Exam 1 scores

Mean = 92.8
Median = 94

1(b) What are $M$, $m$, and $g$?

$M = 1$
$M = 1$
$g = -1$
$g = -1$
$m = -2$
$m = -2$

Three possibilities:
- $m > -2$ Not possible because $1+m$ would replace -1 with a better score.
- $m = -2$ $1+m = 1+2g = -1$
- $m < -2$ Violates $m > 2g$
1(b) Which algorithm?

Semi-global alignment
- Not global: top row has 0's
- Not local: the matrix has some negative values

1(c) What are $M$, $m$, and $g$?

A 0-valued match cell in the first column ->
- Distance function
  - $d(x,x) = 0$
  - First column must be 0's

1(c)

$d(x,x) = 0$
$d(x,_) = 1$

0+g: $\Rightarrow d(x,_) = g = 1$

Require $d(x,y)>0$

1+g: 1+g > 0+g
1(c) Which algorithm?

Semi-global alignment
- Not global: First column must have 0s if the second column has 0s and 1s
- Not local: Local alignment requires a similarity function

3) Consider k sequences \( s_1, s_2, \ldots, s_k \) and let \( M = 1 \) and \( m = g = -1 \) be the scoring function.

Let
- \( X \) be the score of the optimal global alignment of \( s_1 \) and \( s_2 \).
- \( Y \) be the score of the optimal local alignment of \( s_1 \) and \( s_2 \).
- \( Z \) be the score of the pairwise alignment of \( s_1 \) and \( s_2 \) induced by the optimal global alignment of \( s_1, s_2, \ldots, s_k \).

Give an inequality that expresses the relationship between \( X \), \( Y \) and \( Z \).

Explain your answer.

\[ Y \geq X \]

Why? \( Y \) must be positive. \( X \) can be negative.
3) Consider $k$ sequences $s_1, s_2, \ldots, s_k$ and let $M = 1$ and $m = g = -1$ be the scoring fn. Let
- $X$ be the score of the optimal global alignment of $s_1$ and $s_2$.
- $Y$ be the score of the optimal local alignment of $s_1$ and $s_2$.
- $Z$ be the score of the pairwise alignment of $s_1$ and $s_2$ induced by the optimal global alignment of $s_1, s_2, \ldots, s_k$.

Give an inequality that expresses the relationship between $X$, $Y$, and $Z$. Explain your answer.

\[
X \geq Z
\]

Why? The induced alignment reflects relationships in the entire family. The optimal alignment of $s_1$ and $s_2$ with respect to $s_3, \ldots, s_k$ may result in a sub-optimal alignment between $s_1$ and $s_2$.

5(a)

Exact dynamic programming requires time $t_1$ to obtain an MSA of $k$ sequences of length $n$. Let $t_2$ be the time required to obtain an MSA of $k$ sequences of length $3n$ using the same method. What is $t_2/t_1$?

\[
t_1 = O(n^k2^kk^2)
\]
\[
t_2 = O((3n)^k2^kk^2)
\]

\[
\frac{t_2}{t_1} = \frac{O((3n)^k2^kk^2)}{O(n^k2^kk^2)}
\]

\[
\frac{t_2}{t_1} = 3^k
\]
5(b)

Exact dynamic programming requires time $t_1$ to obtain an MSA of $k$ sequences of length $n$. Let $t_3$ be the time required to obtain an MSA of $3k$ sequences of length $n$ using the same method. What is $t_3/t_1$?

$$
t_1 = O(n^k 2^k k^2)
$$

$$
t_3 = O(n^{3k} 2^{3k} (3k)^2)
$$

$$
\frac{t_3}{t_1} = O\left(\frac{n^{3k} 2^{3k} (3k)^2}{n^k 2^k k^2}\right)
$$

5(c)

$O(3^k)$ versus $O(n^2 4^k)$

Increasing $k$ has more impact because

- “increasing $k$ is exponentially increasing the time, while increasing $n$ is adding a constant multiplier”

- “This makes sense because by increasing $n$, you’re simply adding extra boxes in the same dimensions, but by increasing $k$, you’re adding extra dimensions”
Jukes-Cantor model: sequence substitution at a single site

Rate of substitutions
\[ P(xy) = 3\alpha \]

Probability nucleotide \( x \) remains unchanged
\[ P(xx) = 1 - 3\alpha \]

In a discrete time framework, \( \alpha \) is the probability of a given substitution occurring in a single time step.

