CS2220 Introduction to Computational Biology
Lecture 7: Gene Finding by Computational Analysis

Limsoon Wong
Outline

• Gene structure basics

• Gene finding overview

• GRAIL

• Indel & frame-shift in coding regions
Gene Structure Basics

A brief refresher

Some slides here are “borrowed” from Ken Sung
Gene

• A gene is a sequence of DNA that encodes a protein or an RNA molecule
• About 30,000 – 35,000 (protein-coding) genes in human genome
• For gene that encodes protein
  – In Prokaryotic genome, one gene corresponds to one protein
  – In Eukaryotic genome, one gene can corresponds to more than one protein because of the process “alternative splicing”
Introns and Exons

- Eukaryotic genes contain introns & exons
  - Introns are seq that are ultimately spliced out of mRNA
  - Introns normally satisfy GT-AG rule, viz. begin w/ GT & end w/ AG
  - Each gene can have many introns & each intron can have thousands bases

- Introns can be very long
- An extreme example is a gene associated with cystic fibrosis in human:
  - Length of 24 introns ~1Mb
  - Length of exons ~1kb
Typical Eukaryotic Gene Structure

- Unlike eukaryotic genes, a prokaryotic gene typically consists of only one contiguous coding region
Reading Frame

- Each DNA segment has six possible reading frames

<table>
<thead>
<tr>
<th>Forward strand:</th>
<th>Reading frame #1</th>
<th>Reading frame #2</th>
<th>Reading frame #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATG</td>
<td>TGG</td>
<td>GGC</td>
<td></td>
</tr>
<tr>
<td>GCT</td>
<td>CTT</td>
<td>TTA</td>
<td></td>
</tr>
<tr>
<td>TAC</td>
<td>ACG</td>
<td>CGC</td>
<td></td>
</tr>
<tr>
<td>GCT</td>
<td>CTT</td>
<td>TTG</td>
<td></td>
</tr>
<tr>
<td>TGC</td>
<td>GA.</td>
<td>A.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reverse strand:</th>
<th>Reading frame #4</th>
<th>Reading frame #5</th>
<th>Reading frame #6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCA</td>
<td>CAA</td>
<td>AAG</td>
<td></td>
</tr>
<tr>
<td>AGC</td>
<td>GCG</td>
<td>CGT</td>
<td></td>
</tr>
<tr>
<td>GTA</td>
<td>TAA</td>
<td>AAG</td>
<td></td>
</tr>
<tr>
<td>AGC</td>
<td>GCC</td>
<td>CCA</td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>AT.</td>
<td>T.</td>
<td></td>
</tr>
</tbody>
</table>
Open Reading Frame (ORF)

• ORF is a segment of DNA with a start codon and an in-frame stop codon at the two ends and no in-frame stop codon in the middle

• Each ORF has a fixed reading frame

NB: Other definitions are also used. Most important aspect is that there is no stop codon in the middle.
Coding Region

• Each coding region (exon or whole gene) has a fixed translation frame
• A coding region always sits inside an ORF of same reading frame
• All exons of a gene are on the same strand
• Neighboring exons of a gene could have different reading frames
Frame Consistency

- Neighboring exons of a gene should be frame-consistent

Exercise: Define frame consistency mathematically
Overview of Gene Finding

Some slides here are “borrowed” from Mark Craven
What is Gene Finding?

- Find all coding regions from a stretch of DNA sequence, and construct gene structures from the identified exons

- Can be decomposed into
  - Find coding potential of a region in a frame
  - Find boundaries between coding & non-coding regions

Image credit: Xu
Approaches

• Search-by-signal: find genes by identifying the sequence signals involved in gene expression

• Search-by-content: find genes by statistical properties that distinguish protein coding DNA from non-coding DNA

• Search-by-homology: find genes by homology (after translation) to proteins

• State-of-the-art systems for gene finding usually combine these strategies
Relevant Signals for Search-by-Signals

- **Transcription initiation**
  - Promoter
- **Transcription termination**
  - Terminators
- **Translation initiation**
  - Ribosome binding sites
  - Initiation codons
- **Translation termination**
  - Stop codons
- **RNA processing**
  - Splice junction

Image credit: Xu
How Search-by-Signal Works

• There are 2 impt regions in a promoter seq
  –10 region, ~10bp before TSS
  –35 region, ~35bp before TSS

• Consensus for –10 region in E. coli is TATAAT, but few promoters actually have this seq

• Recognize promoters by
  – weight matrices
  – probabilistic models
  – neural networks, …
How Search-by-Content Works

- Encoding a protein affects stats properties of a DNA seq
  - some amino acids used more frequently
  - diff number of codons for diff amino acids
  - for given protein, usually one codon is used more frequently than others

⇒ Estimate prob that a given region of seq was “caused by” its being a coding seq

<table>
<thead>
<tr>
<th>AA</th>
<th>codon</th>
<th>/1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly</td>
<td>GGG</td>
<td>1.89</td>
</tr>
<tr>
<td>Gly</td>
<td>GGA</td>
<td>0.44</td>
</tr>
<tr>
<td>Gly</td>
<td>GGU</td>
<td>52.99</td>
</tr>
<tr>
<td>Gly</td>
<td>GGC</td>
<td>34.55</td>
</tr>
<tr>
<td>Glu</td>
<td>GAG</td>
<td>15.68</td>
</tr>
<tr>
<td>Glu</td>
<td>GAA</td>
<td>57.20</td>
</tr>
<tr>
<td>Asp</td>
<td>GAU</td>
<td>21.63</td>
</tr>
<tr>
<td>Asp</td>
<td>GAC</td>
<td>43.26</td>
</tr>
</tbody>
</table>

