L3: Blast: Alignment Scores etc.
Silly Quiz

- Name a famous Bioinformatics Researcher.

- Name a famous Bioinformatics Researcher who is a woman.
Alignment Cont'd: Linear Space

- Trick discovered by Hirschberg in the 70's.
- Note that the score computation only requires linear space.

For $i = 1$ to $n$

For $j = 1$ to $m$

\[
S[i_1, j_1] = \max \begin{cases} 
S[i, j] &+ C(s_i, t_j) \\
S[i, j - 1] &+ C(s_i, \text{\textdbl}t_j) \\
S[i - 1, j] &+ C(\text{\textdbl}s_i, \text{\textdbl}t_j) \\
S[i - 1, j - 1] &+ C(s_i, \text{\textdbl}t_j) 
\end{cases}
\]

\[
i_1 = i \% 2, \quad i_2 = (i - 1) \% 2
\]

Computing the Alignment itself is harder.
Linear Space Alignment

- We need to compute an opt. path through the DP table.
- In linear space, we can compute the coordinate of the path for any single row.
When you are at position $i=n/2$, you know all of the optimal paths ending at that row.

- $F[j] = S[n/2,j]$
- One of them is the truly optimal path.
Recompute alignments from the bottom up.

- $B[j] = S_b[n/2,j]$
- Clearly the optimum path passes through the point $j^*$ that maximizes $F[j] + B[j]$
- Recurse on $([1,n/2],[1,j^*])$
Linear Space Alignment

- Clearly the optimum path passes through the point $j^*$ that maximizes $F[j] + B[j]$
- Recurse on $([1,n/2],[1,j^*])$, and $([n/2,n],[j^*,m])$
- Space is linear
- $T(nm) = T(nm/2) + O(nm)$
Blast Alignment & HW1

- Blast uses a greedy (sub-optimal) alignment strategy to align.
- In your assignment, you are required to do a linear space local alignment.
The nucleotide bases have structure!
For DNA, we worked with a simple match/mismatch criteria.

- The bases can be grouped into Purines (AG) & Pyrimidines (CT)
- Transitions (nucleotide substitution within a group) are more likely than transversions.
DNA scoring

- Q: What is cDNA?
  - DNA that encodes a protein
  - When aligning cDNA, the third base substitution is more likely, then other positions.
  - $UU[UCAG] = \text{Pro}$
- Q: Can you devise an algorithm to score appropriately?

### The Genetic Code

<table>
<thead>
<tr>
<th>1st base in codon</th>
<th>2nd base in codon</th>
<th>3rd base in codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Ser</td>
<td>Ser</td>
</tr>
<tr>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
</tr>
<tr>
<td>Leu</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>Leu</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>Pro</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Ile</td>
<td>Ile</td>
<td>Asn</td>
</tr>
<tr>
<td>Thr</td>
<td>Thr</td>
<td>Ser</td>
</tr>
<tr>
<td>Thr</td>
<td>Thr</td>
<td>Ser</td>
</tr>
<tr>
<td>Thr</td>
<td>Thr</td>
<td>Ser</td>
</tr>
<tr>
<td>Ala</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Ala</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Ala</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Ala</td>
<td>Ala</td>
<td>Asp</td>
</tr>
</tbody>
</table>

The Genetic Code
Scoring proteins

- Scoring protein sequence alignments is a much more complex task than scoring DNA
  - Not all substitutions are equal
- Problem was first worked on by Pauling and collaborators
- In the 1970s, M. Dayhoff created the first widely used similarity matrices.
  - “One size does not fit all”
  - Homologous proteins which are evolutionarily close should be scored differently than proteins that are evolutionarily distant
  - Different proteins might evolve at different rates and we need to normalize for that
- **PAM1[A,B]:**
  - Probability [Residue A is substituted by residue B] given that the two sequences differ by 1%
PAM1 matrix

