Gene Structure & Gene Finding

Russ B. Altman
BMI 214
CS 274

Gene Finding

Problem: Identify the genes within raw genomic DNA sequence.

Input: Raw DNA sequence

Output: Location of gene elements in the raw sequence, including exons, introns.

Gene finding in bacteria (procaryotes)

Somewhat straightforward to find. Features of genes in bacteria include:

- Promoter sequence in 5' flanking region
- Start with ATG (codes for amino acid MET)
- End with termination codons (e.g. TAA)
- Usually > 300 amino acids in length
- Different composition (AT vs. GC ratio) in coding vs. noncoding regions
- Shine-Dalgarno sequence in 5' UTR allows ribosome to bind the mRNA
Gene finding in eukaryotes (including humans)
Relatively difficult. Features include:

- Transcription signals affect level and location of transcription
- Translational signals affect location of amino acid codons
- Splicing sites define locations of introns/exons
- More variability in size, composition
- More important stuff far away from gene…

Splicing = take out introns to create coding region mRNA

Regulatory Elements

Lots of Factors and Co-Factors

Transcription Signals

Initiator/CAP signal—a sequence motif at the transcription start site (TSS), present in 70%

TATA box—usually about 30 basepairs upstream of TSS, present in 70%

PolyA site—at 3' end, AATAAA hexamer + variable sequence rich in A's 20-30 bp away, present in 50%
Transcription signals
Enhancers—sequence motifs associated with an increased level of expression. Can be far away in the genome.
Silencers—sequence motifs associated with a decreased level of expression.
CpG methylation sites—areas where DNA is modified (rich in GC) to “turn it off.”
Chromatin structure, DNA curvature—how the DNA is folded affects how accessible it is.

Translational signals
“Kozak” signal—upstream sequence motif found before the starting ATG. Consensus = (GCC)GCCGCCATGG
Starting ATG—difficult to identify in eucaryotic genes very often.

Splicing is a killer
Eucaryotic genes have segments of the coding sequence (exons) separated by segments that don’t code for amino acids and are removed after transcription (introns).
The identification of introns is difficult, but there are some clues based on our understanding of biology.
Begin with “GT”
End with “AG”
Splice Sites

Splicing Process

Distribution of bases around DONOR “GT” site...

Alternative Splicing

• First predicted by Walter Gilberts in 1978

• Well documented for several genes such as calcitonin and immunoglobulin mu by the early 1980’s

• Represented a fundamentally new principle about the genetic message
Roles of Alternative Splicing

(a) 

(b) 

(c) 

(d) 

(e) 

Many alt-splicing DBs...

Alternative Splicing DB

DB CONTENT | HOW TO USE | FURTHER WORK | SEARCH

References to the Alternative Splicing Database:

ASDB: databases of alternatively spliced genes


ACOX1 is alternatively spliced.

Genetic Diseases

A

B

C

Use ESTs to find alternatively spliced genes

EST = Expressed Sequence Tag

Genomic DNA

mRNA

(Complementary)
cDNA

Sequence

cDNA

partially.

Transcription by cell...

3' EST

5' EST

Methods for Gene Finding

1. Use metaphor of parsing from linguistics & CS.

2. Use neural networks

3. Use Markov models of sequence elements

4. Use mixed probabilistic models of sequence elements (best performance).
Strict Grammar Approaches

Linguistics:
- subject - verb = I am.
- subject - verb - object = I am Russ.
- subject phrase - verb phrase - object phrase = The person speaking used to be a student.

Biology: exon-intron = XXXXXXIIIIIIIIIIIIII
- exon - intron - exon = etc...

GRAIL = Neural Network Approach
Takes advantage of hexamers (6-tuples), sequences of six DNA bases.

Total = $4^6 = 4096$

Computer probability $p$(hexamer) in each of the three frames on a strand of DNA.
Repeat analysis on reverse complement strand as well...

N-order Markov Chains
What is $P$(ACTGTC)?

N=0:
$p(A)p(C)p(T)p(G)\ldots p(C)$

N=1:
$p(A)p(C|A)p(T|AC)p(G|CT)\ldots p(C|GT)$

N=2:
$p(A)p(C|A)p(T|AC)p(G|CT)\ldots p(C|GT)$

N=3:
$p(A)p(C|A)p(T|AC)p(G|ACT)\ldots p(C|TGT)$
P(ACGGT) = 0.2 * 0.3 * 0.1 * 0.1 * 0.4
P(TTTTT) = 0.4 * 0.4 * 0.4 * 0.4 * 0.4

GENSCAN Algorithmic Overview
(see paper in course reader... Burge & Karlin)

1. A parse is a set of states (q1, q2...qn, n = 27) corresponding to intron, exon, etc...and lengths (d1, d2...dm).
2. Models are built for each state describing the internal structure and length distribution.
3. Given a sequence, find the most likely parse of that sequence using the models.