Image credit: Craven
How Search-by-Homology Works

• Translate DNA seq in all reading frames

• Search against protein db

• High-scoring matches suggest presence of homologous genes in DNA

⇒ You can use BLASTX for this
Search-by-Content Example: Codon Usage Method

- Staden & McLachlan, 1982
- Process a seq w/ “window” of length L
- Assume seq falls into one of 7 categories, viz.
  - Coding in frame 0, frame 1, …, frame 5
  - Non-coding
- Use Bayes’ rule to determine prob of each category
- Assign seq to category w/ max prob
Codon Usage Method

\[ Pr(\text{coding}_i \mid S) = \frac{Pr(S \mid \text{coding}_i) \cdot Pr(\text{coding}_i)}{Pr(S)} \]

- probability that sequence encodes a protein in frame \(i\)

Codon Usage Method

- make simplifying assumption that the codons in a window are independent of one another

\[ Pr(S \mid \text{coding}_i) \approx \prod_{j=1}^{n} Pr(S_i(j) \mid \text{coding}_i) \]

- probability of the \(j\)th codon in frame \(i\) given the sequence is coding

Image credit: Craven

Copyright 2013 © Limsoon Wong
Codon Usage Method

\[ \Pr(\text{coding}_i \mid S) = \frac{\Pr(S \mid \text{coding}_i) \Pr(\text{coding}_i)}{\Pr(S)} \]

probability that sequence encodes a protein in frame i

Codon Usage Method

\[ \Pr(S) = \sum_i [\Pr(S \mid \text{coding}_i) \Pr(\text{coding}_i)] + \Pr(S \mid \text{noncoding}) \Pr(\text{noncoding}) \]

Sometimes this term is dropped since it’s difficult to estimate these statistics

Image credit: Craven
Codon Usage Method

\[ \Pr(\text{coding}_i \mid S) = \frac{\Pr(S \mid \text{coding}_i) \Pr(\text{coding}_i)}{\Pr(S)} \]

- \( \Pr(\text{coding}_i) \) is the same for each frame if window size fits same number of codons in each frame.
- Otherwise, consider relative number of codons in window in each frame.

Image credit: Craven

Copyright 2013 © Limsoon Wong
Codon Usage Method

- By sliding the window, we can generate predictions for the extent of our sequence

![](image)

Predicted Coding Regions

- frame 0
- frame 1
- frame 2
- frame 3
- frame 4
- frame 5

Image credit: Craven
Search-by-Homology Example: Gene Finding Using BLAST

- High seq similarity typically implies homologous genes

⇒ Search for genes in yeast seq using BLAST
⇒ Extract Feature for gene identification

Image credit: Xu
Search-by-Homology Example: Gene Finding Using BLAST

- High seq similarity typically implies homologous genes

⇒ Search for genes in yeast seq using BLAST
⇒ Extract Feature for gene identification

- Searching all ORFs against known genes in nr db helps identify an initial set of (possibly incomplete) genes

sequence

BLAST hits

Image credit: Xu

Copyright © 2004 by Limsoon Wong
A (yeast) gene starts w/ ATG and ends w/ a stop codon, in same reading frame of ORF

- Have “strong” coding potentials, measured by, preference models, Markov chain model, ...
- Have “strong” translation start signal, measured by weight matrix model, ...
- Have distributions wrt length, G+C composition, ...
- Have special seq signals in flanking regions, ...
GRAIL, 
An Important Gene Finding Program

Signals assoc w/ coding regions
Models for coding regions
Signals assoc w/ boundaries
Models for boundaries
Other factors & information fusion

Some slides here are “borrowed” from Ying Xu
Coding Signal

- Freq distribution of dimers in protein seq
- E.g., Shewanella
  - Ave freq is 5%
  - Some amino acids prefer to be next to each other
  - Some amino acids prefer to be not next to each other

