Sequence Analysis '17
Lecture 21

Splicing
Gene finding
HMMs can be modular
Gene Prediction: Computational Challenge

Gene!

Thanks to Pavel Pevzner (UCSD) and Chris Burge (MIT) for slides and images.
Two Approaches to Gene Prediction

- **Statistical**: coding segments (ORFs or exons) have typical sequences on either end and use different subwords than non-coding segments (introns or intergenic regions).

- **Similarity-based**: many human genes are similar to genes in mice, chicken, or even bacteria. Therefore, already known mouse, chicken, and bacterial genes may help to find human genes.
Gene Prediction and Motifs

• Upstream regions of genes often contain motifs that can be used for gene prediction

-35
TTCCAA
Pribnow Box

-10
TATACT

0
ATG

10
GGAGG
Ribosomal binding site

Transcription start site

• Reading frame should be “long enough”.
Promoter Structure in Prokaryotes (E.Coli)

Transcription starts at offset 0.

- Pribnow Box (-10)
- Gilbert Box (-30)
- Ribosomal Binding Site (+10)

Promoter structure in prokaryotes

Consensus sequences
Ribosomal Binding Site

1055 E. coli Ribosome binding sites listed in the Miller book
Central Dogma of biology

DNA

transcription

RNA

translation

Protein

CCTGAGCCCAACTATTGATGAA

CCUGAGCCCAACUAUUUGAUGAA

PEPTIDE
Discovery of Split Genes

• In 1977, Phillip Sharp and Richard Roberts experimented with mRNA of *hexon*, a viral protein.

  – Map hexon mRNA in viral genome by hybridization to adenovirus DNA and electron microscopy

  – mRNA-DNA hybrids formed three curious loop structures instead of contiguous duplex segments
Expanded Central Dogma

exon1  intron1  exon2  intron2  exon3

transcription

splicing

translation

folding

exon = coding
intron = non-coding

replication
Splicing mechanism

GU 

A 

AG 

exon1 

exon2

5’ splice site

“branchpoint”

3’ splice site

“the acceptor”

“the donor”

Spliceosome: Def: A ribonucleoprotein complex, containing RNA and small nuclear ribonucleoproteins (snRNPs) that is assembled during the splicing of messenger RNA primary transcript to excise an intron.

(http://genes.mit.edu/chris/)
Splicing mechanism

sans spliceosome

(1) pre-mRNA

(2) lariat loop forms

Spliceosome (not shown) forms.

(3) exon 1 cleaved from lariat.

3'-OH
Splicing mechanism

(4) exon 1,2 positioned to ligate

(5) liberated lariat + ligated exons

goes to ribosome

gets degraded

Spliceosome (not shown) disassociates.
Splicing mechanism on the web

http://neuromuscular.wustl.edu/pathol/diagrams/splicefunct.html

google: wustl neuromuscular splicefunct

http://neuromuscular.wustl.edu/pathol/diagrams/splicemech.html

google: wustl neuromuscular splicemech

Much thanks to T. Wilson, UCSC!
Evolutionary models for introns

How to find splice points, using BLAST.

1. Translate the DNA in all 6 frames.

2. Search the database of protein sequences using the translations (blastx).

3. Using the complete protein sequence, align it to the translation and find the regions of (near) perfect identity. These will abruptly end at the intron start site.

4. Find the 5’-GT or 3’-AG signal at the point where the identity matches abruptly end.

5. If your translation has an insertion with nearly perfect matches on either side, you have an alternative splicing.
In Class exercise: find the alternative spliced variants

Go to NCBI, search nucleotides for AKAP9 (you should get the sequence with accession number NM_005751.4, GI:197245395)

Select “BLAST sequence”

Select blastx (not tblastx)

Select the nr/nt database. Organism: homo sapiens

Submit.

While waiting, do exercise on the next page....
A sure sign of alternative splicing in blastx output:

<table>
<thead>
<tr>
<th>Score</th>
<th>Identities</th>
<th>Positives</th>
<th>Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>160 bits (404), Expect = 8e-37</td>
<td>85/116 (73%), Positives = 86/116 (74%)</td>
<td>+2</td>
<td></td>
</tr>
</tbody>
</table>

Identical up to the insertion. Identical after the insertion. These must be the same gene.
HMM for pre-mRNA

5'UTR

AUG

exon

stop (UAA|UAG|UGA)

3'UTR

3' splice site (...AG)

5' splice site (GU...)