- Align many proteins that are very similar
  - Is this a problem?
- Estimate the natural frequency $P_a$ for each amino-acid $a$
- Estimate the frequency $P_{ab}$ of alignment for each pair of amino-acids $a,b$
- $PAM$ matrix $P_{b|a} = P_{ab}/P_a$
- $S(a,b) = \log_{10}(P_{ab}/P_a P_b) = \log_{10}(P_{b|a}/P_b)$
### PAM1 Mutation Matrix

1 PAM evolutionary distance

|       | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Ala   | 9867| 2   | 9   | 10  | 3   | 8   | 17  | 21  | 2   | 6   | 4   | 2   | 6   | 2   | 22  | 35  | 32  | 0   | 2   | 18  |
| Arg   | 1   | 9833| 1   | 0   | 1   | 10  | 0   | 10  | 3   | 1   | 19  | 4   | 1   | 4   | 4   | 1   | 0   | 0   | 1   |
| Asn   | 4   | 1   | 9822| 36  | 0   | 4   | 6   | 6   | 21  | 3   | 1   | 13  | 0   | 1   | 2   | 20  | 9   | 1   | 4   | 1   |
| Asp   | 6   | 0   | 42  | 9859| 0   | 6   | 51  | 64  | 4   | 1   | 0   | 0   | 0   | 0   | 4   | 1   | 3   | 3   | 0   | 0   |
| Cys   | 1   | 1   | 0   | 9973| 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 1   | 5   | 1   | 0   | 1   |
| Gln   | 3   | 8   | 4   | 9856| 27  | 1   | 23  | 1   | 3   | 5   | 6   | 4   | 0   | 0   | 2   | 2   | 2   | 0   | 0   | 1   |
| Glu   | 10  | 0   | 7   | 56  | 0   | 15  | 9885| 4   | 2   | 3   | 1   | 4   | 1   | 0   | 3   | 4   | 2   | 0   | 1   |
| Gly   | 21  | 1   | 12  | 11  | 1   | 3   | 9975| 1   | 0   | 1   | 2   | 1   | 1   | 3   | 21  | 3   | 0   | 0   | 5   |
| His   | 1   | 8   | 18  | 1   | 20  | 1   | 0   | 9912| 0   | 1   | 1   | 0   | 2   | 3   | 1   | 1   | 1   | 1   | 4   | 1   |
| Ile   | 2   | 2   | 3   | 1   | 2   | 1   | 2   | 0   | 0   | 9932| 9   | 2   | 12  | 7   | 0   | 1   | 7   | 0   | 1   | 11  |
| Leu   | 3   | 1   | 3   | 9   | 6   | 1   | 1   | 4   | 22  | 9947| 2   | 45  | 11  | 3   | 1   | 3   | 4   | 2   | 15  |
| Lys   | 2   | 17  | 25  | 8   | 0   | 12  | 7   | 2   | 2   | 4   | 1   | 9926| 20  | 0   | 3   | 0   | 11  | 0   | 1   |
| Met   | 1   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 5   | 8   | 4   | 9874| 1   | 0   | 1   | 2   | 0   | 0   | 4   |
| Phe   | 1   | 1   | 1   | 0   | 0   | 0   | 0   | 0   | 2   | 1   | 0   | 0   | 0   | 4   | 9946| 0   | 1   | 0   | 1   |
| Pro   | 13  | 5   | 2   | 1   | 1   | 8   | 3   | 2   | 5   | 1   | 2   | 2   | 1   | 1   | 9926| 12  | 4   | 0   | 0   | 2   |
| Ser   | 28  | 11  | 34  | 7   | 11  | 4   | 6   | 16  | 2   | 2   | 1   | 1   | 7   | 4   | 3   | 17  | 9940| 38  | 5   | 2   |
| Thr   | 22  | 2   | 13  | 4   | 1   | 3   | 2   | 2   | 2   | 11  | 2   | 8   | 6   | 1   | 5   | 33  | 9871| 0   | 2   | 9   |
| Trp   | 0   | 2   | 0   | 0   | 2   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| Tyr   | 1   | 0   | 3   | 0   | 0   | 1   | 0   | 0   | 4   | 1   | 1   | 0   | 0   | 21  | 0   | 1   | 1   | 2   | 9945| 1   |
| Val   | 11  | 2   | 1   | 1   | 3   | 2   | 2   | 3   | 3   | 97  | 11  | 1   | 17  | 1   | 2   | 2   | 10  | 0   | 2   | 9901|
Higher PAMs