<table>
<thead>
<tr>
<th>Name</th>
<th>ala</th>
<th>arg</th>
<th>asn</th>
<th>asp</th>
<th>cys</th>
<th>glu</th>
<th>gln</th>
<th>gly</th>
<th>his</th>
<th>ile</th>
<th>leu</th>
<th>lys</th>
<th>met</th>
<th>phe</th>
<th>pro</th>
<th>ser</th>
<th>thr</th>
<th>trp</th>
<th>tyr</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>ala</td>
<td>9.5</td>
<td>4.1</td>
<td>4.3</td>
<td>5.3</td>
<td>1.2</td>
<td>6</td>
<td>4.8</td>
<td>6.5</td>
<td>2</td>
<td>6.5</td>
<td>11.5</td>
<td>6</td>
<td>2.6</td>
<td>3.7</td>
<td>3.5</td>
<td>6.2</td>
<td>5</td>
<td>11</td>
<td>2.7</td>
<td>6.5</td>
</tr>
<tr>
<td>arg</td>
<td>7.9</td>
<td>5.5</td>
<td>3.9</td>
<td>5.3</td>
<td>1.1</td>
<td>6</td>
<td>5.5</td>
<td>5.9</td>
<td>2.6</td>
<td>6.5</td>
<td>11.4</td>
<td>5</td>
<td>2.2</td>
<td>4.7</td>
<td>3.6</td>
<td>5.5</td>
<td>4.4</td>
<td>1.4</td>
<td>4</td>
<td>6.6</td>
</tr>
<tr>
<td>asn</td>
<td>9.6</td>
<td>4.9</td>
<td>4.2</td>
<td>4.9</td>
<td>1</td>
<td>5.3</td>
<td>5.6</td>
<td>7.4</td>
<td>2.3</td>
<td>6</td>
<td>10</td>
<td>4.9</td>
<td>2</td>
<td>3.5</td>
<td>5.1</td>
<td>6.1</td>
<td>5.5</td>
<td>1.5</td>
<td>3.1</td>
<td>6.1</td>
</tr>
<tr>
<td>asp</td>
<td>9.3</td>
<td>4</td>
<td>4.7</td>
<td>5.1</td>
<td>1</td>
<td>6.7</td>
<td>2.9</td>
<td>7</td>
<td>1</td>
<td>8</td>
<td>7.1</td>
<td>9.6</td>
<td>6.3</td>
<td>2.3</td>
<td>4.3</td>
<td>3.9</td>
<td>5.9</td>
<td>5.1</td>
<td>1.6</td>
<td>3.6</td>
</tr>
<tr>
<td>cys</td>
<td>8.4</td>
<td>4.8</td>
<td>3.3</td>
<td>5.4</td>
<td>1.7</td>
<td>5.6</td>
<td>5.2</td>
<td>8.1</td>
<td>4.3</td>
<td>5.4</td>
<td>10.2</td>
<td>3.8</td>
<td>1.8</td>
<td>4.1</td>
<td>4.5</td>
<td>6.3</td>
<td>4.3</td>
<td>1.6</td>
<td>3.4</td>
<td>6.8</td>
</tr>
<tr>
<td>glu</td>
<td>9.4</td>
<td>5.8</td>
<td>3.6</td>
<td>4.5</td>
<td>0.8</td>
<td>4.9</td>
<td>7</td>
<td>5.8</td>
<td>2.6</td>
<td>5.9</td>
<td>12.7</td>
<td>5</td>
<td>2.4</td>
<td>4</td>
<td>3.5</td>
<td>5.4</td>
<td>5</td>
<td>1.1</td>
<td>2.8</td>
<td>6.8</td>
</tr>
<tr>
<td>gln</td>
<td>10.3</td>
<td>4.9</td>
<td>3</td>
<td>4.4</td>
<td>0.9</td>
<td>4.5</td>
<td>6.8</td>
<td>7</td>
<td>2.7</td>
<td>5.5</td>
<td>12.8</td>
<td>4.1</td>
<td>2</td>
<td>3.9</td>
<td>3.8</td>
<td>5.8</td>
<td>5.3</td>
<td>1.4</td>
<td>3.3</td>
<td>6.9</td>
</tr>
<tr>
<td>gly</td>
<td>8.1</td>
<td>4.8</td>
<td>3.9</td>
<td>5.1</td>
<td>1.2</td>
<td>6</td>
<td>4.6</td>
<td>6.4</td>
<td>2.4</td>
<td>6.8</td>
<td>10.5</td>
<td>5.8</td>
<td>2.7</td>
<td>4.8</td>
<td>2.4</td>
<td>5.8</td>
<td>5.1</td>
<td>1.4</td>
<td>3.7</td>
<td>7.5</td>
</tr>
<tr>
<td>his</td>
<td>7.3</td>
<td>4.7</td>
<td>4</td>
<td>4.8</td>
<td>1.5</td>
<td>4.9</td>
<td>5.6</td>
<td>6.9</td>
<td>3</td>
<td>6.2</td>
<td>10.8</td>
<td>4.8</td>
<td>1.6</td>
<td>5</td>
<td>5.2</td>
<td>6.8</td>
<td>4.9</td>
<td>1.7</td>
<td>4.2</td>
<td>5.1</td>
</tr>
<tr>
<td>ile</td>
<td>11</td>
<td>4.7</td>
<td>4.9</td>
<td>6.5</td>
<td>1.1</td>
<td>6.9</td>
<td>3.6</td>
<td>7.2</td>
<td>2.1</td>
<td>5.3</td>
<td>8.6</td>
<td>5.3</td>
<td>1.8</td>
<td>3.2</td>
<td>4.2</td>
<td>7</td>
<td>5.6</td>
<td>8.9</td>
<td>2.9</td>
<td>6.1</td>
</tr>
<tr>
<td>leu</td>
<td>10.4</td>
<td>4.2</td>
<td>4.3</td>
<td>5.2</td>
<td>1.1</td>
<td>5.2</td>
<td>3.7</td>
<td>6.8</td>
<td>2</td>
<td>5.6</td>
<td>10.6</td>
<td>5.3</td>
<td>2.3</td>
<td>3.8</td>
<td>4.5</td>
<td>7.4</td>
<td>6.2</td>
<td>1</td>
<td>2.6</td>
<td>6.6</td>
</tr>
<tr>
<td>lys</td>
<td>10.6</td>
<td>5.2</td>
<td>3.8</td>
<td>5.2</td>
<td>0.5</td>
<td>5.3</td>
<td>5.9</td>
<td>6.6</td>
<td>2.6</td>
<td>5.2</td>
<td>11.3</td>
<td>4.7</td>
<td>1.9</td>
<td>2.8</td>
<td>4.6</td>
<td>6</td>
<td>5.5</td>
<td>1.2</td>
<td>2.6</td>
<td>7.6</td>
</tr>
<tr>
<td>met</td>
<td>10.8</td>
<td>4.8</td>
<td>3.8</td>
<td>4.6</td>
<td>0.7</td>
<td>4.6</td>
<td>4.9</td>
<td>7</td>
<td>1</td>
<td>7.4</td>
<td>11.4</td>
<td>5</td>
<td>2.8</td>
<td>3.3</td>
<td>5.1</td>
<td>7.4</td>
<td>6.3</td>
<td>0.9</td>
<td>2</td>
<td>6.8</td>
</tr>
<tr>
<td>phe</td>
<td>9.6</td>
<td>3.7</td>
<td>5.2</td>
<td>6.5</td>
<td>1.2</td>
<td>6.4</td>
<td>2.7</td>
<td>7.9</td>
<td>6.7</td>
<td>7.4</td>
<td>5</td>
<td>2.5</td>
<td>3.9</td>
<td>3.6</td>
<td>8</td>
<td>5.8</td>
<td>1.3</td>
<td>3.3</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>pro</td>
<td>8.4</td>
<td>3.6</td>
<td>4.6</td>
<td>5.4</td>
<td>0.7</td>
<td>7.6</td>
<td>5.2</td>
<td>5.4</td>
<td>2.3</td>
<td>6.1</td>
<td>11.2</td>
<td>5.5</td>
<td>2.4</td>
<td>4.2</td>
<td>2.8</td>
<td>6.5</td>
<td>5.4</td>
<td>1.4</td>
<td>2.9</td>
<td>7.5</td>
</tr>
<tr>
<td>ser</td>
<td>9.1</td>
<td>4.6</td>
<td>3.7</td>
<td>5</td>
<td>1</td>
<td>5.4</td>
<td>5.2</td>
<td>7.2</td>
<td>2.6</td>
<td>6</td>
<td>11.6</td>
<td>4.5</td>
<td>2.2</td>
<td>4.1</td>
<td>4.1</td>
<td>6.5</td>
<td>5</td>
<td>1.2</td>
<td>3.2</td>
<td>6.8</td>
</tr>
<tr>
<td>thr</td>
<td>9.1</td>
<td>4.2</td>
<td>3.7</td>
<td>5.6</td>
<td>0.9</td>
<td>5.7</td>
<td>5.7</td>
<td>7.5</td>
<td>2.2</td>
<td>5.5</td>
<td>12</td>
<td>4.2</td>
<td>2</td>
<td>3.5</td>
<td>5.5</td>
<td>6.2</td>
<td>5.3</td>
<td>1.1</td>
<td>2.6</td>
<td>6.7</td>
</tr>
<tr>
<td>trp</td>
<td>7.1</td>
<td>6.3</td>
<td>3.2</td>
<td>4.8</td>
<td>1.3</td>
<td>3.9</td>
<td>8.5</td>
<td>6.6</td>
<td>3.5</td>
<td>6</td>
<td>14.2</td>
<td>3.2</td>
<td>2.4</td>
<td>4.6</td>
<td>3.9</td>
<td>5.8</td>
<td>4.3</td>
<td>1.3</td>
<td>3</td>
<td>6.1</td>
</tr>
<tr>
<td>tyr</td>
<td>7.9</td>
<td>6.5</td>
<td>3.6</td>
<td>4.9</td>
<td>1.2</td>
<td>4.5</td>
<td>7</td>
<td>7.1</td>
<td>2.6</td>
<td>5</td>
<td>11.7</td>
<td>4</td>
<td>1.6</td>
<td>4.7</td>
<td>4.9</td>
<td>6.4</td>
<td>4.6</td>
<td>1.5</td>
<td>3.4</td>
<td>5.7</td>
</tr>
<tr>
<td>val</td>
<td>9.6</td>
<td>4.1</td>
<td>4.4</td>
<td>5.9</td>
<td>1</td>
<td>6.2</td>
<td>3.4</td>
<td>6.4</td>
<td>1.8</td>
<td>6.5</td>
<td>10.2</td>
<td>5.2</td>
<td>2.5</td>
<td>3.7</td>
<td>3.8</td>
<td>7.2</td>
<td>6.1</td>
<td>1.1</td>
<td>2.7</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Image credit: Xu

