L4

Linear space
Scoring matrices
Silly Quiz

• Name an early Bioinformatics Researcher.

• Name an early female Bioinformatics Researcher.
What did we do in the last few lectures?

- Global and Local alignment
- Affine gap costs. Why?
- Space requirement.
Alignment Space?

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

- How much space do we need?
- Is the space requirement too much?
Fig. 1. Dot-plot representation of sample assembly comparison results

Aligning long sequences may require a lot of space

Alignment (Linear Space)

- Score computation

For $i = 1$ to $n$
  For $j = 1$ to $m$

\[
\begin{align*}
S[i, j] &= \max \begin{cases} 
S[i-1, j-1] + C(s_i, t_j) \\
S[i-1, j] + C(s_i, -) \\
S[i, j-1] + C(-, t_j)
\end{cases} \\
i_2 &= i \% 2; \quad i_1 = (i - 1) \% 2;
\end{align*}
\]
Linear Space Alignment

• In Linear Space, we can do each row of the D.P.
• We need to compute the optimum path from the origin \((0,0)\) to \((m,n)\)
Linear Space (cont’d)

• At $i=n/2$, we know scores of all the optimal paths ending at that row.
• Define $F[j] = S[n/2,j]$
• One of these $j$ is on the true path. Which one?
Backward alignment

• Let $S_b[i,j]$ be the optimal score of aligning $s[i+1..n]$ with $t[j+1..m]$

$$S_b[i,j] = \max \begin{cases} 
S_b[i+1,j+1] + C(s_{i+1},t_{j+1}) \\
S_b[i+1,j] + C(s_{i+1},-) \\
S_b[i,j+1] + C(-,t_{j+1}) 
\end{cases}$$

• Boundary cases?
• $S_b[n,j]$? $S_b[m,j]$?
Backward alignment

- Let $S_b[i,j]$ be the optimal score of aligning $s[i+1..n]$ with $t[j+1..m]$
- Define $B[j] = S_b[n/2,j]$
- One of these $j$ is on the true path. Which one?
Forward, Backward computation

- At the optimal coordinate, \( j \)
  \[ F[j] + B[j] = S[n,m] \]
- In \( O(nm) \) time, and \( O(m) \) space, we can compute one of the coordinates on the optimum path.
Linear Space Alignment

- Align(1..n,1..m)
  - For all 1<=j <= m
    - Compute F[j]=S(n/2,j)
  - For all 1<=j <= m
    - Compute B[j]=S_b(n/2,j)
  - j* = max_j {F[j]+B[j] }
  - X = Align(1..n/2,1..j*)
  - Y = Align(n/2+1..n,j*+1..m)
  - Return X,j*,Y
Linear Space complexity

- $T(nm) = c.nm + T(nm/2) = O(nm)$
- Space = $O(m)$
Scoring Matrices

• We have seen that affine gap penalties help concentrate the gaps in small regions.
• What about substitution errors. Are all substitutions alike?
Scoring DNA

- DNA has structure.
DNA scoring matrices

- So far, we considered a simple match/mismatch criterion.
- The nucleotides can be grouped into Purines (A,G) and Pyrimidines.
- Nucleotide substitutions within a group (transitions) are more likely than those across a group (transversions)
Scoring matrices for DNA

- Transversions are more heavily penalized than transitions.
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, Margaret Dayhoff created the first 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
Score function for proteins

- Suppose we are searching with a mouse protein.
- Blast returns proteins ranked by score
  - Top hit is to human
  - Somewhere below is Drosophila
  - Which one will you trust?

```
hum2 ! 75% identity
hum !
mus ! 40% identity
dros
```
It is all about expectations
Score function for proteins

- Paralogs arise via gene duplications
- They rapidly diverge and take different functions
- The expected score is different when looking at human and mouse versus mouse and drosophila
- We need to score drosophila and mouse separately from human and mouse
- In this example, if the expectation is 33% identity, then a 50% identity is great.

\[ \text{hum-paralog} \]

\[ \text{hum} \]

\[ \text{mus} \]

\[ \text{dros} \]

50% identity

75% identity
Frequency based scoring

<table>
<thead>
<tr>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
</tr>
</tbody>
</table>

- Our goal is to score each column in the alignment.
- Comparing against expectation:
  - Think about alignments of pairs of random sequences, and compute the probability that A and B appear together just by chance $P^R(A,B)$.
  - Compute the probability of A and B appearing together in the alignment of related sequences (orthologs) $P^O(A,B)$.
- A good score function? $\log\left(\frac{P^O(A,B)}{P^R(A,B)}\right)$.
Log-odds scoring

\[ S(A,B) = \log \left( \frac{P^O(A,B)}{P^R(A,B)} \right) = \log \left( \frac{P^O(A \mid B)}{P_A} \right) \]

- How can we compute \( P^o_{a \mid b} \)?
  - We need good alignments, but....
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, Margaret Dayhoff created the first 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
END of L4
PAM 1 distance

- Two sequences are 1 PAM apart if they differ in 1 % of the residues.

