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 this lecture
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

Some References
(more on schedule page)

- An extensive online bib
- A good intro survey
- A gene finding bake-off
Biological Basics

Central Dogma:
DNA → RNA → Protein

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

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>Ser</td>
</tr>
<tr>
<td>C</td>
<td>Phe</td>
<td>Ser</td>
</tr>
<tr>
<td>A</td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td>G</td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td>G</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>A</td>
<td>Met/Start</td>
<td>Thr</td>
</tr>
<tr>
<td>U</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>C</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>A</td>
<td>Val</td>
<td>Ala</td>
</tr>
</tbody>
</table>

Translation: mRNA → Protein

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
Why? E.g. histone, enhancer, splice interactions

Recognizing Codon Bias

Assume
Codon usage i.i.d.; abc with freq. f(abc)
a_{1}a_{2}a_{3}…a_{3n+2} is coding, unknown frame
Calculate
p_{1} = f(a_{1}a_{2}a_{3})f(a_{4}a_{5}a_{6})…f(a_{3n-2}a_{3n-1}a_{3n})
p_{2} = f(a_{1}a_{3}a_{4})f(a_{2}a_{5}a_{6})…f(a_{3n-1}a_{3n}a_{3n+1})
p_{3} = f(a_{1}a_{4}a_{5})f(a_{2}a_{3}a_{6})…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 & other signals
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

exchange of U1 for U6

Figure 5. Sequence Characteristics of the Spliceosome's Mechanical Gadgets

(A) Examples of domain structure. DEAD and DEAH, helicase-like domains; C-domain, conserved in the DEAH proteins; S1, a ribosomal motif implicated in RNA binding; RS, rich in serine/arginine dipeptides; RRM, RNA recognition motif; EF-2, elongation factor 2. All factors are from *S. cerevisiae* except for the mammalian factors U2AF65 and HRH1, the human ortholog of Prp22. (B) Sequence motifs of the DExD/H box domains. DEAD, residues identical between Prp5, Prp28, and U5/100 kDa (*Table 1*). DEAH, amino acid residues identical between Prp2, Prp16, Prp22, Prp43, hPRP16, and HRH1 (*Table 1*). x, any amino acid. The specific sequences for the HCV RNA unwindase and Rep are shown for comparison. Pink, residues common to all compared sequences. Yellow, residues common to all superfamily II sequences. Note well: from a structural comparison, motif IV of superfamily II does not align with motif IV of superfamily I; rather, motif IV of superfamily II aligns with a novel superfamily I motif designated IVa (*Figure 8*; [29]).
Figure 6. A Paradigm for Unwinder Specificity and Timing? The DExD/H box protein UAP56 (orange) binds U2AF^{65} (pink) through its linker region (L). U2 binds the branch point. Y’s indicate the polypurine tract; RS, RRM as in Figure 5A. Sequences are from mammals.

Figure 7. A Parallel between the Spliceosome and the Ribosome? The binding of a yeast Phe codon by the anticodon loop of the cognate tRNA is compared with the binding of a 5′ exon by the yeast U5 loop in a hypothetical, yet provocative, configuration. N, any nucleotide.

Hints to Origins?

Tetrahymena thermophila
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></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>1340bp</td>
<td>Selected RefSeq entries (1,804)*</td>
</tr>
<tr>
<td>(CDS)</td>
<td>367 aa</td>
<td>447 aa</td>
<td>Selected RefSeq entries (1,804)*</td>
</tr>
<tr>
<td>Genomic span</td>
<td>14 kb</td>
<td>27 kb</td>
<td>* 1,804 selected RefSeq entries were those with full-length unambiguous alignment to finished sequence</td>
</tr>
</tbody>
</table>

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 = 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 centred 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 centred 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 nucleotide)</th>
<th>Sp (per nucleotide)</th>
<th>Sn (per exon)</th>
<th>Sp (per exon)</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>0.80</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>0.46</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>0.40</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 \( L \)
  - Pick start state \( q_1 \) (~\( \pi \))
  - While \( \sum d_i < L \)
    - Pick \( d_i \)
    - Pick string \( s_i \) of length \( d_i = |s_i| \) ~ submodel for \( q_i \)
    - Pick next state \( q_{i+1} \) (~\( a_{ij} \))
  - Output \( s_1s_2... \)

Decoding

- A “parse” \( \phi \) of \( s = s_1s_2...s_n \) is a pair \( d = d_1d_2...d_k \) \( q = q_1q_2...q_k \) with \( \sum d_i = L \)
- Now use something like the forward/backward algorithms to calculate probabilities like “\( P(\text{seq up to position } i \text{ generated ending in state } q_k) \)” which involves summing over possible predecessor states \( q_{k-1} \) and possible \( d_k \)

\[
Pr(\phi(s)) = \frac{Pr(\phi + s_i)}{Pr(s_i)} ...
\]
**Length Distributions**

- AT-rich avg: 2069
- CG-rich avg: 518

**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 L1+L2</td>
<td>H1+H2</td>
<td>H3</td>
<td>H3</td>
<td></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>63000</td>
<td>36000</td>
<td>5400</td>
<td>2600</td>
</tr>
</tbody>
</table>

**Initial probabilities:**
- Intergenic (N) 0.892 0.867 0.54 0.418
- Introns (I+, I- ) 0.095 0.103 0.338 0.388
- 5' Untranslated region (F+, F-) 0.008 0.018 0.077 0.122
- 3' Untranslated region (T+, T-) 0.005 0.011 0.045 0.072

---

**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?

© W.L.Ruzzo & UW CSE 1994-2006
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.

**$\chi^2$ test for independence**

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

*expected* means expected assuming independence.

* means chi-squared p-value < .001.

<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.
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 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?