Exercise: What is shewanella?
Coding Signal

• **Dimer preference implies dicodon (6-mers like AAA TTT) bias in coding vs non-coding regions**

• **Relative freq of a dicodon in coding vs non-coding**
  - Freq of dicodon X (e.g., AAA AAA) in coding region  
    \[ \text{Freq of dicodon X} = \frac{\text{total number of occurrences of X}}{\text{total number of dicodon occurrences}} \]
  - Freq of dicodon X (e.g., AAA AAA) in noncoding region  
    \[ \text{Freq of dicodon X} = \frac{\text{total number of occurrences of X}}{\text{total number of dicodon occurrences}} \]

Exercise: In human genome, freq of dicodon “AAA AAA” is ~1% in coding region vs ~5% in non-coding region. If you see a region with many “AAA AAA”, would you guess it is a coding or non-coding region?
Why Dicodon (6-mer)?

- Codon (3-mer)-based models are not as info rich as dicodon-based models
- Tricodon (9-mer)-based models need too many data points

⇒ For tricodon-based models, need at least $15 \times 262144 = 3932160$ coding bases in our training data, which is probably not going to be available for most genomes

There are
$4^3 = 64$ codons
$4^6 = 4096$ dicodons
$4^9 = 262144$ tricodons

To make stats reliable, need ~15 occurrences of each X-mer
Coding Signal

• Most dicodons show bias toward either coding or non-coding regions

⇒ Foundation for coding region identification

Regions consisting of dicodons that mostly tend to be in coding regions are probably coding regions; otherwise non-coding regions

⇒ Dicodon freq are key signal used for coding region detection; all gene finding programs use this info
Coding Signal

