Today’s Lecture

- Smith-Waterman special cases
- Word nucleation algorithms
  - BLAST
- Site models
The *Edit Graph* for a Pair of Sequences

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
• Find *imperfect internal repeats* by searching edit graph of sequence against itself
  – i.e. the same sequence labels columns and rows *above (\& not including) the main diagonal*:
    – if include main diagonal, best path will be identity match to self
    – complexity = $O(N^2)$ where $N =$ sequence length.

Graph for finding imperfect internal repeats:
• Find *short tandem repeats* (e.g. microsatellites, minisatellites):
  – scan a *band* just above main diagonal.
  – Complexity = $O(kN)$ where $k$ is width of the band.
  – Manageable even for large $N$, if $k$ small.

Graph for finding short tandem repeats:

```
ACACACACACACACAC
ACACACACACACACAC
```
• Other alignment tasks:
  – EST, or cDNA, to genomic sequence (exons)
  – protein to genomic.
• Can solve by variants of Smith-Waterman:
  – e.g. cDNA vs genomic:
    • set moderately large negative penalty for mismatch and for gap opening,
    • 0 for gap extension.
    • issue of proper placement of splice sites ...
Word Nucleation Algorithms

- Idea: find short (perfect or imperfect) word matches to ‘nucleate’ graph search
  - Each such match defines short diagonal path
  - Only search part of graph ‘surrounding’ this path
- BLAST: allow imperfect short (e.g. length 3) matches.
  - “Neighbors”: set of 3-residue sequences having ≥ min score T against some 3-residue sequence of query
  - Scan database seqs until hit word in neighbor list
  - then do ungapped extension (along diagonal defined by word match)
    - ‘significant’ matches are those with scores ≥ a threshold S
    - Ungapped matches are effective for detecting related proteins:
      - true protein alignments usually include substantial gap-free regions.
BLAST: Word Nucleating Alignment
– If find $\geq 2$ significant ungapped matches in same seq, expand search to connecting region of matrix, allowing gaps:
Other Word Nucleation Programs

• FASTA:
  – look for clusters of short exact matches, on nearby diagonals;
  – when found, extend to gapped alignment

• cross_match:
  – do full search of bands around exact matches

• These all still time complexity $O(MN)$
  – because # word matches proportional to $MN$ but with much smaller constant.
• In database searches, most seqs unrelated to query
• suggests following strategy:
  – Initial rapid pass through database using fast algorithm
    • e.g. just looking for gap-free matches
to get (approximate) score,
  – identify sequences having scores above a threshold
  – use full Smith-Waterman on latter
  – for appropriate (low) threshold can get sensitivity nearly as good as full Smith-Waterman search.
• Important issue: statistical significance for database searches! We will return to this later (Karlin-Altschul theory).
Site Models

- Probability models for short sequences, such as:
  - splice sites
  - translation start sites
  - promoter elements
  - protein “motifs”
(Protein-coding) Gene Structure in Eukaryotes

Transcription start site
Upstream regulatory region

Gene

Exon
Intron

5’ splice site
3’ splice site
Polyadenylation site

mRNA (spliced)

5’ untranslated region

Coding sequence (ORF) – begins with start codon (AUG), ends with stop codon (UAA, UAG, or UGA)

PolyA tail

3’ untranslated region
• Assumptions:
  – different examples of site can be aligned *without gaps* (indels) such that tend to have same residues in same positions
  – drop equal freq assumption: allow *position-specific freqs*
  – retain *independence* assumption (for now)
• Applies to short segments (< 30 residues) where
  – precise residue spacing is structurally or functionally important, and
  – certain positions are highly conserved

• Examples:
  – DNA/RNA sequences binding a single protein or RNA molecule
  – Protein internal regions structurally constrained due to folding requirements; or
  – protein surface regions constrained because bind certain ligands
Construction of Site Models

- Collect examples of site
- Align (without gaps)
- Count occurrences of residues at each position
- Convert to frequencies
### Nucleotide Counts for 8192 C. elegans 3’ Splice Sites

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Exon Counts</th>
<th>Intron Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3276 3516 2313 476 67 757 240 8192 0 3359 2401 2514</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>970 648 664 236 129 1109 6830 0 0 1277 1533 1847</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>593 575 516 144 39 595 12 0 8192 2539 1301 1567</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>3353 3453 4699 7336 7957 5731 1110 0 0 1017 2957 2264</td>
<td></td>
</tr>
</tbody>
</table>

**CONSENSUS**

<table>
<thead>
<tr>
<th>W</th>
<th>W</th>
<th>W</th>
<th>T</th>
<th>T</th>
<th>t</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>r</th>
<th>w</th>
<th>w</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Exon Counts</th>
<th>Intron Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.400 0.429 0.282 0.058 0.008 0.092 0.029 1.000 0.000 0.410 0.293 0.307</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.118 0.079 0.081 0.029 0.016 0.135 0.834 0.000 0.000 0.156 0.187 0.225</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.072 0.070 0.063 0.018 0.005 0.073 0.001 0.000 1.000 0.310 0.159 0.191</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.409 0.422 0.574 0.896 0.971 0.700 0.135 0.000 0.000 0.124 0.361 0.276</td>
<td></td>
</tr>
</tbody>
</table>
3’ Splice Sites – *C. elegans*
### Nucleotide Counts for 8192 C. elegans 5’ Splice Sites

