Outline

• Recap: Prokaryotic gene finding
• Eukaryotic gene finding
• The human gene complement
• Regulation

Gene Finding Questions

• Identify protein coding region
• Identify Open Reading Frame
• Predict mRNA (including UTR’s)
• Predict intron/exon structure
  Eukaryotes only
• Regulatory signals
• Protein sequence

Gene criteria

Snyder and Gerstein, Science 2003

• Open Reading Frames (ORFs)
  Computational
• Sequence features
  Computational
• Sequence conservation
  Computational
• Evidence for transcription
  Experimental
• Gene inactivation induces a phenotype
  Computational

Sequence features

• Coding statistics (e.g. codon bias)
• Gene structure

Pairwise sequence alignment (global and local)
Multiple sequence alignment
Substitution matrices
Database searching
BLAST
Sequence statistics
Evolutionary tree reconstruction
Prokaryotic Gene Finding
Eukaryotic Gene Finding

Online FCE’s: Thru Dec 10

• Tues, Nov 30:
  Gene Finding 1
• Thurs, Dec 2:
  Gene Finding 2, PS5 due
• Tues, Dec 7:
  Project presentations 1
• Thurs, Dec 9
  Project presentations 2
  Final papers due
• Tues, Dec 14:
  DD: Extended office hours: 2:30pm – 5:30pm, MI 650
• Wed, Dec 15
  NS: office hours. DH 1321, noon – 2pm.
• Friday Dec 17
  8:30am Final Exam, Room: TBA
An HMM that finds genes in *E. coli*

Krogh *et al.*, 1995

Outstanding Problems

- Model cannot account for drift in CG content
- Does not take position dependencies into account
- Solution:
  - 6th order Markov chain
  - Looks back 6 positions
- Glimmer (Salzberg *et al.*, 1998)
  - Finds 98% of all genes in a bacterial genome.

Some Problems

- Overlapping genes
- Alternate splicing

Gene Finding Challenges

- Small protein-coding genes (<100 aa’s)
- RNA-coding genes
- Regulatory regions
- Genes with sparse conserved positions and little sequence similarity; e.g., beta-defensins

Some Problems

- Overlapping genes
- Alternate splicing
- Pseudogenes

Gene Finding Challenges

- Small protein-coding genes (<100 aa’s)
- RNA-coding genes
- Regulatory regions
- Genes with sparse conserved positions and little sequence similarity; e.g., beta-defensins

Schutte *et al.*, PNAS, 2001


Outline

• Recap: Prokaryotic gene finding
• Eukaryotic gene finding
• The human gene complement
• Regulation

Prokaryotic vs. Eukaryotic Genes

• Prokaryotes
  – small genomes (0.5Mb to 10Mb)
  – high gene density (90%)
  – no introns (or splicing)
  – no RNA processing
  – simple regulatory regions
  – most long ORF’s are genes

• Eukaryotes
  – large genomes
  – low gene density (3% - 50%)
  – intron/exon structure
  – splicing
  – complex regulatory regions

Genomic data:
Must handle multiple genes and/or gene fragments in input sequence.

Typical human gene sizes

<table>
<thead>
<tr>
<th>Average gene lengths</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding region</td>
<td>1-2kb</td>
</tr>
<tr>
<td>Exon lengths</td>
<td>100 - 200 bp</td>
</tr>
<tr>
<td>Exon count</td>
<td>5-6</td>
</tr>
<tr>
<td>Single exon genes</td>
<td>6%</td>
</tr>
</tbody>
</table>

Genome statistics

<table>
<thead>
<tr>
<th>Size (1 gene per)</th>
<th>Gene number</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>3300Mb</td>
<td>33k</td>
</tr>
<tr>
<td>Fly</td>
<td>150Mb</td>
<td>12.5k</td>
</tr>
<tr>
<td>C. elegans</td>
<td>98Mb</td>
<td>13.1k</td>
</tr>
<tr>
<td>Yeast</td>
<td>12Mb</td>
<td>6.3k</td>
</tr>
<tr>
<td>E. coli</td>
<td>4.8Mb</td>
<td>3.2k</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>1.8Mb</td>
<td>1.7k</td>
</tr>
</tbody>
</table>

http://www.ornl.gov/TechResources/Human_Genome/faq/compgen.html

Typical human gene sizes

<table>
<thead>
<tr>
<th>Size (1 gene per)</th>
<th>Gene number</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>3300Mb</td>
<td>33k</td>
</tr>
<tr>
<td>Fly</td>
<td>150Mb</td>
<td>12.5k</td>
</tr>
<tr>
<td>C. elegans</td>
<td>98Mb</td>
<td>13.1k</td>
</tr>
<tr>
<td>Yeast</td>
<td>12Mb</td>
<td>6.3k</td>
</tr>
<tr>
<td>E. coli</td>
<td>4.8Mb</td>
<td>3.2k</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>1.8Mb</td>
<td>1.7k</td>
</tr>
</tbody>
</table>

http://www.ornl.gov/TechResources/Human_Genome/faq/compgen.html

Genscan

Architecture:
• Individual modules: intergenic region, promoter, 5’UTR, exon/intron, post-translation region
• Semi Hidden Markov Model – various length distributions
• Different statistical models for each module:
  – weight matrices + extensions, 3-periodic 5th order Markov chains