- Dicodon freq in coding vs non-coding are genome-dependent

<table>
<thead>
<tr>
<th>Name</th>
<th>ala</th>
<th>arg</th>
<th>asn</th>
<th>asp</th>
<th>cys</th>
<th>glu</th>
<th>gly</th>
<th>his</th>
<th>ile</th>
<th>leu</th>
<th>lys</th>
<th>met</th>
<th>phe</th>
<th>pro</th>
<th>ser</th>
<th>thr</th>
<th>trp</th>
<th>tyr</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>ala</td>
<td>9.5</td>
<td>4.1</td>
<td>4.3</td>
<td>5.3</td>
<td>1.2</td>
<td>6</td>
<td>4.8</td>
<td>6.5</td>
<td>2</td>
<td>6.5</td>
<td>11.6</td>
<td>6</td>
<td>2.6</td>
<td>3.7</td>
<td>3.5</td>
<td>6.2</td>
<td>5</td>
<td>1.1</td>
<td>2.7</td>
</tr>
<tr>
<td>arg</td>
<td>7.9</td>
<td>5.5</td>
<td>3.9</td>
<td>5.3</td>
<td>1.1</td>
<td>6</td>
<td>5.5</td>
<td>5.9</td>
<td>2.6</td>
<td>6.5</td>
<td>11.4</td>
<td>5</td>
<td>2.2</td>
<td>4.7</td>
<td>3.6</td>
<td>5.5</td>
<td>4.4</td>
<td>1.4</td>
<td>4.4</td>
</tr>
<tr>
<td>asn</td>
<td>9.6</td>
<td>4.9</td>
<td>4.2</td>
<td>4.9</td>
<td>1</td>
<td>3</td>
<td>5.6</td>
<td>7.4</td>
<td>2.3</td>
<td>6</td>
<td>10</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>3.5</td>
<td>5.1</td>
<td>6.1</td>
<td>5.5</td>
<td>1.5</td>
</tr>
<tr>
<td>asp</td>
<td>9.3</td>
<td>4</td>
<td>4.7</td>
<td>5.1</td>
<td>1</td>
<td>6.7</td>
<td>2.9</td>
<td>7</td>
<td>1.8</td>
<td>7.1</td>
<td>9.6</td>
<td>6</td>
<td>3</td>
<td>2.3</td>
<td>4.3</td>
<td>3.9</td>
<td>5.9</td>
<td>5.1</td>
<td>1.6</td>
</tr>
<tr>
<td>cys</td>
<td>8.4</td>
<td>4.8</td>
<td>3.3</td>
<td>5.1</td>
<td>1</td>
<td>5.6</td>
<td>5.2</td>
<td>8.1</td>
<td>4.3</td>
<td>5.4</td>
<td>10.2</td>
<td>3.8</td>
<td>1.8</td>
<td>4.1</td>
<td>4.5</td>
<td>6.3</td>
<td>4.3</td>
<td>1.6</td>
<td>3.4</td>
</tr>
<tr>
<td>glu</td>
<td>9.4</td>
<td>5.8</td>
<td>3.6</td>
<td>4.5</td>
<td>0.8</td>
<td>4.9</td>
<td>7</td>
<td>5.8</td>
<td>2.6</td>
<td>5.9</td>
<td>12.7</td>
<td>5</td>
<td>2.4</td>
<td>3</td>
<td>3.5</td>
<td>5.4</td>
<td>5</td>
<td>1.1</td>
<td>2.8</td>
</tr>
<tr>
<td>gly</td>
<td>10.3</td>
<td>4.9</td>
<td>5</td>
<td>4.4</td>
<td>0.9</td>
<td>4.5</td>
<td>6.8</td>
<td>7</td>
<td>2.7</td>
<td>5.5</td>
<td>12.8</td>
<td>4.1</td>
<td>2</td>
<td>3.9</td>
<td>3.8</td>
<td>5.8</td>
<td>5.3</td>
<td>1.4</td>
<td>3</td>
</tr>
<tr>
<td>his</td>
<td>7.3</td>
<td>4.7</td>
<td>4</td>
<td>4.8</td>
<td>1.5</td>
<td>4.9</td>
<td>5.6</td>
<td>6.9</td>
<td>3</td>
<td>6.2</td>
<td>10.8</td>
<td>4.8</td>
<td>1.6</td>
<td>5</td>
<td>5.2</td>
<td>6.8</td>
<td>4.9</td>
<td>1</td>
<td>4.2</td>
</tr>
<tr>
<td>ile</td>
<td>11</td>
<td>4.7</td>
<td>4.9</td>
<td>6.5</td>
<td>1</td>
<td>6.9</td>
<td>3.6</td>
<td>7.2</td>
<td>2.1</td>
<td>5.3</td>
<td>8.6</td>
<td>5.3</td>
<td>18</td>
<td>3</td>
<td>3.2</td>
<td>4.2</td>
<td>7</td>
<td>5.6</td>
<td>0.9</td>
</tr>
<tr>
<td>leu</td>
<td>10.4</td>
<td>4.2</td>
<td>4.3</td>
<td>5.2</td>
<td>1.1</td>
<td>5.2</td>
<td>3.7</td>
<td>6.8</td>
<td>2</td>
<td>5.6</td>
<td>10.6</td>
<td>5.3</td>
<td>23</td>
<td>3</td>
<td>3.8</td>
<td>4.5</td>
<td>7.4</td>
<td>6.2</td>
<td>1</td>
</tr>
<tr>
<td>lys</td>
<td>10.6</td>
<td>5.2</td>
<td>3.8</td>
<td>5.2</td>
<td>0.5</td>
<td>5.3</td>
<td>5.9</td>
<td>6.6</td>
<td>2.6</td>
<td>5.2</td>
<td>11.3</td>
<td>4.7</td>
<td>19</td>
<td>2</td>
<td>8.4</td>
<td>2.6</td>
<td>6.5</td>
<td>5.5</td>
<td>1.2</td>
</tr>
<tr>
<td>met</td>
<td>10.8</td>
<td>4.8</td>
<td>3.8</td>
<td>4.6</td>
<td>0.7</td>
<td>4.6</td>
<td>4.9</td>
<td>7</td>
<td>1.7</td>
<td>4.7</td>
<td>11.4</td>
<td>5.2</td>
<td>28</td>
<td>3</td>
<td>3.3</td>
<td>5.1</td>
<td>7.4</td>
<td>6.3</td>
<td>0.9</td>
</tr>
<tr>
<td>phe</td>
<td>9.6</td>
<td>5.7</td>
<td>3.2</td>
<td>6.5</td>
<td>1.2</td>
<td>6.4</td>
<td>2.7</td>
<td>7.9</td>
<td>1.9</td>
<td>6.7</td>
<td>7.4</td>
<td>5</td>
<td>25</td>
<td>3</td>
<td>3.9</td>
<td>3.6</td>
<td>8</td>
<td>5.8</td>
<td>1.3</td>
</tr>
<tr>
<td>pro</td>
<td>8.4</td>
<td>5.6</td>
<td>3.5</td>
<td>4.8</td>
<td>0.7</td>
<td>7.5</td>
<td>2.8</td>
<td>6.3</td>
<td>2.1</td>
<td>6.6</td>
<td>11.4</td>
<td>5</td>
<td>22</td>
<td>4</td>
<td>2.1</td>
<td>4.2</td>
<td>8.5</td>
<td>5.4</td>
<td>1.4</td>
</tr>
<tr>
<td>ser</td>
<td>9.1</td>
<td>4.6</td>
<td>3.7</td>
<td>5</td>
<td>1</td>
<td>5.4</td>
<td>5.2</td>
<td>7.2</td>
<td>2.6</td>
<td>6</td>
<td>11.6</td>
<td>4</td>
<td>22</td>
<td>4</td>
<td>2.1</td>
<td>4.2</td>
<td>8.5</td>
<td>5.4</td>
<td>1.4</td>
</tr>
<tr>
<td>thr</td>
<td>9.1</td>
<td>4.2</td>
<td>3.7</td>
<td>5.6</td>
<td>0.9</td>
<td>5.7</td>
<td>5.7</td>
<td>7.5</td>
<td>2.2</td>
<td>5.5</td>
<td>12</td>
<td>4</td>
<td>2</td>
<td>3.5</td>
<td>5.5</td>
<td>6.2</td>
<td>5.3</td>
<td>1.1</td>
<td>2.6</td>
</tr>
<tr>
<td>trp</td>
<td>7.1</td>
<td>6.3</td>
<td>3.2</td>
<td>4.8</td>
<td>1.3</td>
<td>3.9</td>
<td>8.5</td>
<td>6.6</td>
<td>3.6</td>
<td>5</td>
<td>14.2</td>
<td>8.2</td>
<td>2.4</td>
<td>4.6</td>
<td>3.9</td>
<td>5.8</td>
<td>4.3</td>
<td>1.3</td>
<td>6</td>
</tr>
<tr>
<td>tyr</td>
<td>7.9</td>
<td>6.5</td>
<td>3.6</td>
<td>4.9</td>
<td>1.2</td>
<td>4.5</td>
<td>7</td>
<td>7.1</td>
<td>2.6</td>
<td>5</td>
<td>11.7</td>
<td>4</td>
<td>1.6</td>
<td>4.7</td>
<td>4.9</td>
<td>5.4</td>
<td>4.6</td>
<td>1.5</td>
<td>3.4</td>
</tr>
<tr>
<td>val</td>
<td>9.6</td>
<td>4.1</td>
<td>4.4</td>
<td>5.9</td>
<td>1</td>
<td>6.2</td>
<td>3.4</td>
<td>6.4</td>
<td>1.8</td>
<td>6.5</td>
<td>10.2</td>
<td>5.2</td>
<td>25</td>
<td>3</td>
<td>3.7</td>
<td>3.8</td>
<td>7</td>
<td>6.1</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Shewanella

