent?

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

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

Significance: table look up (or approximate as normal)
χ² test for independence of nucleotides in donor sites

<table>
<thead>
<tr>
<th>i</th>
<th>Con 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>
</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>
</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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>

* means chi-squared p-value < .001

Technically – build a 2 x 4 table for each (i,j) pair:
Pos i does/does not match consensus vs pos j is A, C, G, T
calculate χ² as on previous slide, e.g. χ² for +6 vs -1 = 103.8
If independent, you’d expect χ² ≤ 16.3 all but one in a 1000 times.
Summary of Burge & Karlin

Coding DNA & control signals are nonrandom
  Weight matrices, WAMs, etc. for controls
  Codon frequency, MMs, 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
(Of course we can now do better for human, mouse, etc., but what about cockatoos or cows or endangered frogs, or …)
Problems with all methods

- Pseudo genes (~14,000 in human)
- Short ORFs
- Sequencing errors
- Non-coding RNA genes & spliced UTR’s
- Overlapping genes
- Alternative TSS/polyadenylation/splicing
- 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? If that protein is an important player in some pathway, are related genes also present?

Comparative genomics - what does this region look like in related organisms?

And of course – get experimental data