Incorporates:
• Descriptions of transcriptional, translational and splicing signals
• Compositional features of exons, introns, intergenic, C+G regions

Source: http://www.nslj-genetics.org/gene
Genscan

Larger predictive scope than previous models
- Partial genes
- Multiple genes separated by intergenic DNA
- Genes on either/both DNA strands

Proposed pipeline
- Screen for repetitive elements
- Predict protein sequences with GENSCAN
- BLAST predictions to find homologs
- Refine using spliced alignment of prediction with homolog (e.g., Gelfand, Mironov, Pevzner, 96)
- Verify experimentally

GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- Esngl: single exon (intronless) (translation start -> stop codon)
- Einit: initial exon (translation start -> donor splice site)
- Ek: phase k internal exon (acceptor splice site -> donor splice site)
- Eterm: terminal exon (acceptor splice site -> stop codon)
- Ik: phase k intron

How to model sequences with lengths that are not geometrically distributed?

CODON model

$P(L) = (1 - p)p^{L-1}$

Semi-hidden Markov model

- Set of states: $Q_1, Q_2, ...$
- Transition matrix $P(Q(t)|Q(t-1))$
- Initial distribution $P(Q_1)$
- Each state has
  - a length distribution
  - a sequence generating model

Emission:
Each state emits a sequence, according to a particular distribution, of length, $d$, according to a particular length frequency distribution

Semi-hidden Markov model cont’d

- A parse $\varphi$ of length $L$ is
  - A state sequence: $Q_1, Q_2, ...$
  - A sequence of lengths: $d_1, d_2, d_3, ...$

- An observed sequence, $s$, is scored using a modified Viterbi algorithm

$\varphi_{opt} = \text{arg max } P(\varphi | s)$
GenScan Training Set

2.5M base pairs
142 Single Exon Genes (SEGs)
238 multi-exon gene
1492 Exons
1254 Introns
An additional 1619 coding sequences (no introns)

Promotor model based on published sources.

Initial and transition probabilities

Trained separately for four categories of G+C content
- < 43% (G+C)
- 43% - 51% (G+C)
- 51% - 57% (G+C)
- > 57% (G+C)

- Gene density varies with G+C content
- Genes in A+T rich regions had fewer introns

GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- \( E_{\text{sgl}} \): single exon (introns less) (translation start -> stop codon)
- \( E_{\text{init}} \): initial exon (translation start -> donor splice site)
- \( E_k \): phase k internal exon (acceptor splice site -> donor splice site)
- \( E_{\text{term}} \): terminal exon (acceptor splice site -> stop codon)
- \( I_k \): phase k intron:

Intron/Exon model

- Exon phase = phase of previous intron
- Donor and acceptor models incorporated in intron models
- Length and sequence distribution determined empirically

Fig. 3, Burge and Karlin 1997
Intergenic models

- Lengths are geometrically distributed
  \[ \text{mean: } \frac{3 \times 10^9}{60,000} \] (estimated human gene number in 1997)

- Sequence model
  5th order Markov model (highest order trainable with data available).

- Similar models used for untranslated regions

GenScan States

- N: intergenic region
- P: promoter
- F: 5' untranslated region
- \( E_{\text{sgl}} \): single exon (intronsless) (translation start -> stop codon)
- \( E_{\text{ini}} \): initial exon (translation start -> donor splice site)
- \( E_{\text{term}} \): terminal exon (acceptor splice site -> stop codon)
- \( I_k \): phase k intron

Signal Models

TATA box, polyA signal, 5'UTR

- Fixed length models
- PSSMs
- No position dependence

Acceptor Model

- Fixed length model
- "Weight array model"
  - Models dependence on the previous position

Donor Model

- Fixed length model
- "Maximal Dependence Decomposition"
  - Models dependencies between nonadjacent elements

Performance measures

- \[ S_n = \frac{TP}{TP + FN} \]

Fig. 3, Burge and Karlin 1997
Performance measures
Burset & Guigo, 1996

Exon Level

\[
S_n = \frac{TP}{TP + FN} \quad S_p = \frac{TN}{TN + FP}
\]