Dicodon freq in coding vs non-coding are genome-dependent

Bovine

Image credit: Xu
Coding Signal

- In-frame vs any-frame dicodons

- In-frame dicodon freq provide a more sensitive measure than any-frame dicodon freq

**Diagram:**
- **In-frame dicodons:** ATG TTG GAT GCC CAG AAG....
- **Not in-frame dicodons:** TGTTGG, ATGCC, AGAAG, GTTGGA, AGCCCA, AGAAG...

**Legend:**
- **in-frame dicodons**
- **not in-frame dicodons**
- **any-frame**
Dicodon Preference Model

- The preference value $P(X)$ of a dicodon $X$ is defined as

$$P(X) = \log \frac{FC(X)}{FN(X)}$$

where

- $FC(X)$ is freq of $X$ in coding regions
- $FN(X)$ is freq of $X$ in non-coding regions
Dicodon Preference Model’s Properties

• $P(X) = 0$ if $X$ has same freq in coding and non-coding regions

• $P(X) > 0$ if $X$ has higher freq in coding than in non-coding region; the larger the diff, the more positive the score is

• $P(X) < 0$ if $X$ has higher freq in non-coding than in coding region; the larger the diff, the more negative the score is
Dicodon Preference Model Example

• Suppose AAA ATT, AAA GAC, AAA TAG have the following freq:

  FC(AAA ATT) = 1.4%
  FN(AAA ATT) = 5.2%

  FC(AAA GAC) = 1.9%
  FN(AAA GAC) = 4.8%

  FC(AAA TAG) = 0.0%
  FN(AAA TAG) = 6.3%

• Then
  P(AAA ATT) = −0.57
  P(AAA GAC) = −0.40
  P(AAA TAG) = −∞,
  treating STOP codons differently

⇒ A region consisting of only these dicodons is probably a non-coding region
Frame-Insensitive Coding Region Preference Model

• A frame-insensitive coding preference $S_{is}(R)$ of a region $R$ can be defined as

$$S_{is}(R) = \sum_{X \text{ is a dicodon in } R} P(X)$$

• $R$ is predicted as coding region if $S_{is}(R) > 0$

NB. This model is not commonly used
In-Frame Dicodon Preference Model

- The in-frame + i preference value $P_i(X)$ of a dicodon $X$ is defined as

$$P_i(X) = \log \frac{F_{C_i}(X)}{F_{N}(X)}$$

where

$F_{C_i}(X)$ is freq of $X$ in coding regions at in-frame + i positions
$F_N(X)$ is freq of $X$ in non-coding regions

ATG TGC CGC GCT

$P_0$ $P_1$ $P_2$
In-Frame Coding Region Preference Model

• The in-frame + i preference $S_i(R)$ of a region $R$ can be defined as

$$S_i(R) = \sum_{X \text{ is a dicodon at in-frame } + i \text{ position in } R} P_i(X)$$

• $R$ is predicted as coding if $\sum_{i=0,1,2} S_i(R)/|R| > 0$

NB. This coding preference model is commonly used
Coding Region Prediction: An Example Procedure

- Calculate all ORFs of a DNA segment
- For each ORF
  - Slide thru ORF w/ increment of 10bp
  - Calculate in-frame coding region preference score, in same frame as ORF, within window of 60bp
  - Assign score to center of window
- E.g., forward strand in a particular frame...