GENSCAN sources of information
GC content
Introns/exon
Mean lengths of introns/exons
Mean length of transcript
Mean length of inter-gene regions
Expected number of genes
Signal models for coding, transcription, translational, splice
Reverse strand
How does one “rate” gene predictions?
Define a gold standard: experimentally proven gene structure

- Evidence of mRNA in cell with exons in correct order
- Evidence of proteins with predicted sequence in the cell
- Comparative genomic evidence (find conserved exons in related species…)
From CASP to GASP

1. Not quite perfectly controlled as CASP.
2. Set of “known good” genes for drosophila published.
3. All gene finding groups invited to submit their gene predictions.
4. Assess and compare the predictions.

GASP Competition

Exon Level

<table>
<thead>
<tr>
<th>REA D L Y</th>
<th>W R O N G E X O N S</th>
<th>C O R R E C T E X O N S</th>
<th>M I S S I N G E X O N S</th>
</tr>
</thead>
<tbody>
<tr>
<td>P R E D I C T I O N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_T = \frac{\text{number of Correct Exons}}{\text{number of Actual Exons}}$</td>
<td>Sensitivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_S = \frac{\text{number of Correct Exons}}{\text{number of Predicted Exons}}$</td>
<td>Specificity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{FPR} = \frac{\text{number of Wrong Exons}}{\text{number of Predicted Exons}}$</td>
<td>(Sensitivity)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Nucleoside Level

<table>
<thead>
<tr>
<th>R E A D Y</th>
<th>F P N</th>
<th>T P</th>
<th>F P P</th>
<th>T P T</th>
<th>F P F</th>
<th>T P</th>
<th>N T</th>
<th>F N</th>
<th>T N</th>
</tr>
</thead>
<tbody>
<tr>
<td>P R E D I C T I O N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_T = \frac{\text{FP}}{\text{TN} + \text{FP}}$</td>
<td>Sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_S = \frac{\text{TP}}{\text{TP} + \text{FP}}$</td>
<td>Specificity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$CC = \frac{\text{FP} \times \text{TN} - \text{FN} \times \text{FP}}{\sqrt{\text{FP} \times \text{FN} + \text{TP} \times \text{FN} + \text{FP} \times \text{TN} + \text{FN} \times \text{TN}}}$

Correlation Coefficient

$ACF = \frac{\text{TP} \times \text{FP}}{\text{TP} \times \text{FP} + \text{FP} \times \text{FN} + \text{TP} \times \text{TN} + \text{FN} \times \text{TN}}$

Approximate Correlation

Program | Original | Test | Prediction | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity | Mixed errors | Wrong errors |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CASP5A</td>
<td>32</td>
<td>50</td>
<td>49</td>
<td>95</td>
<td>72</td>
<td>76</td>
<td>50</td>
<td>28</td>
<td>26</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>GASP</td>
<td>40</td>
<td>39</td>
<td>40</td>
<td>96</td>
<td>81</td>
<td>16</td>
<td>30</td>
<td>28</td>
<td>26</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Drosophila</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>
GASP questions

1. Are the gene predictions similar to the known gene structures?
2. Are the details of the gene predictions correct (e.g., splice sites)?
3. What other DNA features (besides gene structure) can be reliably identified?
4. Which analysis methods are the most effective?


Human genes by class (from GO)

Overall statistics of human genes

Table 20 Properties of genome and proteome in essentially completed eukaryotic proteomes

<table>
<thead>
<tr>
<th>Class</th>
<th>Homo sapiens</th>
<th>Py.</th>
<th>Worms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of identified genes</td>
<td>~32,000</td>
<td>13,208</td>
<td>18,296</td>
</tr>
<tr>
<td>% with annotated proteomes</td>
<td>31</td>
<td>50</td>
<td>52</td>
</tr>
<tr>
<td>Number of annotated proteomes</td>
<td>1,332</td>
<td>1,055</td>
<td>1,014</td>
</tr>
<tr>
<td>Number of annotated proteome families</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of distinct proteome architectures</td>
<td>1,200</td>
<td>1,000</td>
<td>1,010</td>
</tr>
<tr>
<td>Percentage of % 1-3</td>
<td>1-40</td>
<td>4-60</td>
<td>3-10</td>
</tr>
<tr>
<td>% Signal peptides</td>
<td>30</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>% Transmembrane proteins</td>
<td>30</td>
<td>20</td>
<td>26</td>
</tr>
<tr>
<td>% Repeats containing</td>
<td>10</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>% COG/10</td>
<td>11</td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

Most common protein families in human and friends...

Genes we borrowed from bacteria...
Conclusions

1. Finding genes in eucaryotic organisms is difficult, a bit easier in procaryotes.
2. Current performance is around 90%
3. Gold standard is experimental proof that transcript and protein product exist (or at least one of those).
4. Best methods are based on detailed probabilistic models that include all known effects.