Genscan Performance

<table>
<thead>
<tr>
<th></th>
<th>Sn</th>
<th>Sp</th>
<th>Missing Exons</th>
<th>Wrong Exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genscan</td>
<td>0.93</td>
<td>0.78</td>
<td>0.81</td>
<td>0.09</td>
</tr>
<tr>
<td>FGENEH</td>
<td>0.77</td>
<td>0.61</td>
<td>0.64</td>
<td>0.15</td>
</tr>
<tr>
<td>GENEID+</td>
<td>0.91</td>
<td>0.73</td>
<td>0.70</td>
<td>0.07</td>
</tr>
<tr>
<td>GENEARSER3</td>
<td>0.86</td>
<td>0.56</td>
<td>0.58</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Proportion of genes with all exons correctly predicted:

\[
\frac{243}{570} = 0.43
\]

Outline

• Recap: Prokaryotic gene finding
• Eukaryotic gene finding
• The human gene complement
• Regulation

Gene Prediction in the Human Genome
International Human Genome Consortium, Nature 2001

Initial Gene Index (IGI)

• Ensembl
  – GENSCAN predictions confirmed by homology to EST’s, mRNAs, proteins and protein motifs
• Genie
  – HMM-based
  – Extends homology with EST’s/mRNAs using ab initio approach
• Previously known genes in databases

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl alone</td>
<td>4057</td>
</tr>
<tr>
<td>Ensembl + Genie</td>
<td>14882</td>
</tr>
<tr>
<td></td>
<td>31,778</td>
</tr>
</tbody>
</table>

Gene Prediction Count

IGI contains:
15,000 known genes
17,000 predicted genes

Assuming
20% overprediction
fragmentation rate of 1.4

yields 32,000 genes

Assuming
40% of novel genes are not predicted

yields 31,000 genes

Validation

• Comparison with 31 newly discovered genes
  – 28/31 in draft human genome sequence
  – 19/28 were detected by gene prediction
  – IGI contains ~60% of novel genes in human genome (19/31)
• Comparison with mouse cDNA’s
  – 81% of mouse cDNA’s similar to draft human sequence
  – 69% of mouse cDNA’s similar to predicted genes
• Problems:
  – Overprediction
  – Fragmentation: 1 true gene corresponds to >1 prediction
  – Partial prediction: only part of gene is correctly predicted

Predicting Human Genome Count

yields 24,500 genes

yields 31,000 genes
Outline

• Recap: Prokaryotic gene finding
• Eukaryotic gene finding
• The human gene complement
• Regulation

Examples of binding site motifs

<table>
<thead>
<tr>
<th>Conserved genes</th>
<th>Organism 1</th>
<th>Conserved regulatory elements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Organism 2</td>
<td></td>
</tr>
</tbody>
</table>

Identifying binding sites
Comparative genomic approach

Global alignment of upstream sequences

Kellis et al., Nature, 2003
Comparative genomics for gene finding


Using genomic sequence from S. paradoxus, S. bayanus, S. mikatae
Kellis et al. (Nature, 2003) found
503 false predictions
43 new small genes
42 new regulatory motifs
Requires sequences in species that are close but not too close.

Will this approach work in higher eukaryotes?

One of the first examples of this approach:


We describe a simple algorithm for computing a homology score for Escherichia coli promoters based on DNA sequence alone. The homology score was related to 31 values, measured in vitro, of RNA polymerase selectivity, which we define as the product KBk2, the apparent second order rate constant for open complex formation. We found that promoter strength could be predicted to within a factor of +/-4.1 in KBk2 over a range of 10(4) in the same parameter. The quantitative evaluation was linked to an automated (Apple II) procedure for searching and evaluating possible promoters in DNA sequence files.

<table>
<thead>
<tr>
<th>Whole genome sequencing</th>
<th>Gene finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>Assessment of protein coding measures (Fickett &amp; Tung)</td>
</tr>
<tr>
<td>1995</td>
<td>H. influenzae, 1st whole genome sequence</td>
</tr>
<tr>
<td>1997</td>
<td>Yeast, 1st eukaryote</td>
</tr>
<tr>
<td>1998</td>
<td>C. elegans 1st multicellular organism</td>
</tr>
<tr>
<td>1999</td>
<td>Glimmer: Higher order Markov models for prokaryotes (Salzberg et al)</td>
</tr>
<tr>
<td>2000</td>
<td>GeneScan: Prediction of complete gene structures in human genomic DNA (Huang &amp; Kadam)</td>
</tr>
<tr>
<td>2001</td>
<td>Human, 1st mammal</td>
</tr>
<tr>
<td>2002</td>
<td>Mouse</td>
</tr>
<tr>
<td></td>
<td>Kellis et al, 2003, comparative genomics for finding regulatory regions</td>
</tr>
</tbody>
</table>