Image credit: Xu
Problem with Coding Region Boundaries

• Making the call: coding or non-coding and where the boundaries are

⇒ Need training set with known coding and non-coding regions to select threshold that includes as many known coding regions as possible, and at the same time excludes as many known non-coding regions as possible
Types of Coding Region Boundaries

- Knowing boundaries of coding regions helps identify them more accurately
- Possible boundaries of an exon

- Splice junctions:
  - Donor site: coding region | GT
  - Acceptor site: CAG | TAG | coding region

- Translation start
  - in-frame ATG

What do you expect at translation stop?
Signals for Coding Region Boundaries

- Splice junction sites and translation starts have certain distribution profiles
- For example, ...
Acceptor Site (Human Genome)

- If we align all known acceptor sites (with their splice junction site aligned), we have the following nucleotide distribution

<table>
<thead>
<tr>
<th></th>
<th>-14</th>
<th>-13</th>
<th>-12</th>
<th>-11</th>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11.1</td>
<td>12.7</td>
<td>3.2</td>
<td>4.8</td>
<td>12.7</td>
<td>8.7</td>
<td>16.7</td>
<td>16.7</td>
<td>12.7</td>
<td>9.5</td>
<td>26.2</td>
<td>6.3</td>
<td>100</td>
<td>0.0</td>
<td>21.4</td>
</tr>
<tr>
<td>C</td>
<td>36.5</td>
<td>30.9</td>
<td>19.1</td>
<td>23.0</td>
<td>34.9</td>
<td>39.7</td>
<td>34.9</td>
<td>40.5</td>
<td>40.5</td>
<td>36.5</td>
<td>33.3</td>
<td>68.2</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
<tr>
<td>G</td>
<td>9.5</td>
<td>10.3</td>
<td>15.1</td>
<td>12.7</td>
<td>8.7</td>
<td>9.5</td>
<td>16.7</td>
<td>4.8</td>
<td>2.4</td>
<td>6.3</td>
<td>13.5</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
<td>62.7</td>
</tr>
<tr>
<td>U</td>
<td>38.9</td>
<td>41.3</td>
<td>58.7</td>
<td>55.6</td>
<td>42.1</td>
<td>40.5</td>
<td>30.9</td>
<td>37.3</td>
<td>44.4</td>
<td>47.6</td>
<td>27.0</td>
<td>25.4</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
</tbody>
</table>

- Acceptor site: CAG | TAG | coding region

Image credit: Xu
Donor Site (Human Genome)

- If we align all known donor sites (with their splice junction site aligned), we have the following nucleotide distribution

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>C</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>G</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>U</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

- Donor site: coding region | GT

Image credit: Xu
What Positions Have “High” Info Content?

• For a weight matrix, information content of each column is calculated as

\[ -\sum_{X \in \{A,C,G,T\}} F(X) \times \log \left( \frac{F(X)}{0.25} \right) \]

• When a column has evenly distributed nucleotides, its information content is lowest

• Only need to look at positions having high information content
Information Content Around Donor Sites in Human Genome

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>C</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>G</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>U</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

- **Information content**
  - column \(-3\) = \(-0.34 \times \log (0.34/0.25) - 0.363 \times \log (0.363/0.25) - 0.183 \times \log (0.183/0.25) - 0.114 \times \log (0.114/0.25) = 0.04\)
  - column \(-1\) = \(-0.092 \times \log (0.92/0.25) - 0.03 \times \log (0.033/0.25) - 0.803 \times \log (0.803/0.25) - 0.073 \times \log (0.73/0.25) = 0.30\)

Image credit: Xu
Weight Matrix Model for Splice Sites

- Weight matrix model
  - Build a weight matrix for donor, acceptor, translation start site, respectively
  - Use positions of high information content

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>C</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>G</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>U</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

Nucleotide distribution around human donor sites

Image credit: Xu
Just to make sure you know what I mean ...

- Give me 3 DNA seq of length 10:
  - Seq\(_1\) = ACCGAGTTCT
  - Seq\(_2\) = AGTGTACCTG
  - Seq\(_3\) = AGTTCGTATG
- Then the weight matrix is ...

<table>
<thead>
<tr>
<th>1-mer</th>
<th>pos1</th>
<th>pos2</th>
<th>pos3</th>
<th>pos4</th>
<th>pos5</th>
<th>pos6</th>
<th>pos7</th>
<th>pos8</th>
<th>pos9</th>
<th>pos10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3/3</td>
<td>0/3</td>
<td>0/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0/3</td>
<td>2/3</td>
<td>0/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0/3</td>
<td>0/3</td>
<td>2/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exercise: Fill in the rest of the table
Splice Site Prediction: A Procedure

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>C</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>G</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>U</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

Nucleotide distribution around human donor sites

- **Add up freq of corr letter in corr positions:**

  - **AAGGTAAGT:** $0.34 + 0.60 + 0.80 + 1.0 + 1.0 + 0.52 + 0.71 + 0.81 + 0.46 = 6.24$
  - **TGTGTCTCA:** $0.11 + 0.12 + 0.03 + 1.0 + 1.0 + 0.02 + 0.07 + 0.05 + 0.16 = 2.56$