- $PAM_1(a,b) = \Pr[\text{residue } b \text{ substitutes residue } a, \text{ when the sequences are 1 PAM apart}]$
PAM1 matrix

- Align many proteins that are very similar
  - Is this a problem?
- 1 PAM evolutionary distance represents the time in which 1% of the residues have changed
- Estimate the frequency $P_{b|a}$ of residue $a$ being substituted by residue $b$.
- $P_{a|b} = Pr(b \text{ will mutate to an } a \text{ after 1 PAM evolutionary distance})$
- Scoring matrix
  - $S(a,b) = \log_{10}(P_{ab}/P_a P_b) = \log_{10}(P_{a|b}/P_a)$
PAM 1

- Top column shows original, and left column shows replacement residue = PAM1(a,b) = Pr(a|b)

|     | 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   | 22  | 35  | 32  | 0   | 2   | 18  |
| Arg | 1   | 9913| 1   | 0   | 1   | 10  | 0   | 0   | 10  | 3   | 1   | 19  | 4   | 4   | 6   | 1   | 8   | 0   | 1   |
| Asn | 4   | 1   | 9822| 36  | 0   | 4   | 6   | 6   | 21  | 3   | 1   | 13  | 0   | 1   | 2   | 20  | 9   | 1   | 4   | 0   |
| Asp | 6   | 0   | 42  | 9859| 0   | 6   | 53  | 6   | 4   | 1   | 0   | 3   | 0   | 0   | 1   | 5   | 3   | 0   | 1   | 1   |
| Cys | 1   | 1   | 0   | 0   | 9973| 0   | 0   | 0   | 0   | 1   | 1   | 0   | 0   | 0   | 1   | 5   | 3   | 0   | 1   | 1   |
| Gln | 3   | 9   | 4   | 5   | 0   | 9876| 27  | 1   | 23  | 1   | 3   | 6   | 4   | 0   | 6   | 2   | 2   | 0   | 0   | 1   |
| Glu | 10  | 0   | 7   | 56  | 0   | 35  | 9865| 4   | 2   | 3   | 1   | 4   | 1   | 0   | 3   | 4   | 2   | 0   | 1   | 1   |
| Gly | 21  | 1   | 12  | 11  | 1   | 3   | 7   | 9935| 1   | 0   | 1   | 2   | 1   | 1   | 1   | 3   | 21  | 3   | 0   | 0   |
| His | 1   | 8   | 18  | 3   | 1   | 20  | 1   | 0   | 9912| 0   | 1   | 1   | 0   | 2   | 3   | 1   | 1   | 1   | 4   | 1   |
| Ile | 2   | 2   | 3   | 1   | 2   | 1   | 2   | 0   | 2   | 9   | 2   | 12  | 7   | 0   | 1   | 7   | 0   | 1   | 33  |
| Leu | 3   | 1   | 3   | 0   | 0   | 6   | 1   | 1   | 4   | 22  | 9947| 2   | 45  | 13  | 3   | 1   | 3   | 3   | 4   | 2   | 15  |
| Lys | 2   | 37  | 25  | 6   | 0   | 12  | 7   | 2   | 2   | 4   | 1   | 9926| 20  | 0   | 3   | 8   | 11  | 0   | 1   | 1   |
| Met | 1   | 1   | 0   | 0   | 2   | 0   | 0   | 5   | 0   | 8   | 9874| 1   | 0   | 1   | 2   | 0   | 0   | 0   | 0   | 4   |
| Phe | 1   | 1   | 0   | 0   | 1   | 0   | 1   | 2   | 8   | 6   | 0   | 4   | 9946| 0   | 2   | 1   | 3   | 28  | 0   |
| Pro | 13  | 5   | 2   | 1   | 1   | 8   | 3   | 2   | 5   | 1   | 2   | 2   | 1   | 1   | 9926| 12  | 4   | 0   | 0   | 0   |
| Ser | 28  | 11  | 34  | 7   | 11  | 4   | 6   | 16  | 2   | 2   | 1   | 7   | 4   | 3   | 17  | 9840| 38  | 5   | 2   | 2   |
| Thr | 22  | 2   | 13  | 4   | 1   | 3   | 2   | 2   | 1   | 11  | 2   | 8   | 6   | 1   | 5   | 32  | 9871| 0   | 2   | 9   |
| Trp | 0   | 2   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 1   | 0   |
| Tyr | 1   | 0   | 3   | 0   | 3   | 0   | 1   | 0   | 4   | 1   | 1   | 0   | 0   | 21  | 0   | 1   | 1   | 1   | 2   | 9945|
| Val | 13  | 2   | 1   | 1   | 3   | 2   | 2   | 3   | 3   | 57  | 11  | 1   | 17  | 1   | 3   | 2   | 10  | 0   | 2   | 9901|
PAM and evolutionary time