intron (0|1|2)

exon

intron

exon

intron

XXX XXXX ATG ... XXX GUX ... XX AG XX ... XX GUX ... XX AG X
X ... XX TAA XXXX
Genescan HMM module for introns

short variable length intron sub-model

DSS \[ \text{donor site } = GU \]

\[ p \]

\[ 1-p \]

\[ I_{\text{fixed}} \]

\[ I_{\text{short}} \]

\[ 1 \]

\[ 1-q \]

\[ q \]

\[ I_{\text{geo}} \]

\[ I_{\text{ASS}} \]

\[ \text{acceptor site } = AG \]

fixed length + variable length intron sub-model

Intron model for mammals

A genefinding HMM: Genescan

internal exon model

middle exon model

intron models

initial exon model

terminal exon model

single exon model

Intergentic Regions

Mirrored models for reverse complement strand
Exon models, frame 1

Sequence enters the sub-model after first base in the codon, cycles through non-stop codons (Stop codons have zero transition probability), and transitions with small probability $p$ to final base, exiting in frame 1.
TwinScan

• Aligns two sequences and marks each base as gap ( - ), mismatch ( : ), match ( | ), resulting in a new alphabet of 12 letters: \( \Sigma \{ A-, A:, A |, C-, C:, C |, G-, G:, G |, T-, T:, T| \} \).

• Run Viterbi algorithm using emissions \( e_k(b) \) where \( b \in \{ A-, A:, A|, \ldots, T| \} \).
TwinScan

Exon (E) state emits matches (N|), Intron state (I) emits mismatches (N:) and gaps (N-). Gamma values (forward backward values) predict introns, exons. May be combined with motif models for intron, codon models.
Exercise X: twinscan HMM

- Assign states
- Run Viterbi algorithm
- Run F/B algorithm
- Predict intron/exon
Splicing-related motifs

ESEs, ESSs, ISEs, ISSs

**ESE** = Exonic Splicing Enhancers: sequence in the *exons* that *promote* splicing

**ESS** = Exonic Splicing Silencers: sequence in the *exons* that *inhibit* splicing

**ISE** = Intronic Splicing Enhancers: sequence in the *introns* that *promote* splicing.

**ISS** = Intronic Splicing Silencers: sequence in the *introns* that *inhibit* splicing
How were ESEs found?

(1) Training database was constructed of exonic mRNA (post-spliced) that was (a) constitutively spliced (not alternatively spliced), and (b) from an internal non-protein-coding exon.

(2) Database of ‘control’, non-ESE sequences was constructed.

(3) The relative abundance of all “8-mers” was found.

(4) 8-mers with high relative abundance were tested by mutating the putative ESE 8-mers and determining the splicing efficiency by gel electrophoresis.

Some of the motifs found by Zhang & Chasin using relative abundance analysis of 8-mers, after clustering.
Splicing facts

**Exons** average 145 nucleotides in length  
- Contain regulatory elements:  
  - ESEs: Exonic splicing enhancers  
  - ESSs: Exonic splicing silencers

**Introns** average more than 10x longer than exons  
- Contain regulatory elements (bind regulatory complexes)  
  - ISEs: Intronic splicing enhancers  
  - ISSs: Intronic splicing silencers

**Splice sites**

- **5' splice site**  
  - Sequence: AGguragu (r = purine)  
  - U1 snRNP: Binds to 5' splice site

- **3' splice site**  
  - Sequence: yyyyyyy nagG (y = pyrimidine)

- **Branch site**  
  - Sequence: ynyuray (r = purine)  
  - U2 snRNP: Binds to branch site via RNA:RNA interactions between snRNA and pre-mRNA
Alternative splicing facts

**Alternative splicing**
- Definition: Joining of different 5' and 3' splice sites
- ~80% of alternative splicing results in changes in the encoded protein
- Up to 59% of human genes express more than one mRNA by alternative splicing
- Functional effects: Generates several forms of mRNA from single gene
- Allows functionally diverse protein isoforms to be expressed according to different regulatory programs

**Structural effects:**
- Insert or remove amino acids
- Shift reading frame
- Introduce termination codon
- Gene expression effects
- Removes or inserts regulatory elements controlling translation, mRNA stability, or localization

**Regulation**
- Splicing pathways modulated according to:
  - Cell type
  - Developmental stage
  - Gender
  - External stimuli
1. What information is used to find genes in prokaryotes?
2. How were introns discovered?
3. What is the better way to find genes: by homology, or by statistical models?
4. Where is the ribosome binding site relative to the transcription start site?
5. Where is the Pribnow Box? Gilbert box?
6. How can HMMs be linked together to make a larger model?
7. What happens to the intron once it is removed?
8. What is alternative splicing?
9. What do DNA alignments tell you about introns/exons?
10. What is a splicing enhancer? silencer?