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>A (3404)</th>
<th>4644</th>
<th>1518</th>
<th>0</th>
<th>0</th>
<th>4836</th>
<th>5486</th>
<th>837</th>
<th>1632</th>
<th>2189</th>
<th>2278</th>
<th>2355</th>
</tr>
</thead>
<tbody>
<tr>
<td>C (1850)</td>
<td>1224</td>
<td>583</td>
<td>0</td>
<td>14</td>
<td>118</td>
<td>588</td>
<td>237</td>
<td>801</td>
<td>771</td>
<td>889</td>
<td>986</td>
<td></td>
</tr>
<tr>
<td>G (1562)</td>
<td>912</td>
<td>4891</td>
<td>8192</td>
<td>0</td>
<td>1890</td>
<td>672</td>
<td>6164</td>
<td>589</td>
<td>962</td>
<td>1056</td>
<td>827</td>
<td></td>
</tr>
<tr>
<td>T (1376)</td>
<td>1412</td>
<td>1200</td>
<td>0</td>
<td>8178</td>
<td>1348</td>
<td>1446</td>
<td>954</td>
<td>5170</td>
<td>4270</td>
<td>3969</td>
<td>4024</td>
<td></td>
</tr>
</tbody>
</table>

#### CONSENSUS

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>a</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.416</td>
<td>0.567</td>
<td>0.185</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>C</td>
<td>0.226</td>
<td>0.149</td>
<td>0.071</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>G</td>
<td>0.191</td>
<td>0.111</td>
<td>0.597</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T</td>
<td>0.168</td>
<td>0.172</td>
<td>0.146</td>
<td>0.000</td>
<td>0.998</td>
</tr>
</tbody>
</table>

The 5’ splice site (5’ ss) is marked by the decrease in nucleotide counts from exons to introns.
5’ Splice Sites – *C. elegans*

![Graph showing splice sites for C. elegans with nucleotide frequencies for A, C, G, and T.]
Conserved Domain in RecR and Class I Topoisomerasases

RecR  RLAEEKITEVILATNPTVEGEATANYIAELC
RecM  RLQDDQVTEVILATNPNIEGEATAMYISRLLL
RecR  RVDDVGITEVIIATDPNTEGEATATYLVRMV
TrsI  IFKENKIDEVIATDPAREGENIAAYKILNQL
TOP1  KQLAEKADHIYLATDLDREGEAIAWRLREVII
ORF1  AELLLKQANTIIIVATDSDREGENIAWSIIHKA
TOP1  KDALKDADELILATDEDREGKVISWHLLQLL
TOP1  TIFDKRVKTIILATDAAELEGYIGRNILYRL
TOP3  KREARNADYLMIWTDCDREGEYIGWEIWQEA
TOP3  KRFLHEASEIVHAGDPDREGQQLVDEVLDYL
RGYR  RNLAVEADEVLIGTDPDTEGEKIAWDLYLAL

CONSENSUS  xxxxxxxxxxxxU&uatDxxxxEGexxxxxxUxxxxu

Consensus key:

Uppercase: all residues chemically similar
lowercase: most are
U,u: bulky aliphatic (I,L,V)
&: bulky hydrophobic (I,L,V,M,F,Y,W)

From RL Tatusov, SF Altschul, and EV Koonin, PNAS 91: 12091-12095
Probability Models for Sites (assuming independence!)

- For each position $i$, $1 \leq i \leq n$, let $P_i$ be a prob dist’n on the alphabet of residues
  - e.g. constructed using counts at that position in a sample of sites.
  - $P_i(r)$ for each residue $r$ is the probability that $r$ occurs at position $i$ in a sequence.
- Prob dist’n $P$ on the space $S$ of sequences of length $n$ is defined by
  $$P(s) = \prod_{1 \leq i \leq n} P_i(s_i)$$
  where $s = s_1 s_2 \ldots s_n$
Zero Probabilities

- If \( P_i(r) = 0 \) for some \( i \) and \( r \), then \( P(s) = 0 \) for some sequences.
  - may or may not be desirable

- If due to failure to observe residue because of small sample size,
  - should perform “small-sample correction” to change \( P_i(r) \) to a small non-zero value.
  - usually done by adding ‘pseudocounts’ to each value in the counts matrix;
    - e.g. add 1 to each cell (has justification in Bayesian statistics)
  - Particularly an issue with proteins, due to larger alphabet size.

- If reflects real biological constraints
  - then leave as 0.
  - e.g. requirement for G at position +1 (first intronic base) in 5’ss