• Assume that mutations occur at a constant rate (molecular clock assumption).
• Therefore if 2 sequences are 1PAM apart, they have diverged for some (say, N) years
PAM distance

- Two sequences are 1 PAM apart when they differ in 1% of the residues.
- When are 2 sequences 2 PAMs apart?
Generating Higher PAMs

- \( \text{PAM}_2(a,b) = \sum_c \text{PAM}_1(a,c) \cdot \text{PAM}_1(c,b) \)
- \( \text{PAM}_2 = \text{PAM}_1 \times \text{PAM}_1 \) (Matrix multiplication)
- \( \text{PAM}_{250} = \text{PAM}_1 \times \text{PAM}_{249} \)
- \( \text{PAM}_{250} = \text{PAM}_1^{250} \)
Note: This is not the score matrix:
What happens as you keep increasing the power?
Scoring residues

- A reasonable score function $C(a,b)$ is given as follows:
  - Look at ‘high quality’ alignments
  - $C(a,b)$ should be high when $a,b$ are seen together more often than is expected by chance
  - $C(a,b)$ should be low, otherwise.
- How often would you expect to see $a,b$ together just by chance?
  - $P_a P_b$
- Let $P_{ab}$ be the probability that $a$ and $b$ are aligned in a high-quality alignment
- A good scoring function is the log-odds score
  - $C(a,b)= \log_{10} \left( \frac{P_{ab}}{P_a P_b} \right)$
Scoring alignments

• To compute $P_{ab}$, we need ‘high-quality’ alignments
• How can you get quality alignments?
  – Use SW (But that needs the scoring function)
  – Build alignments manually
  – Use Dayhoff’s theory to extrapolate from high identity alignments
Scoring using PAM matrices

- Suppose we know that two sequences are 250 PAMs apart.
- \( S(a,b) = \log_{10}(P_{ab}/P_a P_b) = \log_{10}(P_{a|b}/P_a) = \log_{10}(PAM_{250}(a,b)/P_a) \)
- How does it help?
  - \( S_{250}(A,V) \gg S_1(A,V) \)
  - Scoring of hum vs. Dros should be using a higher PAM matrix than scoring hum vs. mus.
  - An alignment with a smaller % identity could still have a higher score and be more significant
PAM250 based scoring matrix

\[
\begin{array}{ccccccccccccccc}
C & 12 \\
S & 0 2 \\
T & -2 1 3 \\
P & -3 1 0 6 \\
A & -2 1 1 1 2 \\
G & -3 1 0 -1 1 5 \\
N & -4 1 0 -1 0 0 2 \\
D & -5 0 0 -1 0 1 2 4 \\
E & -5 0 0 -1 0 0 1 3 4 \\
Q & -5 -1 -1 0 0 -1 1 2 2 4 \\
H & -5 -1 -1 0 -1 -2 2 1 1 3 6 \\
R & -4 0 -1 0 -2 -3 0 -1 -1 1 2 6 \\
K & -5 0 0 -1 -1 -2 1 0 0 1 0 3 5 \\
M & -5 -2 -1 -2 -1 -3 -2 -3 -2 -1 -2 0 0 6 \\
I & -2 -1 0 -2 -1 -3 -2 -2 -2 -2 -2 -2 -2 2 5 \\
L & -6 -3 -2 -3 -2 -4 -3 -4 -3 -2 -2 -3 -3 4 2 6 \\
V & -2 -1 0 -1 0 -1 -2 -2 -2 -2 -2 -2 -2 2 4 4 \\
F & -4 -3 -3 -5 -4 -5 -3 -6 -5 -5 -2 -4 -5 0 1 2 -1 9 \\
Y & 0 -3 -3 -5 -3 -5 -2 -4 -4 -4 0 -4 -4 2 -1 -1 -2 7 10 \\
W & -8 -2 -5 -6 -6 -7 -4 -7 -7 -5 -3 2 -3 -4 -5 -2 -6 0 0 17 \\
\end{array}
\]

\[S250(a,b) = \log_{10}(Pab/PaPb) = \log_{10}(PAM250(a,b)/Pa)\]
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% similarity. Then, compute the substitution probability.
  - In practice, BLOSUM62 seems to work very well.
  - [Blast Parameters](#)
PAM vs. BLOSUM

- What is the correspondence?

- PAM1
- PAM2
- Blosum1
- Blosum2
- Blosum62
- PAM250
- Blosum100
END of L4