- **Make prediction on splice site based on some threshold**

Image credit: Xu
Other Factors Considered by GRAIL

- G+C composition affects dicodon distributions
- Length of exons follows certain distribution
- Other signals associated with coding regions
  - periodicity
  - structure information
  - ......
- Pseudo genes
- .........
Info Fusion by ANN in GRAIL

Exon Candidate Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>DNA</th>
<th>Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td># Exons</td>
<td>TP</td>
<td>%</td>
</tr>
<tr>
<td>Short</td>
<td>229</td>
<td>171</td>
</tr>
<tr>
<td>Long</td>
<td>600</td>
<td>575</td>
</tr>
<tr>
<td>Total</td>
<td>829</td>
<td>746</td>
</tr>
<tr>
<td># Bases</td>
<td>134814</td>
<td>122885</td>
</tr>
</tbody>
</table>

Image credit: Xu
Remaining Challenges in GRAIL

- Initial exon
- Final exon
- Indels & frame shifts
Indel & Frame-Shift in Coding Regions

Problem definition
Indel & frameshift identification
Indel correction
An iterative strategy

Some slides here are “borrowed” from Ying Xu
Indels in Coding Regions

- Indel = insertion or deletion in coding region
- Indels are usually caused by seq errors

ATG GAT CCA CAT ..... → ATG GAT CA CAT .....  
ATG GAT CTCA CAT .....
Effects of Indels on Exon Prediction

• Indels may cause shifts in reading frames & affect prediction algos for coding regions

Image credit: Xu
Key Idea for Detecting Frame-Shift

• Preferred reading frame is reading frame w/ highest coding score
• Diff DNA segments may have diff preferred reading frames

⇒ Segment a coding sequence into regions w/ consistent preferred reading frames corr well w/ indel positions
⇒ Indel identification problem can be solved as a sequence segmentation problem!
Frame-Shift Detection by Seq Segmentation

- Partition seq into segs so that
  - Chosen frames of adjacent segs are diff
  - Each segment has >30 bps to avoid small fluctuations
  - Sum of coding scores in the chosen frames over all segments is maximized
Frame-Shift Detection: A Simplified Treatment

- Given DNA sequence \( a_1 \ldots a_n \)
- Define key quantities

\[
C(i, r) = \max \text{ score on } a_1 \ldots a_i, \quad w/ \text{ the last segment in frame } r
\]

- Then

\[
\max_{r \in \{0, 1, 2\}} C(n, r) \text{ is optimal solution}
\]
Frame-Shift Detection: $C(i, r)$

- To calculate $C(i, r)$, there are 3 possible cases for each position $i$:
  - Case 1: no indel occurred at position $i$
  - Case 2: $a_i$ is an inserted base
  - Case 3: a base has been deleted in front of $a_i$

$\Rightarrow C(i, r) = \max \{ \text{Case 1, Case 2, Case 3} \}$
Frame-Shift Detection: Case 1

• No indel occurs at position $i$. Then

$$C(i,r) = C(i-1, r') + P_r(a_{i-5}...a_i)$$

<table>
<thead>
<tr>
<th>$r''$</th>
<th>$r'$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Frame-Shift Detection: Case 2

- $a_{i-1}$ is an inserted base. Then

$$C(i,r) = C(i-2, r') + P_r(a_{i-6}...a_{i-2}a_i)$$

![Diagram](image-url)
Frame-Shift Detection: Case 3

- A base has been deleted in front of $a_i$. Then

\[
C(i, r) = C(i-1, r'') + P_r'(a_{i-5} \ldots a_{i-1} C) + P_r(a_{i-4} \ldots a_{i-1} Ca_i)
\]

Exercise: why is “C” is best choice for the purpose above?
Frame-Shift Detection: Initiation

• Initial conditions,

\[ C(k, r) = -\alpha, \ k < 6 \]
\[ C(6, r) = P_r(a_1 \ldots a_6) \]

• This is a dynamic programming (DP) algorithm; the equations are DP recurrences

Exercise: How to modified the recurrence so that each fragment is at least 30bp?
Frame-Shift Detection: Determining Indel Positions

- Calculation of $\max_{r \in \{0, 1, 2\}} C(i, r)$ gives an optimal segmentation of a DNA sequence
- Tracing back the transition points---viz. case 2 & case 3---gives the segmentation results

Image credit: Xu
Frame-Shift Detection: Determine Coding Regions

- For given $H_1$ and $H_2$ (e.g., $= 0.25$ for noncoding and $0.75$ for coding), partition a DNA seq into segs so that each seg has >30 bases & coding values of each seg are consistently closer to one of $H_1$ or $H_2$ than the other.

Image credit: Xu
Frame-Shift Detection: Finally…

- Overlay “preferred reading-frame segs” & “coding segs” gives coding region predictions regions w/ indels

Image credit: Xu
What Happens When Indels Are Close Together?

- Our procedure works well when indels are not too close together (i.e., >30 bases apart)
- When indels are too close together, they will be missed...
Handling Indels That Are Close Together

- **Employ an iterative process, viz**
  - Find one set of indels
  - Correct them
  - Iterate until no more indels can be found

- **actual indels**

- **predicted indels**

- **predicted indels in iteration 2**
About the Inventor: Ying Xu

- Regents-GRA Eminent Scholar Chair Professor, Dept. of Biochem & Mol Biol, Univ of Georgia, Athens
- Director, Inst of Bioinformatics, Univ of Georgia, Athens

Ying Xu’s family with Huiqing Liu’s family.
Image credit: Huiqing Liu
Any Question?
Acknowledgements

• I “borrowed” a lot of materials in this lecture from Xu Ying (Univ of Georgia) and Mark Craven (Univ of Wisconsin)
References

• D. J. States, W. Gish, “Combined use of sequence similarity and codon bias for coding region identification”, JCB, 1:39--50, 1994
References

• V. Solovyev et al. "Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames", NAR, 22:5156--5163, 1994