Given ancestral nucleotide \( z \), the probability of observing nucleotide \( x \) after time \( t \) is given by

\[ P_{xx}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} , \quad P_{zx}(t) = \frac{1}{4} - \frac{3}{4}e^{-4\alpha t} , \quad y \neq x \]

Note that:
- At \( t=0 \), \( P_{xx}(0) = 1 \) and \( P_{xy}(0) = 0 \). This makes sense because if no time has elapsed, then no substitution can have occurred (yet).
- As \( t \to \infty \), \( P_{xx}(t) = P_{yy}(t) = 0.25 \). This says that the steady state distribution of nucleotide frequencies is uniform under the Jukes-Cantor model.

A more complex model

Different probabilities for transitions and transversions

2. Given an alignment of \( n \) nucleotides that differs at \( m \) positions, the expected number of substitutions since the divergence of the two sequences is given by

\[ E_{\text{sub}} = -\frac{3}{4} \ln \left( 1 - \frac{4m}{3n} \right)n. \]

For example, if we observe 200 mismatches in an alignment of 1000 nucleotides, then the number of actual substitutions is

\[ -\frac{3}{4} \ln \left( 1 - \frac{4 \times 200}{3 \times 1000} \right) \times 1000 = 233 \text{ substitutions} \]

Note that:
- If \( m = 0 \), then \( E_{\text{sub}} = 0 \) and the distance between the sequences is zero.
- \( m/n \leq 0.75 \) in sequences governed by the Jukes Cantor model.
- As \( m/n \to 0.75 \), \( E_{\text{sub}} \to \infty \). This is because once we reach the steady state distribution of nucleotide frequencies, \( m/n \) provides no information about how long the sequences have been diverging.
Transitions and Transversions

**Pyrimidines** have one ring

**Purines** have two rings

Transitions: substitutions within the same class of nucleotide (purine – purine or pyrimidine-pyrimidine)

Transversions: substitutions between classes (purine – pyrimidine or pyrimidine-purine)

Kimura 2 Parameter model

Rate of substitutions

\[ P(xy) = \alpha + 2\beta \]

Probability nucleotide \( x \) remains unchanged

\[ P(xx) = 1 - \alpha - 2\beta \]

In a single time step

\[ \sigma^* = \left\{ \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right\} \]

From this, we can derive several quantities of interest...

Jukes-Cantor model (1969)

- All substitutions have equal probability
- Base frequencies are equal

\[ p(A) = 0.25 \]
\[ p(C) = 0.25 \]
\[ p(T) = 0.25 \]

Given an alignment of \( n \) nucleotides that differs at \( m = m_t + m_v \) positions, where

\[ m_t = \text{number of transitions}, \]
\[ m_v = \text{number of transversions}, \]

the expected number of substitutions is given by

\[ E[\text{sub}] = \left[ -\frac{1}{2} \ln \left( 1 - \frac{m_t}{n} - \frac{m_v}{n} \right) - \frac{3}{4} \ln \left( 1 - \frac{4 m_v}{3 n} \right) \right] n \]
Kimura 2 parameter model (K2P) (1980)

- Transitions and transversions have different probabilities
- Base frequencies are equal

$$p(A) = 0.25$$
$$p(G) = 0.25$$
$$p(C) = 0.25$$
$$p(T) = 0.25$$

Hasegawa, Kishino & Yano (HKY) (1985)

- Transitions and transversions have different probabilities
- Unequal base frequencies

$$p(A) = \pi_A$$
$$p(G) = \pi_G$$
$$p(C) = \pi_C$$
$$p(T) = \pi_T$$

Felsenstein (1981)

- All substitutions have equal probability
- Unequal base frequencies

$$p(A) = \pi_A$$
$$p(G) = \pi_G$$
$$p(C) = \pi_C$$
$$p(T) = \pi_T$$

General Time Reversible model

- All six pairs have different substitution frequencies
- Unequal base frequencies
DNA substitution models

- Four states (A, C, G, T)
- Model specifies the probability of substitution for all possible pairs of nucleotides

<table>
<thead>
<tr>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DNA substitution model (e.g., JC, K2P, GTr)

Amino acid substitution models

- Twenty states (A, C, ... Y)
- Model specifies the probability of substitution for all possible pairs of amino acids

Amino acid substitution matrix (e.g., PAM, WAG, JTT, MtREV etc)