- Two sequences are 1 PAM apart if they differ in 1% of residues
- When are two sequences 2 PAMs apart?

2 PAMs apart does not mean 2% residues are different
Higher PAMs

- Two sequences are 1PAM apart if they have 1% divergence.
- Two sequences s and t are PAM250 units apart if there exists a sequence s’ s.t.
  - s and s’ are PAM1 apart
  - s’ and t are PAM249 units apart

\[
\begin{align*}
P_{250}[A \square B] &= \sum_{X} P_{1}[A \square X]P_{249}[X \square B] \\
&= P_{1}^{250}[A \square B]
\end{align*}
\]
Note: This is not the score matrix: What happens as you keep increasing the power? (This is not an easy question, so think about it)
PAM matrices and scoring

- Suppose we know that two sequences are PAM$_{250}$ apart. Then, the score for a residue
  - $S[a,b] = \text{PAM}_{250}[b|a] / \text{Pr}[b]$
BLOSUM series of Matrices

- Henikoff & Henikoff: Sequence substitutions in evolutionarily distant proteins do not seem to follow the PAM distributions.
- A more direct method based on hand-curated multiple alignments of distantly related proteins from the BLOCKS database.
- BLOSUM60 Merge all proteins that have greater than 60% identity. Then, compute the substitution probability.
  - In practice BLOSUM62 seems to work very well.
What is the correspondence?

PAM1  Blosum1
  :          :
  :          :

PAM100  Blosum60

PAM250  Blosum100
Why is Blast Fast?
Large database search

Database size $n=10M$, Query size $m=300$.
$O(nm) = 3 \times 10^9$ computations
Observations

- Much of the database is random from the query’s perspective
- Consider a random DNA string of length n.
  - \( \Pr[A] = \Pr[C] = \Pr[G] = \Pr[T] = 0.25 \)
- Assume for the moment that the query is all 1’s (length m).
- What is the probability that an exact match to the query can be found?
Basic probability

- Probability that there is a match starting at a fixed position $i = 0.25^m$
- What is the probability that some position $i$ has a match.
- Dependencies confound probability estimates.
Basic Probability: Expectation

Q: Toss a coin: each time it comes up heads, you get a dollar

- What is the money you expect to get after \( n \) tosses?
- Let \( X_i \) be the amount earned in the \( i \)-th toss

\[
E(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p
\]

- Total money you expect to earn

\[
E\left( \sum_{i} X_i \right) = \sum_{i} E(X_i) = np
\]
Expected number of matches

- Expected number of matches can still be computed.

- Let $X_i = 1$ if there is a match starting at position $i$, $X_i = 0$ otherwise.

  $$\Pr(\text{Match at Position } i) = p_i = 0.25^m$$

  $$E(X_i) = p_i = 0.25^m$$

- Expected number of matches is

  $$E(\prod_i X_i) = \prod_i E(X_i) = n\left(\frac{1}{4}\right)^m$$
Expected number of exact Matches is small!

- Expected number of matches = \( n \times 0.25^m \)
  - If \( n=10^7 \), \( m=10 \),
    - Then, expected number of matches = 9.537
  - If \( n=10^7 \), \( m=11 \)
    - expected number of hits = 2.38
  - \( n=10^7, m=12 \),
    - Expected number of hits = 0.5 < 1
- Bottom Line: An exact match to a substring of the query is unlikely just by chance.